DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master´s Thesis in Informatics: Computational Science and Engineering

Approximating many-electron wave function with physics-aware surrogate models

Dawid Pasterny

DEPARTMENT OF INFORMATICS

TECHNISCHE UNIVERSITÄT MÜNCHEN

Master´s Thesis in Informatics: Computational Science and Engineering

Approximating many-electron wave function with physics-aware surrogate models

Author:	Dawid Pasterny
Supervisor:	Dr. Felix Dietrich
Advisor:	Chinmay Datar
Submission Date:	March 29th. 2023

I confirm that this master's thesis in informatics: computational science and engineering is my own work and I have documented all sources and material used.

Munich, March 29th. 2023

Dawid Pasterny

Dawid Pasterny

Abstract

With the ability to obtain sufficiently accurate and computationally accessible solutions to the many-body Schrödinger equation most of modern chemistry and material science could renounce phenomenology and enter the era of *ab-initio* modeling. Since its publication in 1926, the equation received a lot of attention with regards to finding an accurate, yet efficient approximation which would capture the most relevant quantum behavior while remaining polynomial in computational complexity. To this end, one often resorts to methods based on the mean-field approximation, in particular Hartree-Fock (HF) or the highly popular and flexible Kohn-Sham density-functional theory (DFT). Central to both methods lies the interpretation of the nonlinear eigenvalue problem constituted by the stationary Schrodinger equation for the ground state energy as variational minimization procedure. Both methods, in order to tackle that minimization analytically, restrict their computational basis to a linear combination of orthogonal one-electron wave functions only, which has a severe impact on the overall efficacy of those algorithms. A different family of electronic structure methods known as Quantum Monte Carlo (QMC), overcomes this limitation allowing for an arbitrarily complex wave function ansatz which in recent years has led to a surge in interest from deep learning community which lead to formulation of so called Neural Quantum States (NQS). The price one pays for this freedom, however, is the necessity of solving high-dimensional integrals in order to evaluate the energy expectation objective and a subsequent stochastic optimization procedure of the variational parameters. Given the flexibility of neural networks, that is, their inherent lack of any underlying discretization, other than, perhaps the machine precision, of crucial importance becomes performance of the Monte Carlo simulation, which determines which regions of the wave function get assigned most representational resources. In this work, we address the topic of sampling using physics-aware deep learning surrogates of the wave function Ψ within the Quantum Monte Carlo framework, in particular our aim was to develop a sampling algorithm which could deal with the pathologies of the so called Born probability density $|\Psi|^2$. As a secondary endeavour, we have gone to great lengths in providing an extensive, although not exhaustive, review of the state-of-the-art methodology of this young field alongside the traditional, computational approaches of quantum chemistry and physics of fermionic systems published hitherto. Through this work, we wish to improve upon prevailing accuracy-efficiency dilemma in computational quantum mechanics and to further tighten the progressing integration of deep learning and scientific computing which promises significant improvements over the often purely phenomenological elements of the current state-of-the-art methods, closing the gap to the elusive problems of practical significance encountered in engineering, chemistry or medicine.

Contents

1. Introduction

Crucial to the understanding of emergence of macroscopic properties of matter is the knowledge of its electronic structure. It has proven instrumental in developing theories like the free-electron model of A.Sommerfeld and H.Bethe explaining conductivity in crystalline solids, the molecular orbitals theory description of covalent bonding of atoms developed by Hund, Muliken, Hueckel and others or the theoretical underpinnings of magnetic, thermal and optical properties to name the few [1]. The outburst of successful material models in the second half of the twentieth century has been completely facilitated by the revolutionary ideas that came to be known as quantum physics. By the year 1929, three years after the publication of the much celebrated wave equation by Erwin Schrödinger, the field has not only had thorough experimental underpinnings but also a successful theoretical description, leading Paul Dirac to utter the following words:

> "The underlying physical laws necessary for the mathematical theory of a large part of physics and the whole of chemistry are thus completely known, and the difficulty is only that the exact application of these laws leads to equations much too complicated to be soluble. It therefore becomes desirable that approximate practical methods of applying quantum mechanics should be developed, which can lead to an explanation of the main features of complex atomic systems without too much computation." - P.Dirac (1929)

Although quantum mechanics has been first consistently formulated by Werner Heisenberg, Max Born and Pascual Jordan in 1925 using the language of matrix algebra [2] it is usually the so-called Schrödinger picture which we assimilate with the theory of quanta. Its general form reads:

$$i\hbar\frac{\partial}{\partial t}\Psi(\mathbf{r},t) = \hat{H}\Psi(\mathbf{r},t) = \frac{-\hbar^2}{2m}\nabla^2\Psi(\mathbf{r},t) + U(\mathbf{r})\Psi(\mathbf{r},t) \tag{1.1}$$

and describes time evolution of a complex-valued wave function $\Psi(\mathbf{r},t) : \mathbb{R}^n \to \mathbb{C}$. In the above, $\hat{H}$ is the system's Hamiltonian - the total energy operator introduced by R. Hamilton in his alternative formulation of classical mechanics [3] and consists of the kinetic $\frac{(i\hbar\partial_x)^2}{2m}$ and potential $U(r)$ energy terms which are themselves quantum operators. Hamiltonian mechanics will be dealt with in greater detail in later sections with regards to Hamiltonian Monte Carlo, an efficient sampling technique used as a subroutine in the Quantum Monte Carlo method. For now, it is only important to realize it is a system-dependent, *hermitian* operator which *generates* its evolution in time.

With no reference to the underlying physics, the wave function is supposed to represent, one can further state that equation 1.1 [4]:

- is a *linear* differential equation - if $\Psi_1(r,t), \Psi_2(r,t)$ solve the equation $\alpha\Psi_1(r,t) + \beta\Psi_2(r,t)$ does too,
- is *unitary* - time evolution of Ψ preserves its "length" since a complex exponent of a one parameter hermitian operator is unitary (denoted $\hat{U}$) [5]

$$\Psi(r,t) = \Psi(r,0)e^{\frac{-i}{\hbar}\hat{H}t} = \hat{U}(t)\Psi(r,0) \tag{1.2}$$

- is *deterministic* - given unique initial conditions and evolution equation unambiguously determines future states.

To further elaborate on Schrödinger's equation, it shall be mentioned that he himself actually did not have an interpretation for his own creation, this is perhaps best attested by existence of mockery like:

> "Gar Manches rechnet Erwin schon
> Mit seiner Wellenfunktion.
> Nur wissen m'ocht man gerne wohl,
> Was man sich dabei vorstell'n soll"

Indeed, he was not found of the already existing Heisenberg's matrix formulation of quantum mechanics which treated nature as discrete and hoped to prove him wrong by working with smooth waves instead of vectors, an intuition induced by de Brogile's idea of wave-particle duality and motivated by the lack of a corresponding equation for their evolution [6]. He obtained his equation from exceptionally beautiful, although considered unneeded, yet another reformulation of classical mechanics - the Hamilton-Jacobi partial differential equation. As reported by Cornelius [3], the path to the formulation of wave equation led from the Hamilton's opto-mechanical investigations, through Delunay's treatment of separable, multiply-periodic mechanical systems, Sommerfeld-Wilson quantum conditions, their invariant formulation by Einstein, de Brogile's resonance interpretation thereof and finally Schrödinger's logarithmic transformation from the phase function S to the wave function Ψ.

Although Schrödinger wanted his wave to have physical meaning, he could not find one, and a common interpretation nowadays is that there is none, at least in Copenhagen school of quantum mechanics, with N. Bohr at its forefront [7]. The wave function is thought to describe the *state* of a quantum system, that is, it contains all the necessary information, but a one-to-one correspondence with the actual behavior of matter is not present - instead it is defined only probabilistically. According to so called *Born rule*, the modulo squared of Ψ represents the probability density function of observing a system in any particular eigenstate of a quantum operator $\hat{O}$ associated with a physical observable o [8]:

$$|\Psi(\mathbf{r},t)|^2 = Re(\Psi(\mathbf{r},t))^2 + Im(\Psi(\mathbf{r},t))^2 = p(\mathbf{r},t) \tag{1.3}$$

More explicitly, by considering the energy norm of Ψ under $\hat{O}$ [1] we obtain [4]:

$$\begin{gathered}\Psi(r) = \int_O do\ c_o \phi_o(r) \quad or \quad \sum_o c_o \phi_o(r) \\ \Downarrow \\ \langle \Psi | \hat{O} | \Psi \rangle = \left\langle \int_O do\ c_o \phi_o \middle| \hat{O} \middle| \int_O do\ c_o \phi_o \right\rangle = \int_O do\ |c_o|^2 o \, \langle \phi_o | \phi_o \rangle = \int_O do\ |c_o|^2 o := \mathbb{E}[\hat{O}]\end{gathered} \tag{1.4}$$

where ϕ_o is an eigenstate corresponding to any particular realization of $\hat{O}$. The above, begs an interpretation of the modulo squared of the expansion coefficients $|c_o|^2$ as the probability density function of a random variable O with realizations o.

Since c_o contain the same information as the wave function does - indeed, they *are* the wave function, simply expressed in different basis - the above conveys unsettling results, namely that talking about properties of matter is generally meaningless. What we call a property is instead an incompletely defined potentiality realized in more definite form only via interaction with other systems, such as a measuring apparatus [4]. In the words of N. Bohr:

> "It is wrong to think that the task of physics is to find out how Nature is. Physics concerns what we say about Nature"

Such an interpretation came much to everyone's despise, especially Schrödinger's, because not only did his equation failed at getting rid of the discreteness of quantum mechanics but also turned out to yield probabilistic outcomes. This, gave rise to a furry of philosophical debates regarding the nature of reality and the role of physics, very much unsettled to this day [6]. Einstein and Schrödinger stood in firm opposition to Bohr's epistemological claims, especially the absurdity of including an observer in a physical theory, insisting quantum mechanics in such form cannot be considered complete [9]. Although certain strides towards ontological theory of quanta have been made, most notably by D. Bohm [7], they will not be covered within this thesis. So won't the further developments of Quantum Mechanics sometimes referred to as *second quantization*. The group theoretic language thereof, the inclusion of relativistic effects or quantization of the electromagnetic field, developed to a large extent by P. Dirac himself [10], are not of crucial importance for development of the electronic structure methods covered in the following chapters.

The pragmatic success of the wave equation formalism left philosophical unease behind, and with an exception of spin introduced by W. Pauli in 1929, the theory of quantum mechanics took its final form. It not only yields correct description of the seemingly discrete, quantum behavior occurring at length scales on the order of the Planck's constant $h \approx 6.63 \times 10^{-34} [\frac{m^2 kg}{s}]$, but also remains consistent in classical limit, a feature known as the *correspondence principle* [4].

Let us now turn to the many-body Schrödinger equation, which is considered relevant for the description of virtually all of chemistry and material science and per se plays central

[1] which is equivalent to the L2 norm of Ψ expressed in eigenbasis of $\hat{O}$

role in all that is to come in the remainder of this thesis. First and foremost, consider that the "waves" of the wave equation live in an abstract *phase space* - or more technically, a submanifold of the cotangent bundle of the configuration manifold the dimensionality of which is determined by the number of degrees of freedom present in the system. Contrary to classical mechanics, quantum systems are characterized by the non-commutativity of certain operators, like the momentum and position, and therefore, a complete description of the wave function requires only the mutually commutative subset of the phase space. Such formulation however, in which each significant co-ordinate is assigned a unique dimension, is responsible for so-called *curse of dimensionality* and lies at the heart of the practical difficulties in applying quantum mechanics to many-body systems.

Furthermore, before formulating the many-body Schrödinger equation it is important to realize the following, in a system of many identical particles we have no means to practically distinguish which particle we are talking about. Hence, the joint probability distribution of finding any particle in a given place must be invariant with respect to permutation $\mathcal{P}$ of its arguments. Considering a minimal example of two identical particles we have:

$$Pr(\mathbf{r}_1, \mathbf{r}_2) = Pr(\mathbf{r}_2, \mathbf{r}_1) = \mathcal{P}Pr(\mathbf{r}_1, \mathbf{r}_2) \tag{1.5}$$

It is obvious that swapping any two arguments twice yields the initial function back, it is therefore convenient to think of $\mathcal{P}^2$ as of identity operator. Now, recalling that the probabilistic interpretation of the wave function involves taking the square of the modulus, and since $\mathcal{P}$ is commutative with that operation, the above invariance condition leaves us with two cases to consider:

$$\mathcal{P}\Psi(\mathbf{r}_1, \mathbf{r}_2) = \pm\Psi(\mathbf{r}_1, \mathbf{r}_2) \tag{1.6}$$

A simple separable wave function $\Psi(\mathbf{r}_1, \mathbf{r}_2) = \phi_1(\mathbf{r}_1)\phi_2(\mathbf{r}_2)$ clearly does not fulfill it, but an (anti)symmetrized product of one-electron wave functions does:

$$\Psi(\mathbf{r}_1, \mathbf{r}_2) = \frac{1}{\sqrt{2}}\left(\phi_a(\mathbf{r}_1)\phi_b(\mathbf{r}_2) \mp \phi_a(\mathbf{r}_2)\phi_b(\mathbf{r}_1)\right) \tag{1.7}$$

and whether it is symmetric (+) or antisymmetric (-) has profound physical implications. On the one hand we have to do with *bosons*, which tend to group together - the probability of finding two bosons at the same position $\mathbf{r}$ is twice the probability of finding them spread apart. Antisymmetric wave functions on the other hand corresponds to *fermions* which behave in an exactly opposite way:

$$\Psi(\mathbf{r}, \mathbf{r}) = \frac{1}{\sqrt{2}}\left(\phi_a(\mathbf{r})\phi_b(\mathbf{r}) - \phi_a(\mathbf{r})\phi_b(\mathbf{r})\right) = 0 \quad \rightarrow \quad |\Psi(\mathbf{r}, \mathbf{r})|^2 = 0 \tag{1.8}$$

whereby for simplicity of this particular argument we have ignored the spin eigenstates σ. A rigorous treatment of the connection between spin and antisymmetry is only possible within the framework of quantum field theory. There, the so-called *spin statistics theorem* states that the particles of integer spin - bosons - are described by symmetric wave functions and obey Bose-Einstein statistics, on the contrary, particles with half integer spin - fermions - are described by antisymmetric wave functions and obey Fermi-Dirac statistics [11].

Electrons, are 1/2-spin particles and are therefore described by wave functions which are antisymmetric with respect to the interchange of any two particles. This, leads to yet another computational issue, the usual treatment of which, although expensive, is to represent the many body wave function in terms of a so called *Slater determinant*, we shall deffer a detailed discussion to section 2.1.1, where we introduce the Hartree-Fock method.

Finally, the last challenge from the point of view of physics of the many-body Schrödinger equation is to define the energy operator - the Hamiltonian - which determines what kind phenomena can be modelled provided a perfect solution could be obtained. Considering systems of interacting atoms - either ordered in a lattice or bounded into molecules - there is a number of sophisticated Hamiltonians that can be formulated for the study of various properties, from electromagnetic to superconductivity, usually however, one only focuses on the Coulomb interactions due to electric charge of both the nuclei and the surrounding them electrons [12]:

$$\hat{H} = -\underbrace{\sum_i \frac{\hbar^2}{2m_e}\nabla_i^2}_{E_k\ electrons} - \underbrace{\sum_I \frac{\hbar}{2M}\nabla_I^2}_{E_k\ nuclei} + \underbrace{\frac{1}{2}\sum_{i\neq j}\frac{e^2}{4\pi\epsilon_0|\mathbf{r_i}-\mathbf{r_j}|}}_{E_p\ el-el} \underbrace{-\sum_{i,I}\frac{e^2 Z_I}{4\pi\epsilon_0|\mathbf{r_i}-\mathbf{R_I}|}}_{E_p\ nuc-el} + \underbrace{\frac{1}{2}\sum_{I\neq J}\frac{e^2 Z_I Z_J}{4\pi\epsilon_0|\mathbf{R_I}-\mathbf{R_J}|}}_{E_p\ nuc-nuc} \tag{1.9}$$

where i, j and I, J are the electron and nuclei indices respectively, Z is the atomic number, $e \approx -1.67 \cdot 10^{-19}C$ is the charge of an electron and the $\frac{1}{2}$ factor in front of the 3^{rd} and 5^{th} term is there in order not to count electron interactions twice.

Furthermore, for the study of the ground state properties of mater we consider only the time-independent Schrödinger equation which is essentially nothing but a nonlinear, energy eigenvalue problem:

$$\hat{H}\Psi(\mathbf{r}_1, ..., \mathbf{r}_N, \mathbf{R}_1, ..., \mathbf{R}_M) = E\Psi(\mathbf{r}_1, ..., \mathbf{r}_N, \mathbf{R}_1, ..., \mathbf{R}_M) \tag{1.10}$$

It might seem like a very rigid formulation, but it has sufficient complexity to provide a faithful model for i.a. establishing equilibrium compositions, studying band structures or strength properties of crystalline solids as well as chemical bounding mechanisms and therefore formation of molecules. The accuracy with which we are able to capture these phenomena, however, resides ultimately in the scope of the treatment of mutual interaction between electrons, which due to Born's rule is often also refered to as the *electronic correlation* [13, 14]. Below we list some of the most common approximations which will define utility of the methods we shall present shortly in the chapter to follow:

- Born-Oppenheimer approximation (adiabatic approximation) - Nuclei are large, heavy and almost static compared to electrons, due to this discrepancy in scales we can decouple the Schrödinger equation into two terms:

  $$\Psi(\mathbf{r}_1, ..., \mathbf{r}_N, \mathbf{R}_1, ..., \mathbf{R}_M) = \Psi_e(\mathbf{r}_1, ..., \mathbf{r}_N|\mathbf{R})\Psi_{ion}(\mathbf{R}_1, ..., \mathbf{R}_M) \tag{1.11}$$

  The rationale is that since the electrons are so nimble, they always have enough time to readjust their lowest energy state to any movement of atomic nuclei. The mutual evolution of nuclei and electrons is hence adiabatic, the energy exchange between them is negligible. Provided we consider dynamics at reasonably cold temperatures it is a reasonable assumption, otherwise we need to resort to *ab-initio* molecular dynamics [15]. The conditioning on $\mathbf{R}$ indicates that the electron wave function still depends parametricaly on the positions of ions.

- Clamped nuclei approximation - We can further diminish the role of nuclei by ignoring their kinetic energy - again provided we consider sufficiently cold systems. Since the potential energy due to nuclei-nuclei interactions becomes a constant now, we move it over to the other side, for brevity, since electrons are indistinguishable, we also replace $\Psi(\mathbf{r_1}, ..., \mathbf{r_N})$ with $\Psi(\mathbf{r})$:

$$\Bigg[\underbrace{-\sum_i \frac{\hbar^2}{2m_e}\nabla_i^2}_{E_k\ electrons} + \underbrace{\frac{1}{2}\sum_{i\neq j}\frac{e^2}{4\pi\epsilon_0|\mathbf{r}_i - \mathbf{r}_j|}}_{E_p\ el-el} \underbrace{-\sum_{i,I}\frac{e^2 Z_I}{4\pi\epsilon_0|\mathbf{r}_i - \mathbf{R}_I|}}_{E_p\ nuc-el}\Bigg]\Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \tag{1.12}$$

  - Only valence electrons - Since the valence electrons have the largest impact on the reactivity of atoms via covalent bonding [16], one often incorporates the electrons occupying the inner shells into the nuclei, creating so called *pseudopotentials*. This way the number of degrees of freedom is reduced to just the electrons on the valence orbitals.

- Independent electrons approximation - It is possibly the harshest approximation and if used, will require introduction of certain corrections to reintroduce the electron interactions responsible for most of the interesting behavior of matter. Nonetheless, it is an efficient cure for the curse of dimensionality for it reduces the computational effort from solving a $3N$ dimensional problem to solving N, decoupled, three-dimensional ones. It treats the many-electron wave function as a product of one-electron wave functions and therefore makes it separable:

$$\Psi(\mathbf{r}_1, ..., \mathbf{r}_N) = \phi_1(\mathbf{r}) \cdots \phi_N(\mathbf{r}) \tag{1.13}$$

  and the many body Schrödinger equation simplifies even further to:

$$\Bigg[\underbrace{-\sum_i \frac{\hbar^2}{2m_e}\nabla_i^2}_{E_k\ electrons} \underbrace{-\sum_{i,I}\frac{e^2 Z_I}{4\pi\epsilon_0|\mathbf{r}_i - \mathbf{R}_I|}}_{E_p\ nuc-el}\Bigg]\Psi(\mathbf{r}) = E\Psi(\mathbf{r}) \tag{1.14}$$

  From the point of view of probability theory, this approximation makes the joint, Born probability density completely independent.

Lastly, let us touch upon the topic of electronic correlations more elaborately, since their faithful resolution is indeed what we aim to target with our improved sampling algorithms. Physically, there are only two mechanisms through which electrons get correlated: the correlation due to Fermi-Dirac statistics and due to Coulomb repulsion [14], both manifest themselves marvellously through mathematical peculiarities of the Schrodinger's equation. On the one hand, the former, known as *Fermi correlation*, arises solely from the antisymmetry requirements and can be readily seen by computing the Born probability density of the simplest, valid wave function from eq. 1.7, upon doing so it's apparent it must contain a mixed term $\pm\phi_a(\mathbf{r}_1)\phi_b(\mathbf{r}_2)\phi_a(\mathbf{r}_2)\phi_b(\mathbf{r}_1)$ and therefore the joint probability density will not be independent. Importantly this has nothing to do with the charge of the electrons, but rather the fact they are *identical* fermions. Because, electrons are thought to occupy so called *spin orbitals*, that is a spatial function $\phi(\mathbf{r})$ with an attached spin indicator σ, in order to comply with the antisymmetry requirements, symmetric spatial configurations may be only combined with antisymmetric spin factors, and vice versa. Depending on the spin therefore, this will lead to either so called *Fermi holes* - where the spatial probability density decreases to zero when the electrons approach each other - or conversely, *Fermi heaps* - a bosonic-like behaviour where the probability density actually doubles. On the other hand, second mechanism of correlation - the *Coulomb correlation* - is a result of singularity of our Hamiltonian, eq. 1.9, whenever two charged bodies approach each other, that is, whenever the denominators in the potential energy function: $|\mathbf{r_i} - \mathbf{r_j}|$ or $|\mathbf{r_i} - \mathbf{R_I}|$ go to zero. Unless the wave function at such a spot equals zero itself [2], the only possibility for the eigenvalue problem 1.10 to remain defined, is to have an "equal" and opposite infinity in the kinetic energy term. Since, the kinetic energy operator is a Laplace operator, infinity can occur only if the first derivative of the wave function is discontinuous - we call such places *cusps*. Practically, cusps will arise asymptotically, through considerations of energy minimization when the function is expanded as a superposition of all possible *virtual excitations* which is often referred to as *configuration interaction*. Inclusion of other symmetrized orbital products of course also refines the *nodal structure* of the wave function [3], a faithful resolution of which, is particularly important for accurate representation fermionic systems [13, 17], but as discussed, it can only occur approximately. Finally, for historical reasons one sometimes distinguishes between *static* and *dynamic* Coulomb correlations although physically they are not any different. What we have just described is essentially the case of dynamic correlation, whereas static correlation occur whenever there are degeneracies among orbitals, resulting in configurations of very similar energies but different orbital occupations. In mathematical terms, it corresponds to geometric multiplicity of the energy eigenvalues and requires specialized, multi-configuration methods to deal with.

[2] which is the case for antisymmetric spatial orbital i.e. the case of a Fermi hole, which will, however, nonetheless lead to a discontinuity in the wave function, just in its second, not first, derivative

[3] A node of a wave function is generally nothing else than a node of any other wave, a place where $\Psi(\mathbf{x}) = 0$ but $\Psi'(\mathbf{x}) \neq 0$

2. State of the art

This review aims to present the current, although not-comprehensive, landscape of first-principle simulations for prediction and design of material properties - so called *electronic structure methods*. The major challenge is to find approximations to the exponentially hard, multi-body Schrödinger equation 1.10 that, in the best case would have only polynomial scaling, yet could capture the most important correlations between electrons. In the words of P. W. Anderson, *more is different* [18]; scale therefore, somewhat ironically, is the major roadblock as well as the key to understanding the *emergent* properties of solids [1], chemical compounds [19] and therefore virtually all phenomena of interest in molecular biology, material science or nanotechnology [20].

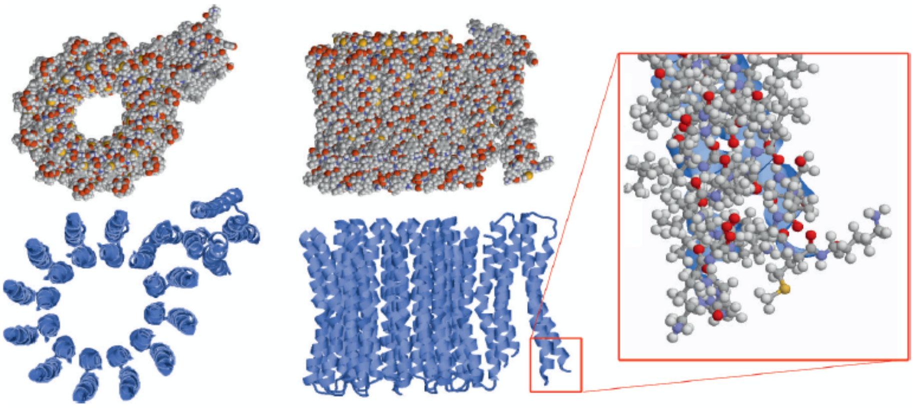

Figure 2.1.: Rotor of a ATP synthase, known to consist of 16649 atoms. Source: [12]

In the first part, the well established and widely used self-consistent-scheme methods will be presented as well as the variational approach of Quantum Monte Carlo, the latter actually builds on the developments of the former and in turn defines theoretical grounds for what is to come in the second part, namely the recent approaches utilizing deep learning. Only a relatively brief account of utility of all these methods will be presented and we shall focus primarily on computational issues leaving the applicability domains - which are indeed broad - aside. An inquisitive reader is advised to consult the various reviews [21–23] and further resources cited in the sections to follow.

2.1. Classical electronic structure methods

A common example given when introducing electronic structure methods are the two approaches to the study of molecular hydrogen H_2. The prior, by Heitler and London, treated the two electrons of the system as being highly correlated by explicitly *excluding* the ionic configurations in their ansatz i.e. those with both electrons on the same hydrogen:

$$\Psi_{HL} = \frac{1}{\sqrt{2}} \left(\phi_{1,\uparrow}(\mathbf{r}_1)\phi_{2,\downarrow}(\mathbf{r}_2) + \phi_{2,\uparrow}(\mathbf{r}_2)\phi_{1,\downarrow}(\mathbf{r}_1)\right) \tag{2.1}$$

with $\phi(\mathbf{r})$ taken as the $1s$ orbital function. Hartree, Fock and Slater, gave an alternative approach in which the electrons are treated as independent and the ansatz is taken as the product of linear combination of spin up and spin down atomic orbitals:

$$\Psi_{HF} = \frac{1}{\sqrt{2}} \left(\phi_{1,\downarrow}(\mathbf{r}_1) + \phi_{2,\downarrow}(\mathbf{r}_2)\right) \frac{1}{\sqrt{2}} \left(\phi_{1,\uparrow}(\mathbf{r}_1) + \phi_{2,\uparrow}(\mathbf{r}_2)\right) \tag{2.2}$$

Since Ψ_{HL} involves only non-ionic configurations it becomes accurate as the distance between both atoms diverges to infinity, that is not the case for the Hartree-Fock ansatz which includes both ionic and non-ionic configurations and it becomes accurate when the electron electron repulsion is ignored [24]. In conclusion, the exact result must lie somewhere in between these two extremes, it turns out however, that many chemically bounded systems are weakly bounded and so the latter is a good starting point [21].

2.1.1. Hartee-Fock and post Hartee-Fock methods

The most important conceptual features of the Hartree-Fock (HF) theory are perhaps the introduction of the mean field approximation and an iterative scheme to obtain the lowest energy, single-particle wave functions subsequently combined using the formalism of Slater determinant. As discussed before a general separable wave function does not obey the antisymmetry requirement for fermionic systems. In mathematical terms, a correct ansatz for a system with N *valence* electrons is obtained with the *exterior product* $\wedge$ of singe-particle wave functions $\phi_i(\mathbf{r})$ [25]. Using determinant definition of $\wedge$ [26] it reads:

$$\Psi_{HF}(\mathbf{r}) = \phi_1 \wedge \phi_2 \wedge \cdots \wedge \phi_N(\mathbf{r}_1, \mathbf{r}_2, ..., \mathbf{r}_N) := \det(\Phi), \quad where \quad \Phi_{ij} = \phi_i(\mathbf{r}_j) \tag{2.3}$$

and it is also known under the name of Slater determinant.

For the derivation of Hartree-Fock method we start our discussion at the stationary, many body Schrödinger equation 1.10 with the Born-Oppenheimer and clamped nuclei approximations as in eq. 1.12. Furthermore we consider the problem of obtaining just the lowest energy eigenstate of $\hat{H}$ and use the Slater ansatz outlined above. The main challenge remains in establishing the actual functional form of the single-particle wave functions $\phi(\mathbf{r})$ and the answer lies in so called *variational principle* which plays a central role in all electronic structure approaches outlined in this and following sections. The key insight is that Ψ_{HF} is not any wave function but one corresponding to the minimal energy state, therefore we can perform an explicit variational minimization of energy as a functional of Ψ_{HF} [12]:

$$\begin{gathered} \frac{\delta E}{\delta \phi_i} = 0 \quad where \quad E = \frac{\langle \Psi_{HF} | \hat{H} | \Psi_{HF} \rangle}{\langle \Psi_{HF} | \Psi_{HF} \rangle} \\ \int \phi_i^*(\mathbf{r})\phi_j(\mathbf{r})d\mathbf{r} = \delta_i^j \end{gathered} \tag{2.4}$$

where the definition of energy, expressed using Dirac notation, follows exactly from the stationary Schrödinger equation and the orthonormality constraint between single-particle wave functions is introduced for pragmatic reasons.

Substituting 2.3 into 2.4 and performing the variational minimization using the *Euler-Lagrange* formula and the formalism of Lagrange multipliers to enforce orthonormality [3] leads to [12]:

$$\left[-\frac{\hbar^2}{2m_e}\nabla^2 + U_H(\mathbf{r}) + U_{ion}(\mathbf{r})\right]\psi_i(\mathbf{r}) + \int U_X(\mathbf{r},\mathbf{r}')\psi_i(\mathbf{r}')d\mathbf{r}' = \varepsilon_i\psi_i(\mathbf{r}) \qquad i = 1,...,N \tag{2.5}$$

which concludes the variational minimization problem.

Notice, the single-particle wave functions have been renamed to $\psi_i(\mathbf{r})$, that is due to *diagonalization* of the system of non-linear equations in Lagrange multipliers λ_{ij} which we obtain from the above minimization procedure. It is a crucial and computationally demanding step, which needs to be performed numerically every iteration of so called *self-consistent scheme*. The energy eigenvalues ε_i as well as $\psi_i(\mathbf{r})$ of the new of the decoupled system are obtained from:

$$\mathbf{S}\begin{bmatrix}\lambda_{11} & \cdots & \lambda_{1N}\\ \vdots & & \vdots\\ \lambda_{N1} & \cdots & \lambda_{NN}\end{bmatrix}\mathbf{S}^{-1} = \begin{bmatrix}\varepsilon_1 & 0 & \cdots & 0\\ 0 & \varepsilon_2 & & \\ \vdots & & \ddots & \vdots\\ 0 & & \cdots & \varepsilon_N\end{bmatrix}$$
$$\psi_i = \sum_j S_{ij}\phi_j \tag{2.6}$$

and since $\hat{H}$ is hermitian the matrix of Lagrange multipliers also is, and according to spectral theorem, it has orthonormal eigenvectors $\mathbf{S}\mathbf{S}^T = \mathbf{I}$ and real eigenvalues ε_i. The total electron energy is simply given as the sum:

$$E = \sum_i^N \varepsilon_i \tag{2.7}$$

The new potential energy terms introduced in the process, $U_H(\mathbf{r})$ and $U_X(\mathbf{r},\mathbf{r}')$, define the "Hartree" and "exchange" potential energies respectively and are a direct result of the electron-electron Coulomb interactions and the antisymmetric ansatz used. The U_{ion} term is simply a restatement of the nuclei-electron interaction already present in eq. 1.12 but here as mentioned earlier we include also the inner shell electrons.

$$U_{ion}(\mathbf{r}) = -\sum_I \frac{e^2 Z_I}{4\pi\epsilon_0|\mathbf{r}-\mathbf{R_I}|} \qquad U_X(\mathbf{r},\mathbf{r}') = -\frac{e^2}{4\pi\epsilon_0}\sum_j \frac{\psi_i^*(\mathbf{r}')\psi_j(\mathbf{r})}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}'$$
$$U_H(\mathbf{r}) = \frac{e^2}{4\pi\epsilon_0}\int\frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' \qquad n(\mathbf{r}) = \sum_j |\phi_j(\mathbf{r})|^2 \qquad \int n(\mathbf{r})d\mathbf{r} = N \tag{2.8}$$

It has been mentioned in the introduction that the independent electron approximation is too drastic, here we see one possible remedy, the Hartree potential, introducing what is known as the *mean field approximation*. It defines *electron density* $n(\mathbf{r})$ - a probability density function of finding any electron - treated classically - at $\mathbf{r}$ and therefore simplifies the electrons' potential energy from all-to-all to a potential energy of *an* electron submerged in a "charge cloud". The crucial aspect is that we maintain the computational advantages of the independent electron approximation but at the same time include at least rudimentary interaction between electrons. Somewhat problematic, however, is that circular dependence of $n(\mathbf{r})$ on $\psi_j(\mathbf{r})$, this leads to a fixed-point like iterative scheme - the *self-consistent scheme* - which will be covered soon, in section 2.1.2.

Post Hartree-Fock methods

The single determinant ansatz Ψ_{HF} of the HF theory gave rise to a mean field approximation in which only the average effect of all electrons is considered. Although this simplification reduces the computational effort from exponential to polynomial, it completely disregards the electronic correlations, which although usually amount to just a small fraction of total energy E are crucial in explaining chemical bounding or properties of metals [21].

The unifying idea behind all post HF methods is the usage of more Slater determinants, or more precisely, a linear combination thereof. Previously, the Slater determinant $\det(\Phi)$ for a system of N valence electrons has been introduced as an exterior product of N single-particle wave functions $\phi_i(\mathbf{r})$, one can therefore think of it as of a differential N-form on a M-dimensional manifold $\mathcal{M}$, whereby for HF Slater determinant, $M = N$. In general, however, nothing restricts the number of wave-functions $\phi_i(\mathbf{r})$ used, in the language of differential geometry, we can increase M and obtain a generic expression for such a differential form by [26]:

$$\Psi(\mathbf{r_1}, ..., \mathbf{r_N}) = \sum_I \alpha_I \phi^I \tag{2.9}$$

where ϕ^I correspond to the *basis N-forms*, which together span the space of all differential N-forms on $\mathcal{M}$ denoted $\bigwedge_N T^*\mathcal{M}$, or equivalently, the space of all antisymmetric $(0, N)$-tensors denoted $\Omega(\mathcal{M})$. The capital index I is a multi-index running over all combinations $\binom{M}{N}$:

$$\phi^I := \phi^{I_1} \wedge \cdots \wedge \phi^{I_N} \tag{2.10}$$

s.t. the indices I_i are the elements of the I-th combination ordered increasingly.

Bringing the discussion back to electronic structure, equation 2.9 defines the ansatz of so called *configuration interaction* family of methods and in particular, if all possible configurations are exhausted - so called *full configuration interaction* (FCI) - it is capable of exactly capturing all electronic correlations. It should be clear though, that this approach has a rather unfavourable scaling, first of all, because the binomial coefficient scales approximately exponentially with M, secondly, because every determinant itself is an effort on the order of $\mathcal{O}(N^3)$, and most importantly, the achieved increase in the correlation energy does not scale consistently with M but more like $\mathcal{O}(\sqrt{M})$ [21]. Nonetheless, the above approach allows for theoretically exact solution of the many-electron Schrödinger equation 1.10 and in a *multi-reference configuration interaction* variant establishes the state-of-the-art baseline, especially for quantum chemical applications [27]. For more details, as well as the omitted, but related *coupled cluster* expansions, refer to [19] or [28].

2.1.2. Density Functional Theory

Density Functional Theory (DFT) is another, and perhaps the most popular electronic structure method owing to its flexibility and its formal exactness. The core contribution of the method is the result of the so called Hohenberg-Kohn theorem which states that if E is the ground state energy of a system, then E is a functional of electron density $n(\mathbf{r})$ *only* [12]. The repercussions of this statement are profound, it implies one can obtain many ground state *properties* of a many-body system by considering an auxiliary, non-interacting one with the same electron density as the real system. On a flip side however, we do not obtain the corresponding ground state *wave function*, we can only use the single-electron orbitals $\psi_i(\mathbf{r})$ which correspond to the auxiliary system and define the density $n(\mathbf{r})$. Nonetheless, these orbitals are sometimes used as the basis for the Quantum Monte Carlo method covered shortly, but by themselves are not particularly useful. The Hohenberg-Kohn theorem is remarkable also because, as presented in eq. 2.4, total energy is in general a functional of $\Psi(\mathbf{r_1}, ..., \mathbf{r_N})$ which is a function of $3N$ variables whereas the electron density is a function of merely 3 spatial coordinates.

The Hohenberg-Kohn energy functional reads:

$$E[n(\mathbf{r})] = \underbrace{e\int n(\mathbf{r})U_{ion}(\mathbf{r})d\mathbf{r}}_{E_p \text{ el-nuc}} \underbrace{-\frac{\hbar^2}{2m_e}\sum_i^N \int \phi_i^*(\mathbf{r})\nabla^2\phi_i(\mathbf{r})d\mathbf{r}}_{E_k} + \underbrace{\frac{e^2}{8\pi\epsilon_0}\int\int \frac{n(\mathbf{r})n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}d\mathbf{r}'}_{\text{Hartree potential energy}} + E_{XC}[n(\mathbf{r})] \tag{2.11}$$

which to a large extend is analogous to the Hartree-Fock functional with an exception that it is defined almost explicitly in terms of $n(\mathbf{r})$, but also, it unifies all the problematic, non-local exchange energy $U_X(\mathbf{r}, \mathbf{r}')$ and the additional correlation energy $U_C(\mathbf{r})$ into one *exchange-correlation energy functional* $E_{XC}[n(\mathbf{r})]$. The accuracy with which this functional can be approximated defines eventually the accuracy of the DFT itself.

Using the variational principle again, but this time in electron density:

$$\frac{\delta E[n(\mathbf{r})]}{\delta n(\mathbf{r})} \overset{!}{=} 0 \tag{2.12}$$

and minimizing, leads to the famous *Kohn-Sham equations* [12]:

$$\begin{gathered}\overbrace{\left[-\frac{\hbar^2}{2m_e}\nabla^2 + U_H(\mathbf{r}) + U_{ion}(\mathbf{r}) + U_{xc}(\mathbf{r})\right]}^{\text{"Kohn-Sham Hamiltonian" } \hat{H}_{KS}} \psi_i(\mathbf{r}) = \varepsilon_i\psi_i(\mathbf{r}) \qquad \forall i \\ U_H(\mathbf{r}) = \frac{e^2}{4\pi\epsilon_0}\int \frac{n(\mathbf{r}')}{|\mathbf{r}-\mathbf{r}'|}d\mathbf{r}' \qquad U_{xc}(\mathbf{r}) = \frac{\delta E_{XC}[n(\mathbf{r})]}{\delta n(\mathbf{r})}\end{gathered} \tag{2.13}$$

for which, all of the considerations regarding diagonalization and fixed-point iteration as in the case of the HF method also apply, in particular, let us now investigate the later in more details.

Self-consistent scheme

Although direct optimization approaches for optimization of the DFT or HF single-electron wave functions are conceivable by constraining the updates of $\phi_i(\mathbf{r})$ to the Stiefel manifold [29], we shall focus on the usual approach of fixed point iteration here. Moreover without loss of generality we cover explicitly only the DFT iteration scheme for the sake of brevity.

The starting point is always the atomic structure, which needs to be known apriori either from modelling, X-ray crystallography or simulation - it sets up the potential due to electron-nuclei interaction U_{ion}. Secondly, we need an initial guess for electron density $n(\mathbf{r})^{(0)}$ to compute U_H and U_{xc} and start the iteration. A good initial guess could be a sum of electron densities around each nuclei as if completely isolated from one another [12]. Having calculated the entire Kohn-Sham Hamiltonian we can then solve the resulting *eigenvalue problems* and obtain first approximation for $\psi_i^{(1)}$, which we use to calculate new electron density $n(\mathbf{r})^{(1)}$ and the entire process repeats. A pseudo algorithm is outlined in Figure 2.2. The moment the eigensolutions ψ_i we have used to obtain $n(\mathbf{r})$ and set up the Kohn-Sham equations coincide with the ones that we get as the solution of these equations, we have reached convergence or self-consistency and the iteration terminates.

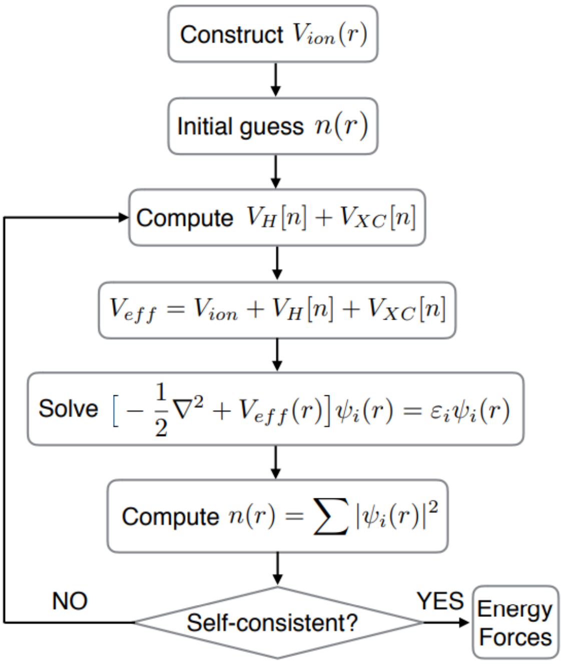
Construct $V_{ion}(r)$
Initial guess $n(r)$
Compute $V_H[n] + V_{XC}[n]$
$V_{eff} = V_{ion} + V_H[n] + V_{XC}[n]$
Solve $\left[-\frac{1}{2}\nabla^2 + V_{eff}(r)\right]\psi_i(r) = \varepsilon_i\psi_i(r)$
Compute $n(r) = \sum|\psi_i(r)|^2$
NO
Self-consistent?
YES
Energy
Forces

Figure 2.2.: Self consistent DFT algorithm. Source: [30]

Chosing the basis

To solve the Kohn-Scham equations we need to settle for certain functional representation of ϕ_i. In periodic systems, abundant in solid-state physics, this is usually means a *plane* *wave expansion*, for molecular systems, of interest in quantum chemistry, we wish to provide more freedom to each individual electron and therefore resort to an atom centered basis like the *Slater type orbitals* (STO) or the *Gaussian type orbitals* (GTO). Surely, one can also resort to traditional mesh discretization within some computational cell Ω_i with either periodic or "infinite-well" boundary conditions and resort to finite difference approximation of the Hamiltonian, but this choice is a rarity [12]. Nonetheless, what we are left with, is a generalized eigenvalue problem in the coefficients of the basis expansion, and if we consider for example the case of molecular systems, this yields so called *Roothan equations*:

$$\mathbf{H}\mathbf{c}_m = \varepsilon_m \mathbf{S}\mathbf{c}_m \tag{2.14}$$

$$H_{ij} = \langle \phi_i | \hat{H} | \phi_j \rangle \qquad S_{ij} = \langle \phi_i | \phi_j \rangle \tag{2.15}$$

$$\langle \psi_m | = c_m^i \langle \phi_i | \tag{2.16}$$

These equations need to be solved, every iteration of the self-consistent scheme, but fortunately, the Hamiltonian matrix **H** is usually endowed with either a band or an orthogonal structure, so it can be dealt with efficiently.

Exchange-correlation potentials

In the language of probability theory, the mean field approximation decorrelates the electrons forming the electron cloud. In reality, however, as we have discussed in the introduction, electrons do interact and they correlate their movement via two mechanisms: the Coulomb repulsion and the Pauli's exclusion principle. Both of these terms are thought to be encapsulated in the exchange-correlation functional, whereby the non-local interactions are further simplified so that we have to deal with just the electron cloud of one argument.

A standard approach is the so called *Local density approximation* (LDA) which assumes a suitable approximation can be obtained from a volumetric integral of the exchange-correlation energy per electron $\epsilon_{xc}^{hom}(n)$ which is in turn established from calculations of an *uniform electron gas* enclosed in some volume V such that its density is n [12]:

$$E_{xc}^{LDA}[n(\mathbf{r})] = \int \epsilon_{xc}^{hom}(n) n(\mathbf{r}) d\mathbf{r} \tag{2.17}$$

If that is not sufficient a *Generalized gradient approximation* (GGA) is used which additionally takes into account the gradient information of $\nabla n(\mathbf{r})$. More sophisticated approximations also exist [31] e.g. including higher derivatives of the electron density, non-local effects or even deep learned exchange-correlation functionals utilizing a fully differentiable self-consistent schemes [32]. Nonetheless, even the best exchange-correlation functional yield usually orders of magnitude lower accuracy than good Quantum Monte Carlo results [21], which shall be covered next.

2.1.3. Quantum Monte Carlo

Quantum Monte Carlo (QMC) is a family of Monte Carlo methods capable of approximating the lowest energy eigenstate of a quantum system through the process of variational minimization. There are many advanced and specialised methods within the QMC framework i.a. the auxiliary-field or path-integral QMC for model Hamiltonians and bosonic systems respectively; there is also the Diffusion Monte Carlo algorithm generating trajectories which sample the ground state wave function without the need for its explicit representation [17]. In this section, however, we will focus on the baseline Variational Monte Carlo (VMC) approach which is restricted to zero-temperature simulations only, but which provides a good template for a review of major components of every Quantum Monte Carlo approach. It is the starting point and the unifying framework for all further approaches presented in this chapter, as well as for the body of this thesis, in which we will consider its unification with machine learning, it is therefore instrumental to understand it thoroughly.

Compared to the two previous electronic structure methods, QMC is the most accurate one, capable of achieving *chemical accuracy* usually defined as within 1 kcal per mole error in total energy, or about 0.04 eV per molecule. In fact, it is even used as reference for designing DFT exchange-correlation functionals [21]. To better comprehend QMC, it should be realized it is much more of a correction on top of mean-field approaches, rather than an electronic structure method on its own - especially for continuum models - and to fully appreciate how meticulous an endeavour it is, one should bear in mind that for most systems, the uncorrelated energy contribution amounts already for well over 90 % of total energy [21]. Nonetheless, these few percent account for virtually all electronic phenomena and hold the key not only to addressing interatomic forces and chemical reactions, but further down the line, phenomena such as superconductivity too.

In a nutshell, VMC is concerned with establishing parameters of a trial wave function Ψ_θ s.t. the expectation value of the energy operator $\hat{H}$ is minimized:

$$\theta^{opt} = \arg\min_\theta < \hat{H} >= \arg\min_\theta \frac{\int \Psi_\theta^* \hat{H} \Psi_\theta d\mathbf{r}}{\int \Psi_\theta^* \Psi_\theta d\mathbf{r}} = \arg\min_\theta \int \rho(\mathbf{r}) E_{loc}(\mathbf{r}) d\mathbf{r} \tag{2.18}$$

where $\rho(\mathbf{r})$ is the normalized $|\Psi_\theta(\mathbf{r})|^2$ and $E_{loc}(\mathbf{r}) = \Psi_\theta^{-1} \hat{H} \Psi_\theta$ is the *local energy* [21]

Physically, equation 2.18 expresses the same *variational principle* presented when considering the HF theory, eq. 2.4, a key distinction, however, is that in the case of Quantum Monte Carlo the expectations are evaluated numerically via Monte Carlo sampling. To confirm such minimization would indeed yield the ground state wave function Ψ_0 consider an energy eigendecomposition of Ψ_θ with the energy eigenvalues ordered increasingly i.e. $E_0 < E_1 < ... < E_n$, then:

$$\Psi_\theta(\mathbf{r}) = \sum_n a_n \Psi_n(\mathbf{r}) \quad \rightarrow \quad < \hat{H} >= \frac{\sum_{n,m} a_n a_m \int \Psi_n^* \hat{H} \Psi_m d\mathbf{r}}{\sum_{n,m} a_n a_m \int \Psi_n^* \Psi_m d\mathbf{r}} = \frac{\sum_n a_n^2 E_n}{\sum_n a_n^2} \geq E_0 \tag{2.19}$$

hence if $a_n = 0 \; \forall n \neq 0$ but $a_0 = 1$ corresponding to the case of $\Psi_\theta = \Psi_0$, the above expectation will be minimal and equal E_0.

The so called *zero variance principle* shall be mentioned at this point, the minimum possible value of the variance of energy σ_E^2 is zero and will be obtained if and only if the Ψ_θ is an energy eigenstate, in particular the ground state. Energy variance therefore, may be used as an alternative objective function for the minimization procedure and is in fact often favoured due to its stability [21]:

$$\sigma_E^2 = \frac{\langle\Psi_\theta|\,(E_{loc} - <\hat{H}>)^2\,|\Psi_\theta\rangle}{\langle\Psi_\theta|\Psi_\theta\rangle} \tag{2.20}$$

Lastly, it shall be clear that the VMC method in not merely suitable for energy ground state approximations. By restricting the the Hilbert space to a subspace, orthogonal to the ground state, the minimization will yield an eigenstate corresponding to the next smallest energy level - the first *excited state*.

Wave function ansatz

The crucial design decision in VMC is the choice of the variational ansatz which would account for antisymmetry and restrict the complete Hilbert space to a tractable subset, yet expressive enough to capture the most important electron correlations. A common choice is the so called *Slater-Jastrow ansatz* [33] - a linear combination of Slater determinants of one-electron wave functions, denoted Ψ_{MF}, with an additional multiplicative term to account for short range correlations [21]:

$$\Psi_\theta(\mathbf{x}_1, ..., \mathbf{x}_N) = \Psi_{MF}(\mathbf{x}_1, ..., \mathbf{x}_N)e^{J(\mathbf{x}_1, ..., \mathbf{x}_N)} \tag{2.21}$$

where $x_i = \{\mathbf{r}_i, \sigma_i\}$, that is, coordinates involving spin $\sigma_i \in \{\uparrow, \downarrow\}$ alongside the usual position in physical space $\mathbf{r}_i \in \mathbb{R}^3$.

For observables which are not directly dependent on spin, the mean-field part can be decomposed into a product of only spin up and only spin down determinants and still yield correct energy expectation, although the wave function becomes antisymmetric only under exchange of electrons with the same spin [21]:

$$\Psi_{MF}(\mathbf{x}_1, ..., \mathbf{x}_N) = \sum_{I=1} \alpha_I \phi_I^\uparrow(\mathbf{r}_1, ..., \mathbf{r}_{N'})\phi_I^\downarrow(\mathbf{r}_{N'+1}, ..., \mathbf{r}_N) \tag{2.22}$$

and ϕ_I are the basis Slater determinants defined as in eq. 2.10.

Similarly, the correlation term - defined in terms of so called *Jastrow factor* J - can be also written in terms of spatial coordinates only and usually takes the form:

$$J(\mathbf{r}_1, ..., \mathbf{r}_N) = \sum_{i=1}^{N} \chi(\mathbf{r}_i|\mathbf{R}) - \frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N} u(\mathbf{r}_i, \mathbf{r}_j) \tag{2.23}$$

with the first term modelling the electron-nuclear, and the latter for electron-electron correlations, both reducing the magnitude of the many-body wave function whenever two particles approach each other creating so called *cusps*. Introduction of just the two-body term u reduces the total energy but unfavourably distorts the charge density by diffusing electrons away from regions with high charge density, introduction of χ counteracts this mechanism and further reduces the energy thereby improving the quality of the optimum while preserving qualitative features of a many-body wave function. A common treatment, which allows for significantly faster convergence, is the *explicit* introduction of cusp conditions γ into $u(\mathbf{r}_i, \mathbf{r}_j)$ by expanding the set of arguments with the electronic distances $|\mathbf{r}_i - \mathbf{r}_j|$, sometimes referred to as the $r12$ modification [14]. On an example of an electron-electron cusp it yields:

$$\gamma(|\mathbf{r}_i - \mathbf{r}_j|) = \sum_{i<j} -\frac{c_{ij}}{1 + |\mathbf{r}_i - \mathbf{r}_j|} \tag{2.24}$$

where the coefficients $c_i j$ are either $\frac{1}{2}$ or $\frac{1}{4}$ depending on the spin of electrons i and j. Although instead of using fixed values, one could define c_{ij} as a variational parameter, practical results show that they very rarely converge to different quantities [21].

For more complicated wave functions, very successful, and particularly useful in fermionic systems, is the technique of *backflow* introduced by Feynman to deal with excitations in liquid helium H_4 [34, 35]. Although computationally expensive it can yield very accurate results and it does so by replacing the electron coordinates $\mathbf{r}_i$ in eq.2.21 with *quasi particle coordinates*:

$$\bar{\mathbf{r}}_i = \mathbf{r}_i + \sum_{j \neq i}^{N} (\mathbf{r}_i - \mathbf{r}_j)\eta(|\mathbf{r}_i - \mathbf{r}_j|) \tag{2.25}$$

with the function η optimized variationally.

Of course, many different possibilities are also conceivable, for example *geminal wave functions*, when used inside Slater determinant ansatz, have the advantage of being equivalent to multi-configuration wave function which turns out useful in studying nearly degenerate ground states [36]. Lately also highly flexible neural network ansaetze have been investigated, we will devote an entire next chapter to this topic, for now however let us proceed to the topic of evaluation of energy expectation through sampling.

Sampling

Although Quantum Monte Carlo owes its accuracy to the flexibility in the choice of the wave function ansatz, that freedom does however come at a price, namely, the integrals appearing in the expressions for energy expectation eq. 2.18 - alternatively its variance eq. 2.20 - are no longer analytically tractable. Approximating expectation values under any general, possibly high dimensional and unnormalized, probability density function (PDF) is in the first place dictated by the ability to generate accordingly distributed samples, then:

$$\mathbb{E}_\rho[E_{loc}] = \int \rho(\mathbf{r};\theta) E_{loc}(\mathbf{r};\theta) d\mathbf{r} \approx \frac{1}{N} \sum_{i}^{N} E_{loc}(\mathbf{r}_i;\theta), \qquad \mathbf{r}_i \sim \rho(\mathbf{r};\theta) \tag{2.26}$$

and arguably one of the most successful and influential algorithms for this purpose belong to the family of *Markov Chain Monte Carlo* techniques (MCMC) [37].

The idea behind the basic MCMC is to define a Markov random process with a limiting distribution which coincides with the PDF we wish to obtain samples from, then, under some constrains on the transition kernel and the Markov Chain itself, the *ergodic theory* ensures the distribution of samples *over time* will match the desired PDF. For a typical, discrete valued, time-homogeneous (stationary) Markov Chain, the limiting distribution $\mathbf{p}^*$ is defined as a fixed point of its transition matrix, thus also called its *invariant distribution*:

$$\mathbf{p}^* = \mathbf{M}\mathbf{p}^* \tag{2.27}$$

where $\mathbf{M}$ is a *left stochastic matrix* with columns summing to unity and $\mathbf{p}^*$ is a vector with entries corresponding to the values of a probability mass function over the nodes of the chain. When it comes to sampling, however, the interest lies predominantly in continuous random variables, as it is the case with the wave function too. By analogy, we define a continuous transition kernel function $T(\mathbf{x}, \mathbf{x}')$ defining the conditional probability density $q(\mathbf{x}'|\mathbf{x})$ under which the expression for a limiting distribution takes the following form [38]:

$$p(\mathbf{x})^* = \int T(\mathbf{x}', \mathbf{x}) p^*(\mathbf{x}') d\mathbf{x}' \tag{2.28}$$

To ensure a desired PDF is an invariant one for particular time-homogeneous transition kernel, it is sufficient to satisfy so called *detailed balance* [38]:

$$T(\mathbf{x}', \mathbf{x}) p^*(\mathbf{x}') = T(\mathbf{x}, \mathbf{x}') p^*(\mathbf{x}) \tag{2.29}$$

which ensures *reversibility* of the Markov Chain. Further properties it should satisfy are *irreducibility* - the possibility to reach any state from any other in finite time - and *aperiodicity* - absence of dead loops.

Considering all of the above, the celebrated *Metropolis-Hastings* algorithm, postulated by N.Metropolis in 1953 and extended by W.K.Hastings in 1970 [38], generates proposals $\mathbf{x}'$ according to a transition distribution $q(\mathbf{x}'|\mathbf{x})$ which are subsequently either accepted or rejected as a sample with acceptance probability:

$$A(\mathbf{x}'|\mathbf{x}) = \min\{1, \frac{p(\mathbf{x}')q(\mathbf{x}|\mathbf{x}')}{p(\mathbf{x})q(\mathbf{x}'|\mathbf{x})}\} \tag{2.30}$$

whereby $q(\mathbf{x}'|\mathbf{x})$ is commonly chosen as an isotropic, multivariate Gaussian centered on the current state $\mathcal{N}(\mathbf{x}, \sigma\mathbf{I})$. The choice of standard deviation σ of the proposal distribution defines a crucial trade-off between the speed of exploration of the state space and the fraction of states that get accepted. Indeed, in the above formulation we have to do with a random walk with an effective step size regulated by σ. If we consider a highly correlated, multivariate target distribution - a case of considerably different length scales on which the entries of $\mathbf{x}$ can vary - it is clear that in order to maintain reasonable acceptance rates one should match σ with the smallest length scale σ_{min} of the target distribution, which greatly decreases exploration efficiency.

Except from the random walk behaviour and the slow convergence rates it might imply, the Metropolis-Hastings algorithm suffers from additional disadvantage - high autocorrelation between samples due to the "locality" of the mechanism that generates them. A straightforward remedy is to pick only every $i - th$ entry of the generated chain - so called *thinning* - but it has obvious consequences in efficiency. We shall elaborate on the topic of sampling and introduce more intelligent remedies to the above problems in the main section of this thesis as by the account of many researchers, QMC is bugged by the poor efficiency of sampling, in some cases with MCMC acceptance rates as low as 0.1% [39]. This becomes particularly problematic when the goal is an accurate resolution of the correlation energy, which as already mentioned, in spite of its immense significance, accounts for only a small portion of the total energy.

Variational optimization

Having sampled the parametric wave function Ψ_θ we can evaluate the optimization objective and approximate the parameter update $\delta\theta$ hoping that such a procedure will be sufficiently well behaved to eventually lead to converge towards the ground state. Traditionally, QMC methods only optimize the variational parameters belonging to the correlation envelopes introduced e.g. with the Jastrow factor - functions $\chi(\mathbf{x})$ and $u(\mathbf{x}_i, \mathbf{x}_j)$ - or the backflow transformation. The single-electron wave functions $\phi(\mathbf{x})$ which constitute Slater determinants in equation 2.21 are usually obtained a priori from the Hartree-Fock or DFT self-consistent-field calculations and remain *unchanged* during the variational optimization procedure [21]. This in not a necessity but rather a pragmatic compromise, the very complex energy landscape of many-body Hamiltonians, especially those describing fermionic systems, poses a formidable challenge in obtaining reasonable updates, and keeping the dimensionality of the optimization problem low is a straightforward way to simplify it. Further still, this complexity leads to very poor results of naive attempts to use stochastic gradient descent [40], which tends to get stuck oscillating back and forth along steep energy wells. Instead, an optimization method of choice is the so called *stochastic reconfiguration*, proposed by Sorella and Capriotti [17, 41], which was de facto developed to stabilize the "sign problem" in Diffusion Monte Carlo (DMC) simulations.

Let us consider an imaginary time Schrödinger equation:

$$-\hbar\frac{\partial}{\partial\tau}\Psi(\mathbf{x}_1, ..., \mathbf{x}_N, \tau) = (\hat{H} - E_T)\Psi(\mathbf{x}_1, ..., \mathbf{x}_N, \tau) \tag{2.31}$$

which is nothing but a change of variables $\tau = it, t \in \mathbb{R}$ from the original Schrödinger equation eq. 1.1. It has a curious property that its solution is no longer oscillatory but rather an exponentially decaying superposition of the eigenstates. This fact can be leveraged to obtain the ground state ψ_{E_0}, since if we chose the energy offset $E_T = E_0 := 0$, it will be the only state that remains stable as $\tau \to \infty$ - in other words, it gets "projected out" [1] of the complete solution:

$$\Psi(\mathbf{x}_1, ..., \mathbf{x}_N, \tau) = \sum_n c_n e^{-\omega_n\tau}\psi_{E_n}(\mathbf{x}_1, ..., \mathbf{x}_N) \qquad \omega_n = \frac{E_n}{\hbar} \tag{2.32}$$

[1] Due to that projection mechanism, diffusion Monte Carlo is often used after optimization of the trial wave function using VMC to further purify the ground state out of the other contaminating eigenstates and even break through the ansatz limit provided it has been able to faithfully capture the true nodal surface of the true ground state.

The stochastic reconfiguration technique, begins with considering a small variation of the variational parameters $\delta\theta$, the corresponding change in the value of our parametric wave function upon such variation will therefore read:

$$\left|\Psi'_{\theta}\right\rangle = \left|\Psi_{\theta}\right\rangle + \sum_{k=1} \delta\theta_k \frac{\partial}{\partial\theta_k} \left|\Psi_{\theta}\right\rangle \tag{2.33}$$

which for the sake of convenince is usually reformulated to:

$$\left|\Psi'_{\theta}\right\rangle = \sum_{k=0} \delta\theta_k O^k \left|\Psi_{\theta}\right\rangle \tag{2.34}$$

with O^0 being an identity and O^k have been slightly reformulated to logarithmic derivative for stability considerations.

At this point we return to the projection property of DMC, from eq. 2.31 we deduce that the *exponential map* of $(\hat{H} - E_T)$ indeed is responsible for the infinitesimal projection of $|\Psi_t heta\rangle$ onto a configuration that has a suitably lower energy in accordance to the energy shift $\Lambda = E_T$ chosen. This projection, we wish to assimilate with the variation of parameters $\delta\theta$ which would hence define a well-founded iterative scheme to reach the minimum possible energy from parameters θ exploring their linear neighbourhood. In conjunction with eq. 2.34, this defines the stochastic reconfiguration relations [36, 41]:

$$\delta\theta = \mathbf{S}^{-1}\mathbf{f} \tag{2.35}$$

where $S_{ij} = \langle\Psi_{\theta}|\, O^i O^j \,|\Psi_{\theta}\rangle$ are elements of the covariance matrix and $f_i = \langle\Psi_{\theta}|\, O^i(\Lambda i\hat{H})\,|\Psi_{\theta}\rangle$, both being computed stochastically from the Monte Carlo samples we obtain during the evaluation of the energy expectation.

Notice the exact correspondence of the form of above equation with the second order optimization schemes like e.g. the Quasi-Newton methods, in which case the matrix $\mathbf{S}^{-1}$ would be a preconditioner of the gradient based on the local curvature of the parameter space. It turns out, such an analogy can indeed be made [42] and the matrix $\mathbf{S}$ actually corresponds to the *Fisher Information Matrix* of the *natural gradient method* postulated by Amari et al. [43, 44]:

$$S_{\alpha\beta} = \left\langle O_{\alpha}^{\dagger} O_{\beta} \right\rangle - \left\langle O_{\alpha}^{\dagger} \right\rangle \left\langle O_{\beta} \right\rangle \qquad \text{with} \qquad O_{\alpha}\left|x\right\rangle = \frac{\partial \log \Psi_{\theta}(x)}{\partial\theta_{\alpha}} \left|x\right\rangle \tag{2.36}$$

The stochastic reconfiguration, therefore gains a new interpretation of a second order optimization algorithm on stochastic manifolds, that is, manifolds where each point corresponds to a parametric probability distribution. The step directions it defines correspond to geodesics of that manifold with a distance determined by the Kullback-Leiber divergence between Ψ_{θ} and $\Psi_{\{\theta+\delta\theta\}}$.

The only problem with natural gradient is that it requires an inverse of the Fisher information matrix (FIM) which quickly becomes computationally intractable whenever the number of variational parameters grows. As a remedy, Martens et al. proposed a Kronecker-factored Approximate Curvature [42] approximation, aimed primarily at optimization of the neural network quantum states which shall be covered soon. Without focusing on too many computational details, the gist of the method they propose is based on the reasoning that in trying to predict a sensitivity with respect to some parameter θ^l_α belonging to l-th layer of the network, the most useful seem to be the parameters from the same layer. Hence the FIM matrix can approximated as a block diagonal one, with block sizes corresponding to widths of individual layers, and computing its inverse requires only computing l inverses of smaller matrices. Alternatively, including also the adjacent layers $l-1$ and $l+1$ offers more accurate but more expensive trade-off motivated by the information flow in the forward and backward pass during the process of training a neural network.

This section concludes our discussion of classical electronic structure methods, in what follows we present the state-of-the-art approaches utilizing neural networks as a variational ansatz within the Quantum Monte Carlo framework.

2.2. Machine learning quantum states

In this section we discuss the state-of-the-art methodology of utilizing deep learning in electronic structure calculations. It has been stated previously that all information about a quantum system is probabilistically contained within the highly multidimensional wave function. Given the success of neural networks in approximating even very high dimensional mappings, it is not completely unjustified to apply the proven methods of information theory to obtain reliable approximations in the realm of quantum mechanics. Of crucial importance is the representational power and training efficiency. In the case of the former, utilizing physically motivated symmetries to introduce weight sharing [45] or model transparency in order to gain physical insight from the structure of the model itself are obvious considerations. Regarding training, speed of convergence as well as quality of the optima are fundamental, but also here one can consider physically motivated update schemes.

In this review, we focus mainly on the self-supervised, or *variational* approach, capable of modelling the wave function directly - surrogates of other kind, like those trained on data sets of electronic structure calculations for various properties [46–49] or generative approaches for inverse material design [50–53] are not considered. The main aim of the variationally trained neural networks for wave function approximation - which bear similarity to reinforcement learning [54] - is to obtain chemically accurate solutions to the many-body Schrödinger equation, eq. 1.10, which is hoped to further improve on the accuracy and fidelity of QMC methods covered so far.

We start with the in-depth coverage of foundational work of Carleo & Troyer [54] on *neural quantum states* which could be classified as a Fock space approach for lattice models. Only later, we will focus on the more recent PauliNet [55] and FermiNet [56] which both approximate the wave function variationally in the real space of electron coordinates. They represent the current state-of-the-art for approaches not constrained to a lattice but vary vastly in terms of their design philosophy. The following material in this chapter will be mostly structured according to the reviews by [57] and [58] but completed with additional details relevant for each chapter.

2.2.1. Restricted Boltzman Machine

The RBM architecture is by far the best studied method with regards to neural quantum states (NQS), mostly due to its precedence since its publication by Carleo and Troyer [54], but also because it is a reasonably simple, graphical model, which in fact, already has some relation to physics - it belongs to a class of *energy-based models* [59]. Generally speaking, Boltzmann Machines represent the joint probability distribution over N visible σ_j, and M hidden h_i, mutually interconnected, binary random variables as a Boltzmann (isothermal) distribution:

$$p(\boldsymbol{\sigma}, \mathbf{h}) = \frac{1}{Z} e^{-E(\boldsymbol{\sigma}, \mathbf{h})} \tag{2.37}$$

whereby $\sigma_j, h_i \in \{-1, 1\}$, Z is the partition function accounting for proper normalization and the energy function $E(\mathbf{v}, \mathbf{h})$ is given by:

$$E(\boldsymbol{\sigma}, \mathbf{h}) = -\sum_{j=1}^{M} a_j \sigma_j - \sum_{i=1}^{N} b_i H_i - \sum_{i=1}^{N} \sum_{j=1}^{M} h_i W_{ij} \sigma_j \tag{2.38}$$

In our setting, visible variables $\boldsymbol{\sigma}$ represent the spin occupations on the nodes of a lattice and the trainable model parameters $\{a_j, b_i, W_{ij}\} \in \mathbb{C}$ determine the magnitude of interactions between the spin states, through the auxiliary hidden layer. It is exactly the presence of $\mathbf{h}$ that distinguishes the Boltzmann Machine from an *Ising model*, furthermore a *restricted* Boltzmann machine (RBM) removes any connections among the hidden and visible variables themselves making the probabilistic graph a bipartite one, in which entanglement between the spin states states is mediated through the hidden variables only [57]. Such formalism makes training not only easier but also provides a straightforward control over the expressiveness of the model - the larger the hidden variables count, the more sophisticated correlation features can be captured.

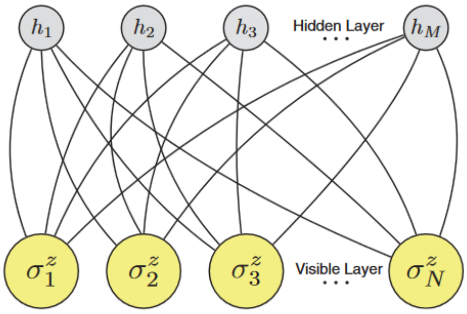

Figure 2.3.: General architecture of a Restricted Boltzmann Machine, source [54]

It is important to realize that the RBM, as proposed by [54], does not represent the wave function $|Psi\rangle$ itself but rather the expansion coefficients, or amplitudes ψ, of the linear combination of tensors spanning a subspace of so called *Fock space*. For a system of identical spin particles, in particular bosons, described by model Hamiltonians such as the transverse-field Ising or the antiferromagnetic Heisenberg models covered in the original publication, we can explicitly list all the discrete degrees of freedom and so, the *full configuration interaction* ansatz reads:

$$|\Psi\rangle = \sum_{\boldsymbol{\sigma}} \psi_{\boldsymbol{\sigma}} |\boldsymbol{\sigma}\rangle \tag{2.39}$$

with $|\boldsymbol{\sigma}\rangle$ being a vector with N elements representing any particular configuration where each entry corresponds to a particular site on the lattice. The sum runs over all 2^N configurations since each state can be either occupied with a spin up $\sigma_i = 2S_i^z = 1$ or spin down $\sigma_i = S_i^z = -1$ particle [2].

[2] whereby peculiarities of the implementation depend on the use case, here we consider the $S = \frac{1}{2}$ Hubbard model

Importantly - although it is clear from the above that the problem size grows exponentially in M - RBMs require only a polynomial number of parameters $\mathcal{O}(NM + N + M)$ which is a great advantage. Notice, however, that for a complete description of a quantum state, both the amplitude and the phase factor of the wave function are needed, therefore the coefficients ψ_{σ} and hence all parameters must admit complex values [54]. Some extensions, however, like the Deep Boltzmann Machines, can explicitly model the phase and amplitude of a wave function coefficients and therefore, get away with only real parameters [57].

Finally, to obtain ψ_{σ} we need to marginalize the hidden units out and, up to a normalization factor, we obtain:

$$|\psi_{\sigma}\rangle \sim \prod_{j=1}^{M} e^{a_j \sigma_j} \prod_{i=1}^{N} 2\cosh\left(b_i + \sum_{j=1}^{M} W_{ij}\sigma_j\right) \tag{2.40}$$

Due to the their proven universality for representing discrete probability distributions [60] as well as precedence in learning neural quantum states, many further investigations have been based on the RBM architecture. The success of the above ansatz is its flexibility, $|\psi_{\sigma}\rangle$ already has the representational power to reproduce the common correlation factors like the the Jastrow or the Gultzwiller ones [61], nonetheless one of the first improvements of the original work was its extension by Nomura et al. [61]. They introduced an additional factor $\langle \boldsymbol{\sigma} | \phi_{ref} \rangle$ choosing the reference state $|\phi_{ref}\rangle$ as a *pair-product* (or *geminal*) wave function to capture some of the non-local entanglement crucial in describing strongly correlated systems. Such modification already allowed the treatment of not only bosonic but also fermionic systems by enabling the ansatz to optimize the nodal structure of $|\Psi\rangle$ which in case of fermions is crucial and naively doubling the number of visible units to allow for double occupations would lead to poor results. Nonetheless, encoding fermionic asymmetry directly into the ansatz does restrict the freedom of the nodal structure to some extend and therefore the exact ground state cannot, in principle, be achieved, even in the limit of infinitely many variational parameters. With a similar goal, but a slightly different approach, Valenti et al. [62] extended the energy functional, eq. 2.38, by explicitly adding visible neurons corresponding to certain spin *products* $\sigma_l \cdot ... \cdot \sigma_k$. In effect, a Jastrow-factor-like term has been added to the ansatz enabling flexibility and transparency in capturing those many-electron correlations which might be of most interest according to physical intuition.

The representation of fermionic quantum states has been further explored by Choo et at. [39] who ingeniously mapped the fermionic problem to an equivalent spin one with the help of the *Jordan-Wigner transformation* often used in fermionic simulations on quantum computers. Using a minimal basis set STO-3G with 20 spin-orbitals their results on dissociation curves of C_2 and N_2 molecules surpassed the standard quantum chemical methods of CCSD and CCSD(T) reaching almost perfect agreement with the exponentially scaling FCI. Using the same technique Yoshioka et al. [63] studied crystalline solids, in particular the 1D hydrogen chain, 2D graphene lattice and the 3D lithium hydride crystal and again, reached very good agreement with FCI. Moreover, they performed first NQS simulation of excited states beyond the first excited state performed primarily by Choo et al. [64]. Under the assumption that single quasiparticle excitations dominate the low-lying band spectrum, they computed the band structure from linear-response behaviour of the ground state represented with RBM and reported good agreement with standard methods for the first valence and conduction bands.

Turning to publications which focus more on the physics of the model itself, Deng et al. [65] studied entanglement properties of RBM quantum states, in particular entanglement entropy and spectrum. They were able to prove all short-range RBM states satisfy area-law of entanglement and bipartition geometry [66] as well as that unlike for tensor networks, the (volume-law) entanglement was not a limiting factor for the efficiency of RBM representation. Park and Kastoryano [67] investigated the geometry of the parameter space of complex-valued RBM. Their analysis revealed that the weights of the model do not reveal much insight because of the multitude of equivalent representations near the ground state, however the spectrum of the quantum Fisher information matrix used in stochastic reconfiguration, c.f. sec. 2.1.3, does convey interesting information about the electronic entanglement. In particular - somewhat analogously to what is often observed in dimensionality reduction with PCA - the largest eigenvalues corresponded to eigenvectors dominated by first moments and hence did not contain much information about correlations in the system.

Regarding extensions to deep architectures, Gao and Duan [68] proved inefficiency of RBM to represent certain quantum states like those originating from constant-depth quantum circuits or ground states of gaped Hamiltonians, unless $\#P \subset P$. They proposed a Deep Boltzman Machine with one more layer of hidden states, which was able to overcome that limitation, but at the same time complicated the inference due to the need of additional Monte Carlo sampling of the hidden spin states. With applications to spin-1/2 Heisenberg model, Kochkov et al. [69] proposed a graph neural network architecture reminiscent of an auto-encoder which, given the spin configuration $\boldsymbol{\sigma}$ and the sublattice encoding, computes the latent representations of the wave-function and eventually, returns the wave function logarithmic amplitudes separately to its phase. They argued that keeping the two apart, apart from having the benefit of real variational parameters, was critical to enable effective generalization of the learned sign structure. Furthermore, the distributed treatment of the lattice, enhanced with the information about its local symmetries as well as the sum-based reduction across graph vertices made the model readily applicable to a large variety of lattice shapes and sizes. On a final note, the work of Sharir et al. [70] could be regarded as a paradigm shift which gets rid of the undesirable features of the Monte Carlo sampling which hinder the applicability of deep architectures in Neural Quantum States. Virtually all of the approaches utilizing machine learning follow the same sampling scheme as VMC (see sec. 2.1.3), which despite being asymptotically well-behaved, in practice suffers from long burn-in times and struggles with multimodal or sharply peaked distributions. In contrast to that, they proposed an *autoregressive model*, adjusted accordingly to treat complex-valued wave functions, and reported not only its qualitative and quantitative advantage over the original work of Carleo and Troyer [54] but also at much shorter simulation times when applied to transverse field Ising and antiferromagnetic Heisenberg models.

2.2.2. FermiNet

It has been previously stated that any antisymmetric function can be represented by an infinite linear combination of Slater determinants built from single-particle wave functions $\phi_i(\mathbf{x})$ but that unfortunately leads to exponential asymptotic complexity. An alternative is to allow $\phi_i(\mathbf{x})$ to be a function of all variables, in which case every antisymmetric function can be represented with just a *single* determinant, which can be computed in $\mathcal{O}(N^3)$ time [56, 71]:

$$\Phi[\phi_{I_1}, ..., \phi_{I_N}](\mathbf{x}_1, ..., \mathbf{x}_N) = \det \begin{pmatrix} \phi_1(\mathbf{x}_1|\{\mathbf{x}_{\neq 1}\}) & \cdots & \phi_N(\mathbf{x}_1|\{\mathbf{x}_{\neq 1}\}) \\ \vdots & \ddots & \vdots \\ \phi_1(\mathbf{x}_N|\{\mathbf{x}_{\neq N}\}) & \cdots & \phi_N(\mathbf{x}_N|\{\mathbf{x}_{\neq N}\}) \end{pmatrix} \tag{2.41}$$

where $\{\mathbf{x}_{\neq i}\} := \{\mathbf{x}_1, ..., \mathbf{x}_{i-1}, \mathbf{x}_{i+1}, ..., \mathbf{x}_N\}$ and if we assume $\phi_i(\mathbf{x}_j|\{\mathbf{x}_{\neq j}\})$ is symmetric in all but the j-th variable [3] - hence the notation - exchanging $\mathbf{x}_i$ and $\mathbf{x}_j$ will remain equivalent to exchanging rows of i and j in the above matrix and therefore Φ remains antisymmetric.

This is exactly the idea postulated by Pfau et al. [56], which they called *generalized Slater* ansatz and which by the argument of Hutter [71] is able to represent any antisymmetric function of n electrons, for all cases where $\mathbf{x}_i \in \mathbb{R}^d,\ d > 1$, however, the proof necessitates discontinuous $\phi_i(\mathbf{x}_j|\{\mathbf{x}_{\neq j}\})$. Recently, Huang et al. [72] undermined the representational power of such an ansatz by considering its algebraic-geometric structure. Although restricting $\phi_i(\mathbf{x}_j|\{\mathbf{x}_{\neq j}\})$ to finite-degree polynomials only, by providing a bound on the dimensions of the target and source spaces they postulated the necessity of at least $\mathcal{O}(N^{3N-N})$ determinants like the above in order to represent a generic, totally antisymmetric polynomial. Although their result remains of limited practicality when the orbitals are modelled by neural networks - dubbed universal function approximators [73] - it hints, however, that approximating a continuous, antisymmetric wave function may be better behaved when more than one determinant is used. This has been also empirically confirmed in the implementation of FermiNet [56, 74] which in practice, performed better with a small (8 to 32) linear combination of determinants. The FermiNet ansatz therefore reads:

$$\Psi_F(\mathbf{x}_1, ..., \mathbf{x}_N) = \sum_{I=1} \alpha_I \Phi_I^{\uparrow}(\mathbf{r}_1, ..., \mathbf{r}_{n^\uparrow}) \Phi_I^{\downarrow}(\mathbf{r}_{n^\uparrow+1}, ..., \mathbf{r}_N) \tag{2.42}$$

and the functions $\phi_i^{I\alpha}(\mathbf{r}_j^\alpha|\{\mathbf{r}_{\neq j}^\alpha\})$ constituting every Φ_I^α are modelled by a neural network as presented eq. 2.45 , where α stands for spin and I is the combination index as explained in section 2.1.1.

The architecture of FermiNet is centered around two streams of information flow, the electron-nuclei and electron-electron interaction, whereby the initial features $\mathbf{h}_i^0$ and $\mathbf{h}_{ij}^0$ respectively are just a concatenation of the relative position vectors $\mathbf{r}_i - \mathbf{R}_k, \forall k$ and $\mathbf{r}_i - \mathbf{r}_j$ their absolute lengths. The latter are fed explicitly as additional inputs because, as the authors argue, that removes the need to separately include Jastrow factor after the determinant, which

[3] which can be trivially achieved by defining it in terms of their pairwise distances as it was done in eq. 2.25. Note also, that rigorously, one should be using the term multi-symmetry or block-symmetry here, as symmetry under exchanging components of any $\mathbf{x}_i \in \mathbb{R}^d,\ d > 1$ is not provided

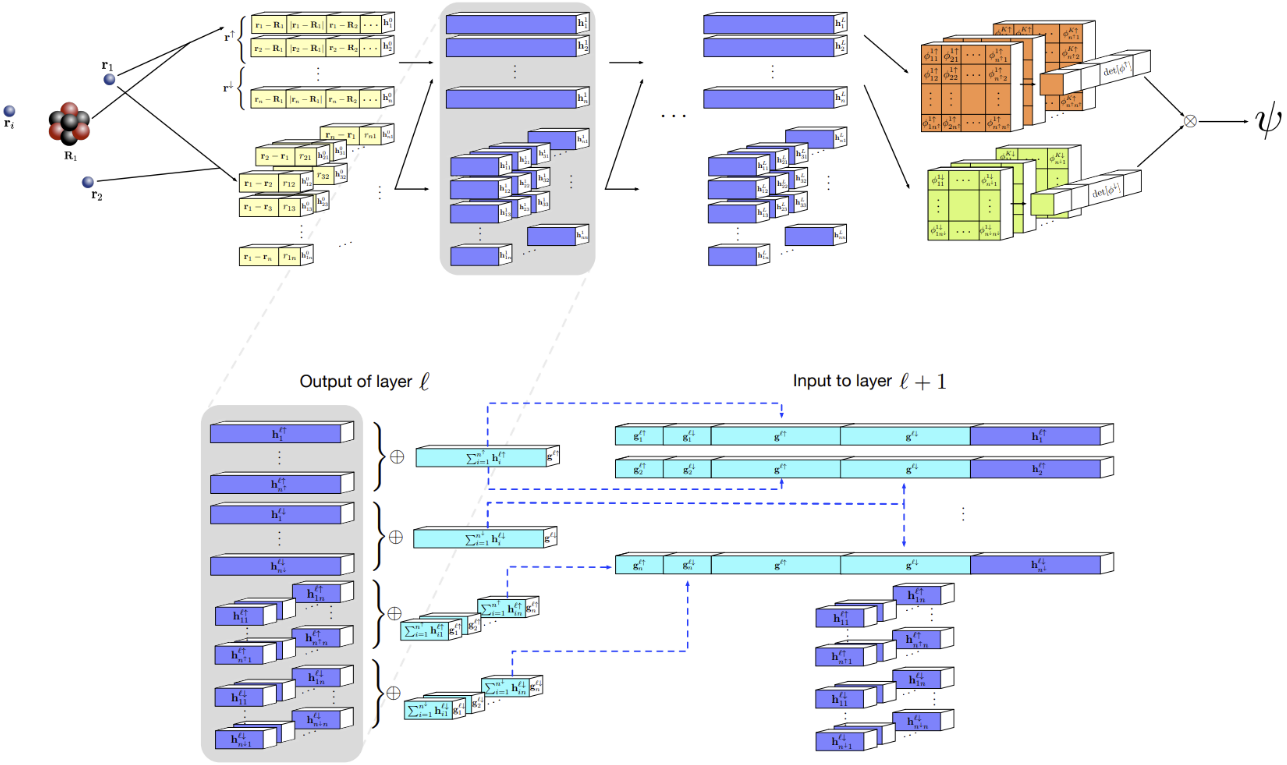

Figure 2.4.: The FermiNet architecture. Top: Global architecture, bottom: detailed view on information streams in each layer. Source: [56]

was found to be extremely unstable when used with orbitals initialized not from Hartree-Fock calculations but from random initial weights. Notice also, that such an explicit incorporation of nuclei positions $\mathbf{R}_k$, although defined invariantly of external coordinate systems, makes the network *single purpose* i.e. not generalizible across different molecules.

The features are then feed through a standard feed forward neural network with tanh non-linearity and residual connection [75]:

$$\mathbf{h}_i^{l+1} = \mathbf{h}_i^l + \tanh\Big(\mathbf{V}^l \mathbf{f}_i^l + \mathbf{b}^l\Big), \qquad \mathbf{h}_{ij}^{l+1} = \mathbf{h}_{ij}^l + \tanh\Big(\mathbf{W}^l \mathbf{h}_{ij}^l + \mathbf{c}^l\Big) \tag{2.43}$$

and we observe the electron-electron interaction streams $\mathbf{h}_{ij}$ are just propagated forward to the next layer without much information exchange whereby the electron-nuclei streams accumulate all the information via concatenation of simple, sum-based reductions of features sorted by spin:

$$\mathbf{h}_i^l \to \mathbf{f}_i^l = \text{concat}\left(\mathbf{h}_i^l, \frac{1}{n^\uparrow}\sum_{j=1}^{n^\uparrow}\mathbf{h}_j^l, \frac{1}{n_\downarrow}\sum_{j=1+n^\uparrow}^{N}\mathbf{h}_j^l, \frac{1}{n^\uparrow}\sum_{j=1}^{n^\uparrow}\mathbf{h}_{ij}^l, \frac{1}{n_\downarrow}\sum_{j=1+n^\uparrow}^{N}\mathbf{h}_{ij}^l\right) \tag{2.44}$$

After a series of L layers the final one electron features $\mathbf{h}_i^L$ undergo one last linear transformation and a multiplication with a number of *exponential envelopes* to enforce correct boundary conditions of each orbital at infinity distance from every nuclei:

$$\phi_i^{I\alpha}(\mathbf{r}_j^\alpha|\{\mathbf{r}_{\neq j}^\alpha\}) = \Big(\mathbf{w}_i^I \cdot \mathbf{h}_j^L + g_i^I\Big) \times \sum_k \pi_{ik}^I e^{-\sigma_{im}^I|\mathbf{r}_j - \mathbf{R}_k|} \tag{2.45}$$

with $\mathbf{w}_i^I$, g_i^I, π_{ik}^I, σ_{im}^I being further learnable parameters, whereby the last one is an isotropic decay rate of the exponential envelope - a simplification introduced in [74] instead of an anisotropic, 3x3 coefficient matrix Σ_{im}^I. Also, compared to the original publication we have dropped the spin indices α, because it is obvious from the ordering rule of electrons i.e. $\alpha =\uparrow$ if $0 < i \leq n^\uparrow$ and $\alpha =\downarrow$ if $n^\downarrow \leq i \leq N$ with $n_\downarrow = N - n^\uparrow$.

The orbitals are then finally assembled into a linear combination of spin-up and spin-down Slater determinants of equation 2.42 as depicted schematically in fig. 2.4 and the entire ansatz is trained variationally in accordance to the general QMC scheme covered in section 2.1.3. For numerical reasons, the determinants were computed in the log domain, therefore yielding log wave function amplitudes, furthermore, only a block diagonal Kronecker-factored Approximate Curvature [42] optimization scheme has been used due to the sheer number of parameters.

Regarding the results, FermiNet approximates wave functions directly in the anti-symmetrized Hilbert space and therefore, it is independent of any restrictions an incomplete basis set might have induces. This flexibility really unfolds when considering systems with significant electron correlations. In particular the H_4 rectangle, the dissociation of N_2 and the H_{10} chain have been studied, in all cases consistently beating the accuracy of the standard quantum chemical methods, like the unrestricted CCSD(T), coming close to the highly specialized, multi-reference methods like R12-MR-ACPF, yet at only *polynomial scaling*.

Nonetheless, the constant prefactor of computational complexity remains high and training times vary between a few hours for the smaller systems, up to a month for bicyclobutane using 8 to 16 GPUs. It has been meaningfully decreased in a follow up paper [74] in which the network architecture has been simplified and implemented efficiently in JAX [76] leading to a much larger GPU utilization and altogether to an order-of-magnitude reduction in the compute time. Moreover, due to that increase in efficiency, better results could have been obtained by increasing the network width which aligns with the theoretical guaranties of universality of such an ansatz [71] but perhaps more importantly, increasing the number of Slater determinants as well as the MCMC steps did prove even more significant. The latter hints that perhaps not only the brute force computational power but also physical priors do matter, which shall be demonstrated even more vividly in the next section. Lastly, regarding efficiency, the work of Ren et al. [77] must be mentioned, in which Diffusion Monte Carlo simulation using the FermiNet ansatz for systems exhibiting various degree and kind of electron correlation has been performed. Crucially, they demonstrated that even an undertrained network captures the nodal surface of the ground state exceptionally well and as an effect, with significantly smaller overall computational effort, FermiNet-DMC can not only be applied to larger systems, but converges to much better results - in the case of the dissociation of the N_2 molecule for example, the most accurate ever reported.

2.2.3. PauliNet

FermiNet is completely general, in the sense that it does not include any known physical features of the wave function besides the antisymmetry and appropriate boundary conditions. Hermann et al. [55] proposed a more traditional VMC approach, where the deep neural network is used only to modify the ansatz, obtained otherwise from self-consistent calculations like Hartree-Fock or CASSCF $\phi_i^{I\alpha}(\mathbf{r}_j^\alpha)$. In particular, the proposed model named PauliNet, implements a backflow transformation (cf. sec. 2.1.3) using a deep neural network, generalizing the original idea of *neural network backflow* proposed by Luo and Clark [78] to continuous space. The major advantage of such an approach is the ability to explicitly encode arbitrarily complex electron correlations, as well as the sign structure, at a relatively small cost. Instead of returning the wave function amplitudes directly, as in the previous two approaches, here we transform the single-electron orbitals in a configuration dependent way, therefore, efficiently and systematically expanding on the bulk information contained within the mean-field approximation. The PauliNet ansatz therefore reads:

$$\Psi(\mathbf{r}) = e^{\gamma(\mathbf{r})+J(\mathbf{r};\theta)} \sum_{I=1} \alpha_I \Phi_I^{\uparrow}(\mathbf{r}_1, ..., \mathbf{r}_{n^\uparrow}) \Phi_I^{\downarrow}(\mathbf{r}_{n^\uparrow+1}, ..., \mathbf{r}_N), \tag{2.46}$$

and the one-electron wave functions which constitute each Slater are given by:

$$\tilde{\phi}_i^{I\alpha}(\mathbf{r}_j^\alpha) = \phi_i^{I\alpha}(\mathbf{r}_j^\alpha) f_{ij}^{I\alpha}(\mathbf{r};\theta) \tag{2.47}$$

Both, the Jastrow factor $J(\mathbf{r};\theta)$ as well as the multiplicative backflow coefficients $f_{ij}^{I\alpha}(\mathbf{r};\theta)$ corresponding to each ij-th entry of I-th, α-spin Slater matrix are functions of *all electrons*

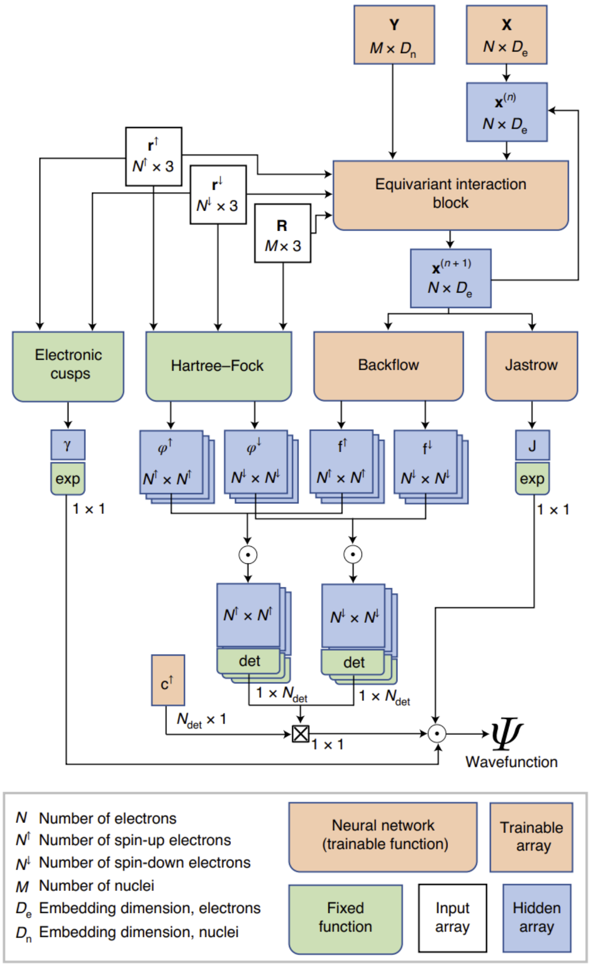
$\mathbf{Y}$
$M \times D_n$
$\mathbf{X}$
$N \times D_e$
$\mathbf{x}^{(n)}$
$N \times D_e$
$\mathbf{r}^{\uparrow}$
$N^{\uparrow} \times 3$
$\mathbf{r}^{\downarrow}$
$N^{\downarrow} \times 3$
$\mathbf{R}$
$M \times 3$
Equivariant interaction block
$\mathbf{x}^{(n+1)}$
$N \times D_e$
Electronic cusps
Hartree–Fock
Backflow
Jastrow
γ
exp
1×1
$\varphi^{\uparrow}$
$N^{\uparrow} \times N^{\uparrow}$
$\varphi^{\downarrow}$
$N^{\downarrow} \times N^{\downarrow}$
$\mathbf{f}^{\uparrow}$
$N^{\uparrow} \times N^{\uparrow}$
$\mathbf{f}^{\downarrow}$
$N^{\downarrow} \times N^{\downarrow}$
J
exp
1×1
$N^{\uparrow} \times N^{\uparrow}$
det
$N^{\downarrow} \times N^{\downarrow}$
det
$\mathbf{c}^{\uparrow}$
$1 \times N_{\text{det}}$
$1 \times N_{\text{det}}$
$N_{\text{det}} \times 1$
1×1
Ψ
Wavefunction
N Number of electrons
$N^{\uparrow}$ Number of spin-up electrons
$N^{\downarrow}$ Number of spin-down electrons
M Number of nuclei
D_e Embedding dimension, electrons
D_n Embedding dimension, nuclei
Neural network (trainable function)
Trainable array
Fixed function
Input array
Hidden array

Figure 2.5.: PauliNet architecture, the "Equivariant interaction block" is a layer of an appropriately adapted SchNet architecture [49] to establish electron embeddings $\mathbf{x}^L$ subsequently fed into Jastrow and backflow neural networks as described in the text. Source: [55]

and modelled by deep neural networks. Additionally, explicit cusp conditions are enforced by the $\gamma(\mathbf{r})$ function with fixed coefficients as explained in sec. 2.1.3. Within the described architecture, to retain antisymmetry and correct behaviour at cusps, both the Jastrow as well as the backflow neural functions must be modelled as *cuspless* and *invariant*, respectively *equivariant*, to exchange of same-spin electrons. To achieve this goal, Hermann et al. adapted a neural network architecture from a slightly different area of supervised learning of electronic properties on molecules, so called SchNet [79] neural network.

A traditional convolutional filter [80], for grid-ordered data, would be a $D \times W \times B$ tensor where $W \times B$ defines, a usually square, receptive field and D is the number of channels in the input, after applying K such filters the output will have K channels and appropriately smaller size based on the size of the receptive field and padding used. A continuous convolution filter, proposed by Shutt et al. [49], aggregates and applies a transformation to input features across continuous directions from the kernel's centre, unrestricted to any grid. A certain simplification is to only consider the radial distance between feature positions $r_{ij} = |\mathbf{r}_i - \mathbf{r}_j|$, another one is to only consider a so called *depth-separable convolutions*, in which the F-dimensional feature vectors $\mathbf{x}_i$ at any point of the D-dimensional space are only multiplied element-wise (and aggregated), and therefore remain constant in size. That amounts to the following:

$$\mathbf{x}_i^{l+1} = \sum_j \mathbf{x}_j^l \odot C^l(\mathbf{r}_i, \mathbf{r}_j), \qquad C^l : \mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^F \tag{2.48}$$

where l is the layer index and C^l is a filter-generating function, which outputs a filter vector based on the relative positions of the current kernel centre $\mathbf{r}_i$ and position of the j-th feature $\mathbf{r}_j$. In actual implementation, C^l is not, however, a map $\mathbb{R}^D \times \mathbb{R}^D \to \mathbb{R}^F$, nor is it $\mathbb{R} \to \mathbb{R}^F$ which the discussion on radial distance would have suggested. Since we want to model C^l with a feed forward neural network, inputting just a single scalar seems unreasonable, we therefore discretize the radial axis (up until some cutoff radii) and additionally use a distance featurization using some radial basis function $e(r_{ij})$ (in the PauliNet, constrained be cuspless at 0) and use that as an input.

With all these considerations, the "Equivariant interaction block" of PauliNet which learns electron embeddings looks as follows:

$$\begin{aligned} \mathbf{z}_i^{l,\alpha} &= \sum_{j \neq i} B^{l,\alpha}(\mathbf{x}_j^l) \odot C^{l,\alpha}(\mathbf{e}(r_{ij})) \\ \mathbf{z}_i^{l,nuc} &= \sum_k \mathbf{X}_k \odot C^{l,nuc}(\mathbf{e}(R_{ik})) \\ \mathbf{x}_i^{l+1} &= \mathbf{x}_i^l + A^{l,\uparrow}\left(\mathbf{z}_i^{l,\uparrow}\right) + A^{l,\downarrow}\left(\mathbf{z}_i^{l,\downarrow}\right) + A^{l,nuc}\left(\mathbf{z}_i^{l,nuc}\right) \end{aligned} \tag{2.49}$$

where A, B, C are small, feed forward neural networks, l is the layer index, α is the spin index, $\mathbf{X}_k$ are fixed nuclei embeddings, $\mathbf{x}_i^l$ are the iterated electron embeddings, R_{ik} is the electron-nuclei distances and r_{ij} are electron-electron distances. The sum in the first row is implicitly taken only over indices corresponding to electrons with the same spin. Notice the three independent streams of information flow, separately for interactions with other spin-up/down electrons and nuclei and a final sum based reduction accompanied with a residual connection in the spirit of ResNet [75].

The final electron features $\mathbf{x}_i^L$ are subsequently used to predict the Jastrow and backflow coefficients, but overall PauliNet remains a rather small model, with only around 100k trainable parameters compared to over 700k of FermiNet [81]. Although FermiNet manages to achieve slightly better results, it does so, even in the highly optimized version, at more than 5x the computational cost of PauliNet [74]. Still, PauliNet is able to recover between 97% and 99.9% of the correlation energy for a number of atomic and diatomic systems, such as H_2, LiH, Be as well as meet the accuracy of the "gold standard" quantum chemical methods on dissociation of H_{10} chain or automerization of cyclobutatiediene. This emphasizes what can be gained in terms of efficiency when we bias a flexible deep learning ansatz with physical intuition.

Gerard et al. [81], in a model which could be considered a hybrid of the two aforementioned ones, further investigated the role of physical priors. They found that explicit symmetries or cusp conditions generally aid performance which should not be surprising as e.g. explicit cusps were already a well established treatment to accelerate and improve convergence of the standard electronic structure methods [14]. Also, short supervised pre-training or physics-inspired initialization of certain parameters, like those in the envelope, does accelerate optimization but taken too far may lead to detrimental biases. In particular, replacing the $\exp(\gamma(\mathbf{r}) + J(\mathbf{r};\theta))$ factor in PauliNet with a physically inspired CASSCF-envelopes led to strong bias that could not have been overcome during training. Altogether, they obtained the best results up to date, on par or better with the aforementioned Ferminet-DMC [77] or specialized chemistry standards MRCI-F12(Q) [27], with 4-6x fewer compute resources compared to the original FermiNet [56] and 5-10x compared with Ferminet-DMC [77]. According to their ablation studies, after proper hyperparameter tuning, the second most important improvement was using a SchNet-like convolution in a FermiNet-like ansatz, hinting that not only what, but also where the electron features are, matters.

3. Approximating many-electron wave function with physics aware surrogate models

It is a strong conviction of the author that modern deep learning should raise above the stigma associated with neural networks as being solely black box models. Although, a generic guarantee on the representational power of neural networks is perhaps impossible, just their universality in approximating virtually any arbitrarily complex, possibly discontinuous mapping, merely through manipulation of relevant information [73], makes them an indispensable tool in the hands of science provided we use it intelligently. In particular, if in conjunction with providing basic uncertainty quantification [82], the field shifts its focus from "generalizability" by extrapolation and turns instead towards generalizability through *group equivariance* [45], the reluctance of scientific applications will be completely dispelled.

To fully comprehend this claim, consider the role of an *ansatz* in scientific computing. Virtually all of what could be described as variational numerics takes place in a subspace of the complete Hilbert space defined apriori and the subsequent solution process aims at minimizing some sort of residual - an error with respect to the the true solution, which in general, lives outside of the computational basis chosen. Over the years, much has been accomplished in terms of extending the flexibility and representational power of this procedure, to name a few, the hierarchical basis provided a computationally efficient mechanism for extendability of the ansatz, sparse grids offered improved asymptotic complexity thanks to a principled accuracy trade-offs which enabled tackling higher-dimensional problems [83, 84] and lately, wavelets gained in popularity, thanks to the ability to construct non-trivial, problem-specific basis functions with desirable properties like orthogonality [85]. What we postulate, is popularizing the usage of neural networks as a computational ansatz - and whithin scientific realms, only as an ansatz - which encompasses all the benefits of the aforementioned methods. Indeed, what is refered to as transfer learning in deep learning community [86] can be seen as a parallel of the mechanism provided by hierarchical basis, like sparse grids, we can allocate computational resources to only what we deem to be the most important details, and finally, they are essentially basis-free i.e. the granularity of the approximation of the Hilbert space will be guided in the end only by machine precision which makes them inherently problem-specific.

First applications of neural networks in such a framework have already proven succesfull, as one of the first examples one should not forget to mention the Physics-Informed Neural Networks (PINN) by Maziar Rassi et al. [87], but important for us, are the FermiNet [56, 74] and PauliNet [55] networks which we elaborated on in the previous chapter and which also follow this philosophy. Since we became accustomed to neural networks being used in massive regression or classification tasks, seeing one, trained to represent a single, particular solution of some partial differential equation might appear very inefficient. We belive, however, this is where group equivariance can prove very succesful facilitating well founded "generalizibility by construction". Most physically motivated problems are characterised by relatively simple principles which govern their behaviour, that is at least the hope, the entire scientific effort of humanity has been conditioned upon to this day. Conservation principles, symmetries and invariances are all different names for what physics has been able to come up with so far and what is comonly refered to as the *laws of physics*, their proper mathematical treatment, which has established geometry as the unifying language of physics, did not appear until the early nineteen hundrets under the name of the *Noether's theorem* [88].

In the early days of deep learning, similar ideas were nonetheless already present, this is probably best asserted by the publication of the first convolutional neural network (CNN) - the LeNet-5 [80] - in the eighties. It successfully applied a neural network to learn translation invariant, local receptive field, later also called convolution filter, which could sense meaningful, visual features such as oriented edges or corners from a 2D pixel array. Accounting for the deficiency of the feed forward network architecture in respecting topology of the input has been fairly well developed since, with examples beeing the recurrent neural networks [89] for sentences, more recently also graph neural networks [90, 91] which could be regarded as an extension of CNNs to non-uniform/non-euclidean grids or even the attention mechanism [92] which essentially "learns" the topology of the input itself. Extending group equivariance aspects to the outputs of the network, however, did not experience such fruitful development. Only very recently, due to Cohen and Welling [93, 94] as well as Worrall et al. [95] the translational invariance of CNN filters has been extended to rotational invariance by leveraging an old concept of *filter steerability* [96]. Later, publications on full SO(3) group equivariance for volumetric data such as voxels [97] and point clouds followed [48], culminating in the appearance of SO(3) equivariant graph neural networks like the already covered SchNet [49, 79], DimeNet [46, 98], and others [47, 99].

Briefly commenting on the latter, when trained on datasets of ab-initio calculations containing various electronic properties with formation energies as a target, thanks to the differentiability of the network also with respect to its inputs, the energy conserving intra- and inter-molecular forces could have been readily obtained. Leveraging the "hard-coded", directional equivariance of the learned features such models could have been directly applied to ab-initio molecular dynamics (AIMD) simulations [15] like in the work of Li et al. [100]. Crucially, provided obviosuly the network has been trained on atoms constituting molecules in the AIMD simulation, virtually no concerns of its generalizibility must have been considered. Another marvelous account on how equivariance can lead to meaningful generalization is the work of Lie et al. [101] which proposed a rotationally equivariant graph neural network trained to represent the DFT Hamiltonian of crystalline materials in order to bypass the expensive self-consistent-field calculations. In one example, they studied twisted van der Waals materials, in which at certain “magic Moire angle", exotic quantum phases such as correlated insulation or topological superconductivity start to appear. Their network trained on a dataset of DFT calculations of *untwisted* materials was able to successfully, and virtually for free, explore desired properties at arbitrary twist angles, in particular, also reproduce the characteristic effects occurring at the "magic angle" which, due to broken symmetry between the twisted lattices, would have otherwise require an enormous computational effort using the common plane-wave basis.

All of the above, we hope, was enough to convince the reader that our initial claim was indeed well founded and that already, a trend at the intersection of scientific computing and machine learning has been started. The goal of this thesis, however, couldn not have possibly been developing a new model capable of competing with the state-of-the-art deep learning algorithms for quantum chemistry and physics. Considering that every QMC algorithm consists of three major components: wave function ansatz, optimization and sampling, we have decided to focus on the latter, being the most underrated and underdeveloped one. The main body of this chapter has been therefore devoted to investigations of sampling algorithms which could deal with the singularities of the Born probability density - the modulo squared of the wave function $|\Psi|^2$. For our final experiments we have chosen PauliNet as the wave function surrogate because of the physical priors it utilizes, which not only is in line with the aforementioned philosophy but additionally offers much faster training times compared to its competition. The majority of experiments, however, have been performed on synthetic examples meant to mimic certain challenging aspects of sampling from the wave function in order to thoroughly investigate the adequacy of various sampling techniques, which themselves, rely on interesting physically-aware techniques.

3.1. Improved sampling of Born probability density

Sampling efficacy is the major bottleneck of QMC's accuracy, which with the advent of expressive, deep learning wave function ansätze became even more pronounced [39, 55, 56, 63, 81]. Indeed, any neural network, although expressive, is eventually limited in its represetational power by the obvious limitations of memory required to store all its parameters. When used in a highly-dimensional, variational optimization task, which will undoubtely require expectation (or simply integration) of the objective functional *under the function being optimized*, the fidelity of monte calro approximation will directly and inevitably influence where that representational power is spent. This kind of neural network optimization is actually prevalent in reinforcement learning (RL) and the problem described here bears similarity to the so called exploration-exploitation dilema, which has no obvious remedies [102]. For variationally optimized wave functions, as a scientific application, we aim however at the most extensive, but at the same time, detailed sampling resolution with an aim of faithfully capturing the electronic correlations, which in practical terms means maintaining sampling precission even in the most constrained regions of Ψ in order to achive high agreement of its nodal hypersurface with reality, so crucial for fermionic systems [17, 41].

After the somewhat naive MCMC method,*Hamiltonian/Hybrid MC* (HMC) [103] is usually the first resort, because thanks to the energy-conserving property of Hamiltonian dynamics and the usage of sympletic integrators (see appendix A), one can successfully follow the *flow* of the sympletic manifold [1] associated with a specially crafted Hamiltonian, generating in the

[1] also referred to as the *phase space*

process even distant proposals in the state/configuration space, yet at very high acceptance rates. The process described requires, computation of the gradients of the Hamiltonian function corresponding to the negative log probability density of the desired distribution, therefore making some assumptions on the smoothness and differentiablity thereof. The Born probability distribution $|\Psi|^2$ for multi-electron systems is however, infested with numerous *cusps* - points with discontinuous derivative in places where two or more electrons would coincide - and therefore not quite well suited even for HMC. To this end, Liu and Zhang [104] proposed a version of HMC, called *Quantum-Inspired HMC* (QHMC), with randomized mass matrix guided by the Heisenberg's uncertainty relations. The approach can be implemented at almost no extra cost compared to the standard HMC yet provides substantial improvements in the cases considered and is easily combined with other HMC improvements. In particular, with the *Riemannian Manifold HMC* proposed by Girolami et al. [105], which provides dynamic adaptation mechanisms essential when sampling from strongly correlated and highly dimensional target densities circumventing the need for costly pilot runs required in standard HMC, which in the end can only provide static adaptation. In what follows, we describe the methods mentioned and present their advantages on representative toy problems before benchmarking their efficiency when applied to sampling the many-electron wave function in section 3.2

Throughout our tests we have used both, qualitative and quantitative analysis of the sampling algorithms we studied. Considering the latter we have mostly relied on the *effective sample size* (ESS) per computation time, as a relevant metric [37, 106]. The rationale behind ESS is that since all MCMC techniques rely on a stochastic process to generate samples from a desired target distribution, their quality can be assessed on the basis of autcorrelations between subsequent realizations. In the end, if the generated samples are very close to one another, even extensive *thinning* will not redeem the fact that only a small portion of the domain has been explored. The relevant formula reads:

$$ESS = N \left(1 + 2 \sum_{i=1}^{N} \rho_i \right)^{-1} \tag{3.1}$$

where N is the total sample count and ρ_i are *autocorrelations* between subsequent samples, commonly computed in the spectral domain. To this end, one usually performs utilizes the Fast Fourier Transform (FFT) [107, 108], takes the modulo squared of the frequency spectrum, and transforms the result back into the time domain with inverse FFT. The code used in this chapter, including all the algorithms and i.a. the aforementioned utility will be made available on a github repo

3.1.1. Hamiltonian Monte Carlo

As already mentioned in sec. 2.1.3, the standard Metropolis-Hastings algorithm exhibits random walk behaviour making the exploration of high-dimensional spaces inefficient, high correlation only exacerbates the problem since obtaining reasonable acceptance rates requires using small step sizes. Proposing smaller transitions, however, solves the problem only seemingly, because it leads to highly autocorrelated trajectories implying small *effective sample size* and hinders sampling of multimodal distributions. Guaranteeing detailed balance and ergodicity of the chain limits what can be achieved to alleviate the mentioned problems, nonetheless, major progress has been achieved by including the gradient information of the target density. In particular, the *Metropolis Adjusted Langevin Algorithm* (MALA) uses a discretized Langevin diffusion process with an appropriately defined drift term to guide the exploration. Although some success has been reported by Scemama et al. [109] in sampling the Born probability density using this method with Ricci-Ciccotti discretization, due the further extensions we wish to implement however, we shall focus on a different approach, namely, *Hamiltonian Monte Carlo* (HMC). We motivate it also with the claim that HMC is in fact a preconditioned equivalent of MALA (under certain choice of integrators) and therefore more general [105].

The discussion of Hamiltonian Monte Carlo, or Hybrid Monte Carlo as it was originally called [110], usually starts with the introduction of the canonical equations of Hamiltonian mechanics as well as some numerical integrators designed specifically for their solution while preserving certain desirable properties. We refer such discussion to app. A and instead focus here on just how it can be used for the purpose of sampling.

Instead of directly sampling a configuration $\mathbf{q}$ Hamiltonian Monte Carlo introduces an auxiliary momentum variable $\mathbf{p}$, which is required to fully specify the energy function which defines Hamiltonian equations of motion. The connection to the target distribution, is then facilitated by the concept of *canonical ensemble* from statistical physics [111]:

$$p(\mathbf{q}, \mathbf{p}) = \frac{1}{Z} e^{-H(\mathbf{q},\mathbf{p})/k_B T} \tag{3.2}$$

where $H(\mathbf{q}, \mathbf{p})$ defines the Hamiltonian, Z is the partition function which provides necessary normalization, T is temperature and k_B is the Boltzmann constant. For general purpose, one usually disregards the $k_B T$ by setting it to unity.

Notice, however, that the above defines a joint probability distribution over the entire phase space whereas we are interested only in the distribution over $\mathbf{q}$. To achieve the desired behaviour, we partition the Hamiltonian into potential $U(\mathbf{q})$ and kinetic $T(\mathbf{p})$ energy terms and furthermore define them as follows [103]:

$$H(\mathbf{q}, \mathbf{p}) = U(\mathbf{q}) + T(\mathbf{p}) = -\log p(\mathbf{q}) + \frac{1}{2} \mathbf{p}^T \mathbf{M}^{-1} \mathbf{p} \tag{3.3}$$

where $p(\mathbf{q})$ is the target probability density function we wish to obtain samples from, and the kinetic energy term takes a quadratic form which leads to $\mathbf{p}$ having zero-mean, multivariate Gaussian distribution with a covariance matrix given by the so called mass matrix $\mathbf{M}$ (provided correct normalization is accounted for in the partition function Z). With a formulation like the one above, the variables of the joint probability density become *independent* and the momentum variable $\mathbf{p}$ accounts only for exploration, which can be additionally biased by $\mathbf{M}^{-1}$, an intelligent choice thereof is however no easy task.

The sampling process then occurs as follows, first, provided an initial position $\mathbf{q}$ is already given, a momentum variable is drawn from its Gaussian distribution - notice that this already determines the total energy value $H(\mathbf{q}, \mathbf{p})$. Then a number of steps according to discretized Hamiltonian dynamics is performed:

$$\frac{d\mathbf{q}}{dt} := \frac{\partial H}{\partial \mathbf{p}} = \mathbf{M}^{-1}\mathbf{p} \qquad \frac{d\mathbf{p}}{dt} := -\frac{\partial H}{\partial \mathbf{q}} = -\nabla_{\mathbf{q}} U(\mathbf{q}) \tag{3.4}$$

whereby for numerical solution of the above system of equations one usually uses the sympletic, second-order *Störmer-Verlet* method, also called *Leapfrog*, covered in appendix A, eq.A.25. Notice also, that in order to simulate Hamiltonian dynamics the target probability must be continuously differentiable.

The above dynamics generate a new proposal state $\mathbf{q}'$ (and momentum $\mathbf{p}'$) which then, just like in the standard Metropolis algorithm (see sec. 2.1.3), gets accepted with probability:

$$A(\mathbf{q}'|\mathbf{q}) = \min\{1, \frac{p(\mathbf{q}')}{p(\mathbf{q})}\} = \min\{1, \exp\big(-H(\mathbf{q}', \mathbf{p}') + H(\mathbf{q}, \mathbf{p})\big)\} \tag{3.5}$$

and the process repeats. Here, two properties of the Hamiltonian dynamics become particularly prominent, due to *time reversibility*, the Hastings correction is not needed - the proposals are symmetric, and more importantly, due to *conservation of energy* the new proposal will be accepted with almost 100% probability since $\frac{p(\mathbf{q}')}{p(\mathbf{q})} \approx 1$, with perfect acceptance rates undermined only by the accuracy of the sympletic integrator. The former property also means Hamiltonian proposals will satisfy the *detailed balance* which as already covered, is a sufficient condition in order for a MCMC process to converge to the desired target PDF [103], the later, on the other hand, indicates the importance of random momentum resampling, without it HMC would only sample constant energy hyper-surfaces (if simulated exactly) i.e. regions of the target distribution with equal probability density.

The genius of HMC resides in how it utilizes a random walk in momentum space to cover the entire spectrum of energy - that is, probability density - while exploiting efficient exploration of the state space degeneracy - that is, the ensemble of states with same energy - proposing nearly independent samples. Still, it suffers from several problems:

- it requires computation of gradients of the potential energy term which is costly but can also be problematic when the target distribution has some form of "spikes" or "cusps",
- typically HMC will be ergodic i.e. it will not get trapped in any subset of the state space [103, 104], nonetheless it might be hard to explore multimodal distributions by resampling the momentum only,
- it introduces additional hyper-parameters controlling the Hamiltonian dynamics, the step size ε, number of steps L and mass matrix M, with no obvious way to tune them [103, 105, 112]

Much can be achieved by appropriately tuning the metric M but we postpone its discussion to later sections, focusing now on more straightforward approaches and heuristics for the purpose of tuning ε and L. When introducing MCMC we mentioned that in order to avoid large rejection rates, the standard deviation σ of the proposal distribution ought to be chosen

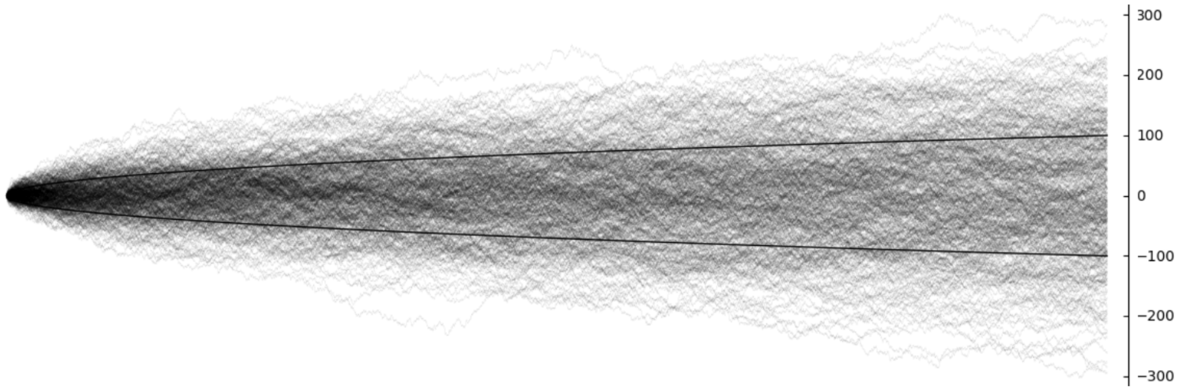

Figure 3.1.: Illustration of the slow, diffusive exploration of space with a 400 standard normal random "walkers" over 10000 steps. Horizontal axis represents time, the bold solid line indicates the ±1σ bounds given by equation 3.6

on the scale of the width of the target distribution in the most constrained dimension, that is, the one corresponding to the *smallest singular value* of its covariance matrix σ_{min}. Notice however, how it influences the ability of the MCMC sampler to explore the remaining dimensions which might vary on scales, orders of magnitude larger than σ_{min}. Due to its random walk behaviour, the typical distance covered after L steps, will only scale as $\sqrt{L}$, in accordance with the formula below:

$$\sigma_{\hat{X}}^2 = \mathbb{E}\big[\hat{X}^2\big] = \mathbb{E}\Big[\Big(\sum_i^L X_i\Big)^2\Big] = \mathbb{E}\Big[\sum_i^L X_i^2 + 2\sum_{i,j>i}^L X_i X_j\Big] = L\sigma_X^2 \tag{3.6}$$

where all $X_i \sim \mathcal{N}(0, \sigma_X^2 I)$ are the random variables corresponding to the steps taken under the proposal distribution of MCMC and the expectation over the sum of mixed terms $X_i X_j$ vanishes because they are all i.i.d.

In HMC, the general heuristic on the choice of step size on the order of σ_{min} also holds, however here due to considerations of the accumulated error in energy caused by discretization of Hamiltonian dynamics - which eventually determines the acceptance rate from eq. 3.5. When the Hamiltonian is separable, as it is often the case in HMC, the leapfrog integrator becomes explicit and therefore only *conditionally stable*, considering a dummy Gaussian target distribution with particular σ, that condition for stability becomes $\varepsilon < 2\sigma$ [103], motivating the $\varepsilon \approx \sigma_{min}$ rule. Still, since the steps taken by HMC tend to follow one direction, the effective distance covered will scale proportionally to the number of steps taken and the advantage of that becomes very apparent when sampling from high-dimensional, highly-correlated distributions. To validate that experimentally, we consider a common, 100-dimensional multivariate Gaussian benchmark with the square roots of the eigenvalues of the covariance matrix chosen as $0.01, 0.02, ..., 0.99, 1$ - the simplest way to construct it, is with a diagonal matrix with entries equal to the square of those values. We chose the step size $\varepsilon \approx 0.01$ in accordance with what has just been discussed and sufficient $L \approx 100$ to match the total HMC path length εL with σ_{max}. Precise values of the parameters have been chosen in accordance

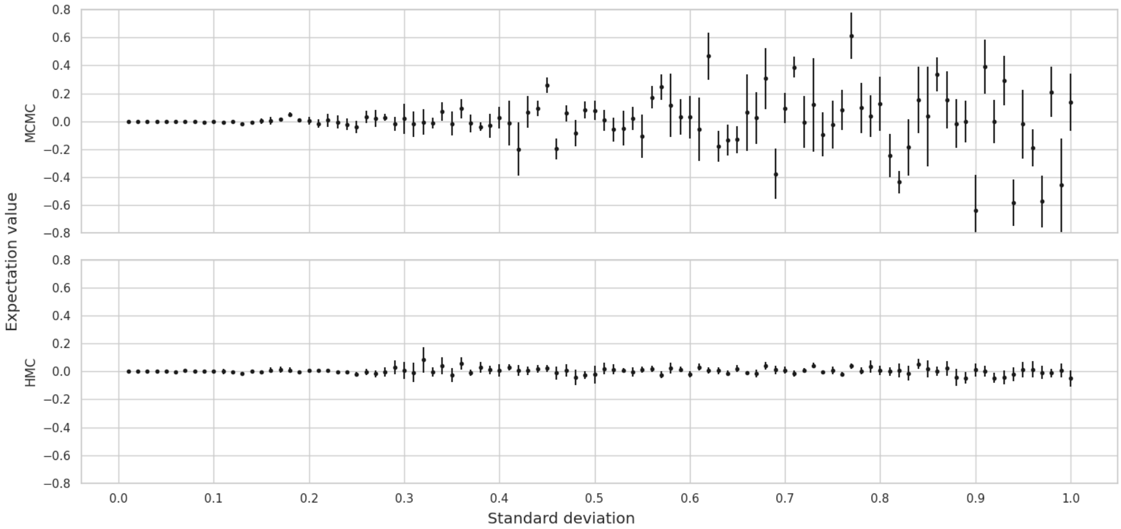

Figure 3.2.: Comparison between MCMC and HMC of sample expectation values for each dimension of a 100-dimensional Gaussian distribution. Standard deviation of the proposal distribution of MCMC iteration was chosen uniformly from the interval [0.0176, 0.0264], whereas for HMC, the leapfrog stepsize was sampled from the interval [0.0104, 0.0156] for each trajectory (not iteration) anew. The number of leapfrog steps was $L = 150$ and the number of MCMC iterations was also scaled accordingly by L to roughly match the effective path length with σ_{max} under the step sizes chosen. The acceptance rates were 0.2477 and 0.6610 for MCMC and HMC respectively which is very close to the optimum and the single-thread runtime for the latter was roughly 2.5 times greater, it is however mostly due to inefficient gradient computation in our particular implementation. It is clear HMC produces superior results, avoiding the inefficiency of random walks in exploring high dimensional distributions which led to insufficient coverage and therefore poor estimate in the case of MCMC.

with an equivalent experiment by Neal et al. [103] and have been stated, together with the results in figure 3.2.

The step sizes of MCMC and HMC have significantly different meaning, the former regulates a random walk whereas the latter, discretization size of deterministic differential equation. This deterministic nature can lead to certain undesirable artifacts like *periodicity* of sample proposals if the total path length generated by leapfrog updates happen to coincide with the scale of the distribution in particular dimension. To avoid that, a common practice is to let the step sizes vary from trajectory to trajectory [103, 106] - so called *step size jittering*. To illustrate its importance we conceived a pathological example with fixed leapfrog stepsize for a 2D Gaussian distribution illustrated in figure 3.3. The expectation estimates with one sigma confidence intervals for a) MCMC sampler, b) HMC sampler with constant step size and c) HMC sampler with random step size were [-0.0026 -0.0443] ± [0.0163 0.0724], [0.0011 -0.0217] ± [0.0097 0.03705] and [0.0095 -0.0221] ± [0.0097 0.02867] respectively with corresponding acceptance rates 0.6482, 0.7557 and 0.8208. At a first glance, these values do not indicate anything wrong about the constant-size HMC estimation which, however, upon qualitative investigation are obviously wrong. Periodicity therefore, may pose a serious threat, reminiscent of the famous Anscombe's quartet [113], which might be hard to discover

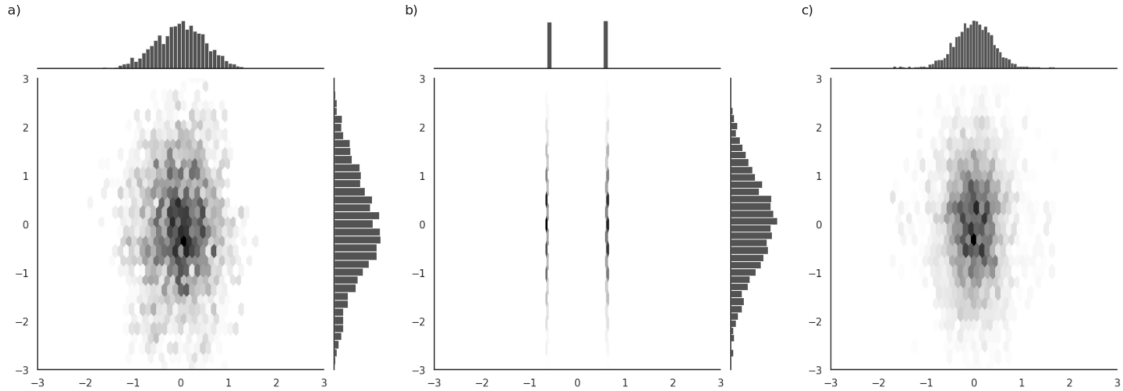

Figure 3.3.: Histograms of 5000 samples for a 2D Gaussian distribution with $\sigma_{1,1} = 0.5, \sigma_{2,2} = 1, \sigma_{1,2} = \sigma_{2,1} = 0$. From the left: a) MCMC sampler, b) HMC sampler with constant step size and c) HMC sampler with random step size. In figure b) the fixed, total path length matched with the standard deviation of the horizontal axis in such a way that the leapfrog trajectories always ended up bouncing periodically from one side to the other, producing very unreliable samples.

post-factum, when all we have is the numerical estimation of expectation value of the highly dimensional target pdf. In fact, the loss in accuracy around $\sigma = 0.3$ in the previous experiment fig. 3.2, despite random step size, could be attributed to exactly this phenomena but we can claim so only because we know the exact scales of this synthetic problem.

Crucially however, in practical situations we often do not have apriori knowledge about the scales at which the target distribution may vary, a common practice is perform few preliminary ("pilot") runs and monitor the influence of hyperparameters on various convergence and sample quality indicators and tweak them accordingly. In fact, this process can be automated, and performed in an online learning fashion throughout entire Monte Carlo simulation with stochastic approximation techniques, provided sufficient care is taken to preserve the ergodic properties of the Markov Chain. Indeed hyperparameters like the step size ε, number of steps L or even the mass matrix M can be learned from the samples generated during the sampling process, if only one ensures the values of those parameters depend less and less on the recently visited states of the chain - a process known as *vanishing adaptation* [114]. In particular Andrieu and Thoms [114] proposed using the well known Robbins-Monro algorithm [40] for the purpose of MCMC step size adaptation in which the acceptance probability objective is optimized against ε to achieve certain desired target acceptance A^*:

$$\varepsilon_{t+1} = \varepsilon_t + \eta_t(A_t(\varepsilon) - A^*) \tag{3.7}$$

where $A_t(\varepsilon)$ denotes the acceptance probability for t-th sample of the chain defined as in eq. 2.30 and $eta_t = t^{-\kappa}$, $0.5 > \kappa \geq 1$ is the learning rate schedule satisfying the usual convergence criteria. Acceptance rate is a good optimization objective, because there usually exist well founded analysis of its optimality based on the minimization of cost to obtain independent sample, for example for MCMC $A^* = 0.23$, whereas for standard HMC $A^* = 0.65$ [103]. In the context of HMC, Hoffman and Gelman [112] proposed an improved update scheme, so called *dual averaging method*, based on the Nestrov's primal-dual algorithm in which, in contrast to the above formulation, the most recent samples are given most importance, which lets it adapt to the samples coming from the equilibrium region of the target distribution rather the ones obtained in the transient stages of sampling. Notice however, that these approaches are well founded only provided the autocorrelation between samples of the chain are small, at best when the samples are i.i.d.

Also autocorrelation, guides the choice of the number of steps L during leapfrog integration, too short paths will produce essentially random-walk-like behaviour. Too long paths however, except from obvious computational overheads - and perhaps sizable error propagation - might also lead to the trajectory reversing its direction, as it can be observed e.g. in figure 3.4. Once that happens we start visiting regions we could have reached with just a fraction of steps taken. To overcome this, Hoffman and Gelman, in the same paper, introduced an adaptation mechanism for L called No-U-Turn (NUTS) in which adaptation is achieved by a recursive algorithm that follows the energy level set to a turning point, doubles that path for time-reversibility reasons, and finally samples a point along that path with weights proportional to the target density. This unfortunately, introduces a very blatant bottleneck when it comes to practical implementations, namely, since each path requires different amount of computation, the parallelization is only possible within SPMD paradigm which means we cannot efficiently leverage vectorization or GPU computing. This has been overcome more recently by Hoffman and Suntsonov [115], which proposed a completely tuning free (generalized) HMC, designed to make good use of SIMD hardware accelerators allowing most chains to be updated in parallel in each iteration.

Returning to step adaptation mechanisms, one particularly important aspect have not been taken into account, the condition on stability of the leapfrog integrator, and hence step size, might vary from one place to another. The "Neal's funnel" distribution [116], is the best benchmark for this phenomena:

$$p(\mathbf{x}, v) = \prod_{i=1}^{n} \mathcal{N}(x_i|0, e^{-v})\mathcal{N}(v|0, 3) \tag{3.8}$$

It emulates many pathological features of popular distributions like those arising in hierarchical Bayesian or latent variable models and inevitably, also the Born distribution $|\Psi|^2$ e.g. along intersections of nodal hypersurfaces. Additionally, it automatically provides a simple diagnostic for the bias in sampling through the marginal distribution of v which by construction is given as $\mathcal{N}(0, 1)$. Figure 3.4 illustrates leapfrog trajectories with four different step sizes on a two dimensional version of the funnel given by $p(x, v) = \mathcal{N}(x|0, e^v)\mathcal{N}(v|0, 1)$. It seems obvious, some sort of adaptivity in step size *over space* is required to avoid divergence in highly constrained regions, this brings us to the mass matrix.

3.1.2. Riemannian HMC

Recall the general form of the Hamiltonian function for Hamiltonian Monte Carlo:

$$H(\mathbf{q}, \mathbf{p}) = U(\mathbf{q}) + T(\mathbf{p}) = -\log p(\mathbf{q}) + \frac{1}{2}\log\Big((2\pi)^D|\mathbf{M}|\Big) + \frac{1}{2}\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p} \tag{3.9}$$

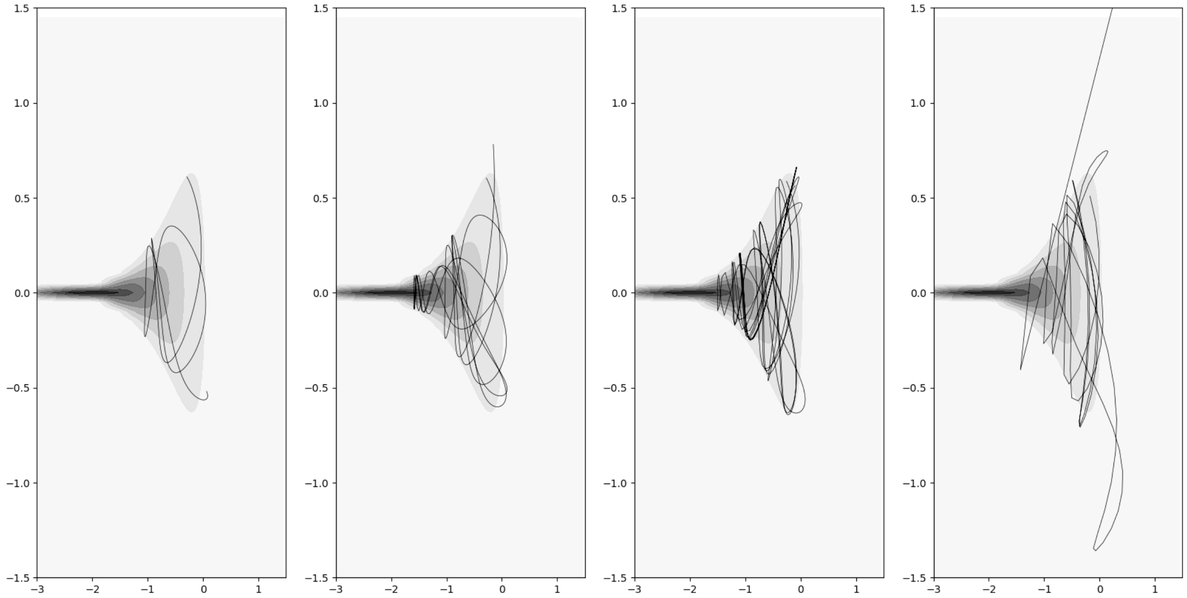

Figure 3.4.: Hamiltonian trajectories integrated with leapfrog method for a two dimensional "funnel", horizontal axis corresponds to v and vertical to x. Starting from the left, experiments were performed with step sizes: 0.02, 0.05, 0.1 and 0.25 respectively, all trajectories were initialized with the same state and momentum and iterated over 500 steps. It is clear from the last figure, how once (and if at all) the trajectory reaches a highly constrained region, the chosen step size might turn out too be no longer within stability bounds and the trajectory diverges.

Compared to the earlier definition there is an additional normalization term $\frac{1}{2}\log\left((2\pi)^D|\mathbf{M}|\right)$ which completes the analogy of the quadratic form of the kinetic energy to a Gaussian distribution of momenta upon exponentiation e^{-H} as explained previously. The term is usually omitted since it is just a constant and can be incorporated into the overall normalization factor of the joint distribution $p(\mathbf{q},\mathbf{p})$ which HMC does not require either way. Situation changes however, when M varies with $\mathbf{q}$ which is what we are about to postulate as a remedy for sampling from highly correlated distributions.

If for a moment we assume our target distribution is a multivariate Gaussian with covariance matrix $\mathbf{\Sigma}$, that is, the potential energy is a quadratic form:

$$U(\mathbf{q}) = \frac{1}{2}\mathbf{q}^T\mathbf{\Sigma}^{-1}\mathbf{q} \tag{3.10}$$

then, an estimate of its covariance matrix $\tilde{\mathbf{\Sigma}}$ could be utilized in order to unscale and uncorrelate the individual dimensions of the distribution by accordingly transforming either $\mathbf{q}$ or $\mathbf{p}$. Before one can proceed with any linear transformations, however, it is necessary to ensure they do not modify the overall dynamics. Indeed, transforming the configuration $\mathbf{q}$ with some non-singular matrix $\mathbf{A}$, is only allowed as long as a corresponding transformation of momenta with $\mathbf{p}$ with $(\mathbf{A}^T)^{-1}$ also takes place [103]. In such a case the potential and kinetic energies expressed in terms of the new variables $\mathbf{q}'$ and $\mathbf{p}'$ read:

$$U'(\mathbf{q}') = U(\mathbf{A}^{-1}\mathbf{q}') + \log|\mathbf{A}| \qquad T'(\mathbf{p}') = T(\mathbf{A}^T\mathbf{p}') = \frac{1}{2}\mathbf{p}'^T(\mathbf{A}\mathbf{M}^{-1}\mathbf{A}^T)\mathbf{p}' \tag{3.11}$$

where we have used the fact $U = -\log p(\mathbf{q})$ and we define $\mathbf{M}' := \mathbf{A}\mathbf{M}^{-1}\mathbf{A}^T$.

Now, provided we have access to $\tilde{\mathbf{\Sigma}}$, we can leverage the information it provides in two ways, either transforming configuration e.g. with $\mathbf{q}' = \mathbf{L}^{-1}\mathbf{q}$, where $\mathbf{L}$ is obtained from the Cholesky decomposition $\mathbf{L}\mathbf{L}^T = \tilde{\mathbf{\Sigma}}$ and keeping $\mathbf{M}'^{-1} = \mathbf{I}$ or alternatively keeping original $\mathbf{q}$ and instead taking $\mathbf{M}'^{-1} = \tilde{\mathbf{\Sigma}}$. Important for the discussion to come is the equivalence of both approaches, it can be confirmed easily by plugging $\mathbf{A} = \mathbf{L}^{-1}$ into the expression for effective mass in $T'(\mathbf{p}')$ in eq. 3.11 which yields $\mathbf{M}'^{-1} = \mathbf{A}\mathbf{M}^{-1}\mathbf{A}^T = \mathbf{L}^{-1}(\mathbf{L}\mathbf{L})^T(\mathbf{L}^{-1})^T = \mathbf{I}$ [103]. Nonetheless, in both cases, HMC should perform very well with just a handful of leapfrog steps sufficing to obtain an independent proposal because the target distribution has been effectively turned into a standard multivariate normal.

It follows that a proper choice of the mass matrix can provide beneficial preconditioning of the dynamics while at the same time preserving the equilibrium properties of the original system. Such ideas were in fact proposed in the context of molecular dynamics simulations already almost 50 years ago by Bennett [117], where replacing classical point masses m_i with a positive definite *mass tensor* M_{ij} has proven effective in slowing down the high frequency motions and speeding up the low frequency ones, increasing the overall computer time efficiency with which configuration space can be explored. The problem we are faced with could be described as an *inverse* problem of statistical mechanics, that is, we are trying to engineer a Hamiltonian, and hence a respective Markov Chain, with an ergodic distribution consistent with the desired target, but also one which is "well-behaved" for HMC simulation. In our case, that means accounting for spatial variation in scale of the target distribution, a characteristic of strong correlations, which may catastrophically affect the quality of numerical integration or even entirely miss certain regions of non-negligible contribution to the expectation value. In essence, what we are looking for is a smoothly varying, positive definite mass tensor which would locally standardize any generic distribution - what we are looking for is a Riemannian metric tensor.

The first account of Riemannian Manifold Hamiltonian Monte Carlo (RMHMC) was by Girolami et al. [105], the particular metric they have used, however, was specific to the task of Bayesian inference and in fact approximated empirically during sampling. Following the paradigm of information geometry [44], they have used the Fisher information metric $\mathcal{I}|_{\boldsymbol{\theta}}$ [2] which is a relevant Riemannian metric on statistical manifolds as we alluded to before in section 2.1.3. A typical setting is rather different though, most importantly the space over which the target probability distribution is defined is assumed to be euclidean - that holds at least for the Born probability distribution over electronic coordinates $\mathbf{x}_i$ [3]. On this account, Betancourt et al. [124, 125] provide a more principled, geometric treatment.

We should first address the fact that a covariance matrix has meaning only as long as we consider a multivariate Gaussian distribution, more precisely when the energy landscape is a convex quadratic form. Although that is rarely the case globally, in a convex neighbourhood the potential energy can always be approximated with:

$$U(\mathbf{q}) \approx \frac{1}{2}\mathbf{q}^T\mathbf{H}\mathbf{q} \tag{3.12}$$

where $H_{ij} = \frac{\partial^2 U}{\partial q^i \partial q^j}$ define the elements of a Hessian matrix. One quickly runs into troubles when the signature of the Hessian changes and the log determinant $\log|\mathbf{H}|$ becomes undefined. If the target distribution is endowed with some natural conditioning variables, like in the case of bayesian posterior, marginalizing over these variables does guarantee to yield a positive (semi)definite metric - indeed, a Fisher information metric. For general usage, Girolami [125] proposes using the *exponential map* [126] - a mapping from the space of all matrices to elements of the general linear group $GL(n)$, which in the case of a symmetric matrix like the Hessian, is isomorphic to the space of positive-definite matrices. Upon considerations of numerical stability, a particular combination of exponential maps can ensure less distortion of the spectral decomposition and additionally provides necessary regularization of small eigenvalues, which turn the Hessian into a theoretically well-behaved metric for RMHMC:

$$\mathbf{M}'|_{\mathbf{q}} = \tilde{\mathbf{H}} = \mathbf{Q}\tilde{\mathbf{\Lambda}}\mathbf{Q}^T, \qquad \tilde{\Lambda}_{ii} = \mathrm{SoftAbs}(\lambda_i, \alpha) \tag{3.13}$$

[2] Just like when talking about differential geometric concepts in the appendix A we follow here the vertical bar notation to indicate the evaluation of the metric tensor field at ` when expressed in some local coordinates. The usual parenthesis notation is reserved for the *action* of the tensor and in the case of a $(0,2)$ metric tensor it means a bilinear transformation $V \times V \to \mathbb{R}$ on elements of a vector space V.

[3] This claim is a rather imprecise one even disregarding the undertakings of unifying general relativity with quantum physics. The notion of a position vector has long been disregarded as meaningless in physics, whereas the wave function is not really a function, but rather a section of a complex line bundle over the configuration manifold. We shall however leave these concerns unaddressed and refer the reader to the literature on geometric quantization [118, 119] and geometric physics in general [120–123]

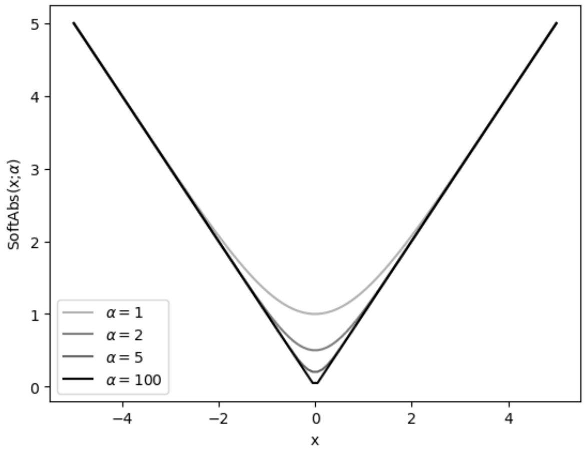

Figure 3.5.: *SoftAbs* function. The parameter α controls the "smoothness" of the approximation of the actual absolute value and in practical terms, limits the scaling of the integration step size which bears similarity to the *trust region* for step size adaptation [127]. In our experiments, setting $\alpha < 1e2$ usually provided most stable results.

with $\mathbf{Q}$ being the matrix of orthogonal eigenvectors of $\mathbf{H}$, which remain unaffected by the mapping, and $\tilde{\mathbf{\Lambda}}$ a diagonal matrix of appropriately "softened" real eigenvalues with a *SoftAbs* map defined as $\text{SoftAbs}(x_i, \alpha) = x_i \coth \alpha x_i$, see figure 3.5.

From geometric perspective, the Riemannian metric defined in this way provides information on the curvature of the graph of the potential $U(\mathbf{q})$, that is, an n-dimensional submanifold in the (n+1)-dimensional manifold with coordinates $(\mathbf{q}, U(\mathbf{q}))$. Hamiltonian evolution in such a setting locally parallels its geodesics [121], which could be given as a reason for the superior efficiency of RMHMC approach, which on the first sight may not be obvious. In the end, the usefulness of this technique is limited by the computational cost of matrix operations which introduce a number of complications, so it must be handled carefully. A primary concern is obtaining partial derivatives with respect to positions and momenta for the Hamiltonian evolution, which now requires third derivatives of the potential function. Provided sufficient continuity, and with a careful numerical treatment [125], the total computational effort can be kept at $\mathcal{O}(N^3)$, matching the cost of the eigendecomposition of the Hessian. Furthermore, dependence of $\mathbf{M}'$ on position variables means the Hamiltonian is no longer separable $H(\mathbf{q}, \mathbf{p}) = U(\mathbf{q}) + T(\mathbf{q}, \mathbf{p})$ and the leapfrog integration becomes implicit, requiring fixed point iteration for both $\mathbf{p}_{n+\frac{1}{2}}$ and $\mathbf{q}_{n+1}$ c.f. eq. A.24. Still, despite its complexity, the method does produce correct results which can be quickly asserted on a two dimensional Gaussian, see fig. 3.6.

Before investigating the viability of RMHMC on more challenging benchmarks let us consider a particular modification of the presented approach, which constitutes our small contribution. As mentioned before, a linear transformation of the variables $\mathbf{q}$ or $\mathbf{p}$ can occur in different ways [103], choosing to modify just the momenta by a spatially varying mass matrix is just one possibility. Nothing prevents the opposite, that is, a transformation of the position $\mathbf{q}$. Such approach would have the big advantage of keeping the leapfrog integration explicit

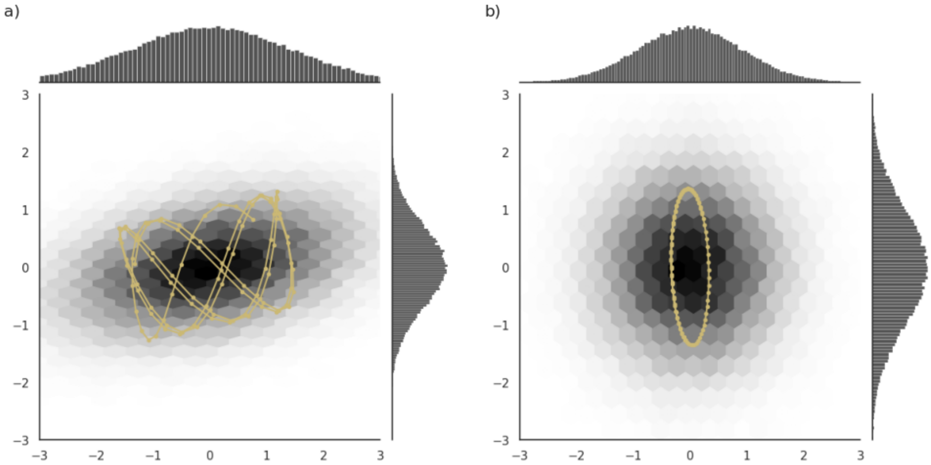

Figure 3.6.: Comparison of a) "euclidean" and b) riemannian HMC on a two dimensional Gausian distribution with non-isotropic covariance. Plots show histograms for 100,000 samples and Hamiltonian trajectories for certain random, but same initial conditions. The RMHMC method successfully standardises the distribution and as a result makes the trajectories much better behaved. Also, compared to figure 3.3, the overall quality of sampling has significantly improved, that is thanks to gains in efficiency that could not have been achieved without the just-in-time compilation of JAX. Indeed, on this two-dimensional Gaussian example, the total runtimes were 2.4s for HMC and 22,3s for RMHMC, whereas the implementation used to create the histogram in fig. 3.3 took around 5min and only for a fraction of the total number of samples.

so it is almost surprising the research community has overlooked this possibility focusing so much on finding a mass preconditioner instead. As far as our simplistic experiments went, however, we did observe the desired behaviour with improved performance and stability thanks to the removal of fixed point iteration, the transformation we propose reads:

$$\mathbf{q}' = \mathbf{A}\mathbf{q} = \sqrt{\tilde{\mathbf{\Lambda}}}\mathbf{Q}^T\mathbf{q} \tag{3.14}$$

where $\mathbf{Q}$ and $\tilde{\mathbf{\Lambda}}$ are defined as before. The kinetic energy remains unchanged i.e. takes the simplest possible form $T'(\mathbf{p}') = T(\mathbf{p}) = \mathbf{p}^T\mathbf{p}$ and distribution of $\mathbf{q}'$ becomes locally standard normal which follows directly by construction from equations 3.12 and 3.11:

$$\begin{aligned} U'(\mathbf{q}') &= U(\mathbf{Q}\sqrt{\tilde{\mathbf{\Lambda}}^{-1}}\mathbf{q}') + \log|\sqrt{\tilde{\mathbf{\Lambda}}}\mathbf{Q}^T| \\ &\approx \frac{1}{2}\mathbf{q}'^T\sqrt{\tilde{\mathbf{\Lambda}}^{-1}}\mathbf{Q}^T\mathbf{Q}\tilde{\mathbf{\Lambda}}\mathbf{Q}^T{}^T\mathbf{Q}\sqrt{\tilde{\mathbf{\Lambda}}^{-1}}\mathbf{q}' + \log|\sqrt{\tilde{\mathbf{\Lambda}}}| \\ &= \mathbf{q}'^T\mathbf{q}' + \sum_i \log\tilde{\lambda}_i \end{aligned} \tag{3.15}$$

where we have dropped the $\frac{1}{2}$ factor in the last line as well as leveraged the fact the determinant of an orthogonal matrix is unity, its inverse is equal to its transpose and a transpose of a diagonal matrix is equal to itself.

What remains to be shown is the viability of RMHMC, for this purpose we turn back to the funnel example. Figure 3.7 presents the samples and autocorrelations of the v variable obtained with MCMC, HMC and RMHMC algorithms respectively for a 20-dimensional funnel as given by eq. 3.8. For the experiment we have used a diagonal version of RMHMC, that is, with diagonal Hessian computed efficiently in JAX utilizing the jacobian-vector-product operation on a gradient vector. All runs were performed for the same number of samples and we have used dual averaging for HMC and RMHMC to establish the best step size automatically, for MCMC it has turned out detrimental as it would decrease the step size virtually to zero as the sampler got stuck in the narrow region and could not leave without access to any directional information. For diagnostic purposes, autocorrelations up to 1000 steps back (lags) were computed with a Fast Fourier Transform as described at the beginning of this chapter and plotted together with their 95% confidence intervals. A descriptive metric for comparing different chains is the lag for which the autocorrelation markers cross with the lines of their confidence interval bounds, upon investigation it is obvious RMHMC produces virtually decorrelated samples whereas MCMC suffers from the random walk behaviour as expected.

Table 3.1 compares all three approaches in a more quantitative manner. In the first column we list the expectations of the v variate together with a 95% confidence bounds on the estimate, since the true expectation equals exactly zero, this value provides a good measure on how biased a sampler is, that is, how well can it explore the highly constrained region of the funnel. Although MCMC manages to do so reasonably well, it does so at the expense of very high autocorrelation. Furthermore, its high acceptance rate is an indicator that the random walk leaves large regions of the configuration space largely unexplored, so even the ESS metric should not be trusted too much in this case. With HMC, we can once again observe how too large step sizes lead to certain regions of space being completely missed, this manifests itself in the high bias of the estimation of v. The step size could not have been chosen differently though, it has been stochastically optimized to meet the desired acceptance rate on the basis of an average trajectory, if we forced it to be smaller the behaviour in all variates other than v would have been essentially turned into a random walk. Finally, RMHCM seems to perform best, both in terms of ESS per computation time and estimate of the expectation, the remaining bias, we believe, can be attributed to a relatively large alpha parameter for the *SoftAbs* function, $\alpha = 10$, which was necessary to obtain stable trajectories for our implementation.

Although the comparison based on efficient sample size per computation time clearly favours RMHMC, it must be noted that with growing dimensionality this advantage might suffer from its cubic asymptotic complexity compared to MCMC but also HMC, Betancourt [125] however does report similar results for a 100-dimensional case. Despite the potential efficiency gains we found the RMHMC method can be very fragile, that is because instabilities may arise from very different mechanisms and it is unclear how to deal with them in an automated manner. This fact is also a reason for a rather poor reception of this approach in probabilistic programming languages (PPL) like Stan [106]. In the next section we will briefly cover another, much more straightforward extension of Hamiltonian Monte Carlo, which we have empirically confirmed to have a stabilizing effect on RMHMC.

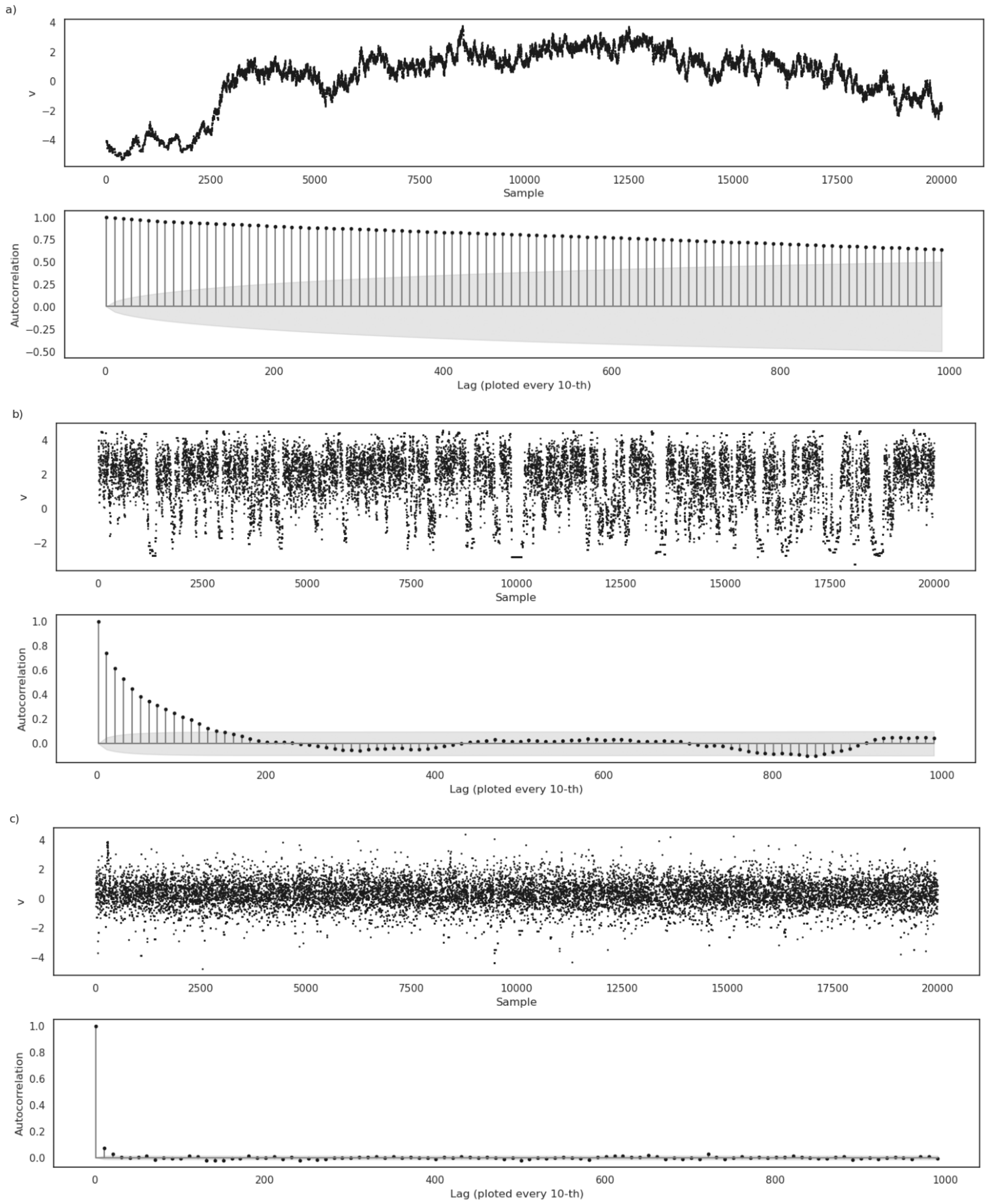

Figure 3.7.: Scatter plots of the v variate of a 20-dimensional funnel together with corresponding autocorrelations for a) MCMC b) HMC and c) RMHMC. The remaining bias in figure c) does probably arise from the inaccuracy of reparametrization in the almost singular narrow neck region of the funnel. Quantitative results can be compared in table 3.1.

Algorithm	Expectation	Final step size	Acceptance rate	ESS	ESS/s
MCMC	0.2893 ± 0.0283	0.1	0.8092	96.81	26.5938
HMC	1.4181 ± 0.0258	0.3	0.6772	2051.6	70.7448
RMHMC	0.0745 ± 0.0137	0.1	0.9101	18234.1	246.4068

Table 3.1.: Result comparison for a 20-dimensional funnel benchmark. All runs were performed for $2 \cdot 10^4$ samples with $5 \cdot 10^3$ initial warm up iterations. In spite of increased computational effort of RMHMC, comparison of effective sample size per second reveals its superiority in practical terms. The initial step size for dual averaging was 0.1 for all cases and the target acceptance rates 0.25, 0.65 and 0.85 respectively in accordance with the conditions discussed earlier in the text and empirical recommendation for higher acceptance rates for RHMMC by Betancourt [125].

Algorithm	Final step size	Acceptance rate	Expectation
HMC, M=1	0.5375	0.5802	3.8535 ± 0.070
HMC, M=2	1.2064	0.5767	11.4073 ± 0.0676
HMC, M=0.5	0.0657	0.5122	-4.2613 ± 0.0313
QHMC	0.2798	0.6325	2.0012 ± 0.0621

Table 3.2.: Comparison of results for a 2D funnel distribution obtained with HMC for various values of mass and D-QHMC. All runs were performed for 50 leapfrog steps with the initial step size $\varepsilon_0 = 0.25$ which has been adjusted with dual averaging with the same target acceptance $A^* = 0.65$ during the simulation. Runtimes for all methods with 1000 burn-in steps and 20,000 samples were virtually equal $\approx 5s$

3.1.3. Quantum-inspired HMC

Proposed by Liu and Zhang [104], Quantum-Inspired Hamiltonian Monte Carlo (QHMC) promises to improve the performance of HMC when sampling from "spiky" and multimodal distributions by leveraging the energy-time uncertainty relations of Quantum Mechanics. Their key contribution is allowing the mass of the virtual particle in phase space to admit random values rather than having fixed value throughout the simulation, which could be seen as employing varying simulation time scales.

In effect, this modification finds a good trade-off between extended exploration of the configuration space facilitated by simulating the dynamics of a light particle and a slow and careful probing of highly constrained regions when using a heavy particle c.f. equation 3.4. Although interestingly, using dual averaging we have observed an exactly opposite behaviour, that is, on a 2D funnel benchmark using small mass led to exploration of its narrow neck whereas simulation with large mass explored mostly its flat region. This, we believe, can be attributed to the interplay of the step size and mass in the explicit leapfrog integration when step size is not fixed, eq.A.25. Since mass appears only in one out of three operations per iteration its effects gets overcompensated by step size, this can be readily confirmed upon investigation of the final step sizes for each case presented in table 3.2.

In its last column we present expectation values of the v variable which is indicative of the bias in sampling, the indicated bounds correspond, again, to 95% confidence intervals. The differences between different mass settings are significant and discernible with naked eye,

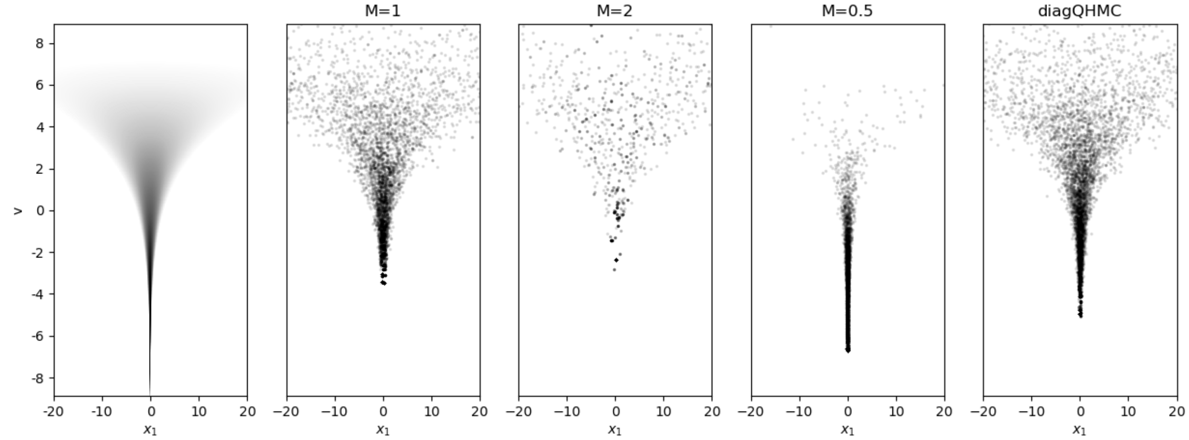

Figure 3.8.: Samples from a 2D funnel distribution obtained with HMC for various values of mass and QHMC with anisotropic but uncorrelated mass distribution. For visual clarity the sample count has been thinned to every fifth. Qualitative results were provided in table 3.2 but it is obvious QHMC offers smaller bias thanks to its improved penetration of the narrow neck of the funnel.

c.f. fig. 3.8 but perhaps surprisingly in neither case did we observe random walk behaviour as it was the case with MCMC, fig. 3.7 a). Even for $m = 0.5$ the Hamiltonian dynamics remained unaffected and well conditioned, they just occurred at a much smaller scale. In the rightmost plot of the aforementioned figure we see samples from the implementation of a random, diagonal mass matrix, dubbed originally D-QHMC [104]. The mass values were chosen according to the distribution $\log m_i \sim \mathcal{N}(-0.1, 0.3)$ to maintain positive definiteness. In higher dimensions we did not observe the robustness to the choice of μ_i and σ_i parameters Liu and Zhang allude to, hence the restriction to this particular distribution with expectation value approximately equal to unity. Importantly, the mass resampling process occurs for every trajectory, not the leapfrog step, implementing the latter would lead to a random walk.

QHMC promises also to be useful for multimodal distributions thanks to an analogy to quantum tunnelling. Quantum objects can climb over potential barriers even with an insufficient total energy because of inherent uncertainty as postulated by the Heisenberg's relations $\Delta q \Delta p \propto \hbar$ or equivalently $\Delta E \Delta t \propto \hbar$. In relation to Plank's time and an interpretation of step size ε as a "quantized" unit of time one can rewrite these relations to $\Delta q \Delta p \propto \varepsilon$ and since momenta are Gaussian distributed like $\propto \exp\left(-p_i^2/2m_i\right)$ their uncertainty is given by $\sqrt{m_i}$ and the uncertainty in position therefore $\Delta q_i \propto \varepsilon/\Delta p_i = \varepsilon/\sqrt{m_i}$.

Considered generally, if isolated modes are found to exist, one cannot naively combine the results of various HMC runs because they are each confined to just a single mode, there is therefore no principled way in which one could perform such combination without introducing an "inter-modal" bias [103]. A traditional way of dealing with multiple modalities is achieved with *tempering*, which in analogy to statistical mechanics varies the temperature parameter from the original formulation of the potential energy function from eq. 3.2. This modification allows for larger fluctuations in potential energy - and hence probability density - beyond random momentum resampling, which by itself introduces only a $\approx N/2$ variation

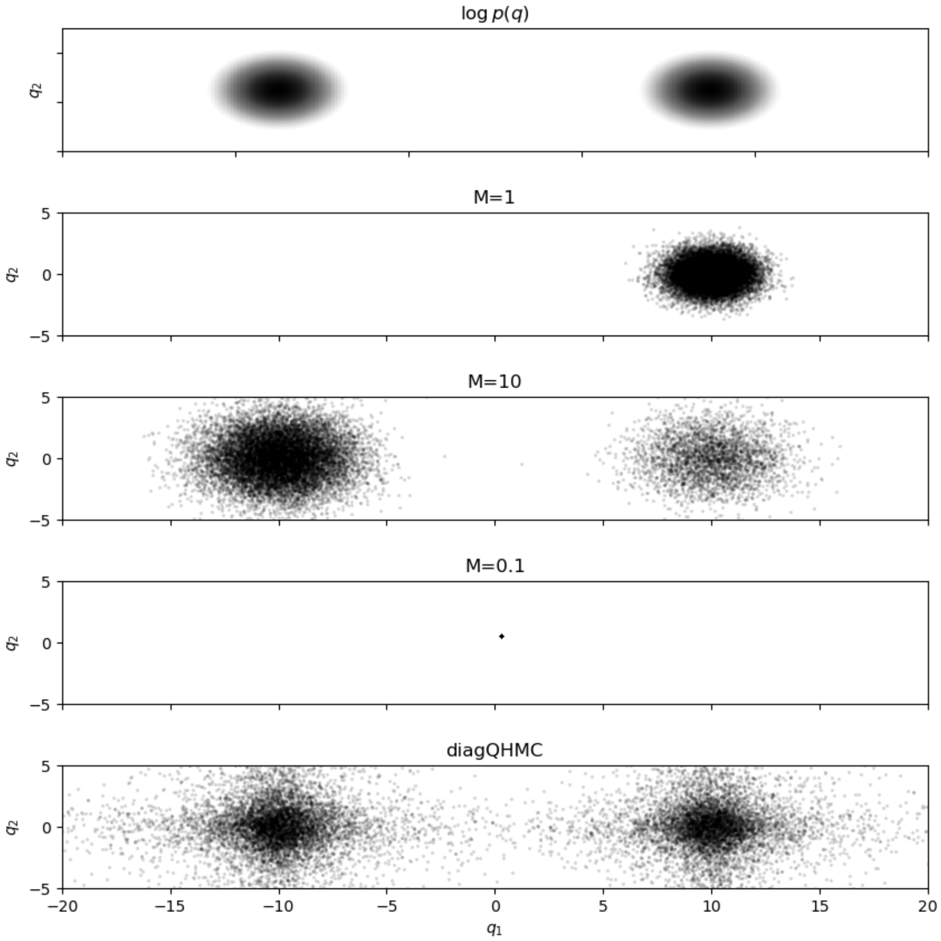

Figure 3.9.: Comparison of sample distributions between HMC with various mass values and the diagonal version of QHMC for a mixture of two Gaussians. The apparent, increased spread of QHMC samples compared to the density plot, we recon, is just a visual artifact of plotting the density with a specific cutoff.

since $\mathbf{p}^T\mathbf{M}^{-1}\mathbf{p}$ is a χ^2 variate. As described, QHMC utilizes a slightly different mechanism compared to tempering, but it nonetheless induces the same behaviour and in end effect even produces arguably better results as measured by the Wasserstein distance between the multimodal sample distribution [104].

Figure 3.9 contains scatter plots of samples obtained with QHMC and HMC for various constant masses. Again, as far us our tests went, we've observed best behaviour when the mass expectation value was kept around one. In the following section, we will provide more details on our final implementation utilizing both QHMC and RMHMC adjustments and test its performance as a sampling subroutine in PauliNet.

3.2. Sampling the PauliNet wave function

As already mentioned at several occasions, we have chosen to implement our algorithms in JAX [128]. This way rather than being tied to any probabilistic programming language (PPL) like PyMC, Pyro or Stan we could have focused solely on the design and performance of the sampling algorithms themselves. Moreover, since the negative log probability density function - as defined by the PauliNet ansatz - has also been implemented in JAX it was very straight forward to insert our code as its subroutine. The strengths of JAX further include its support for reverse- and forward-mode differentiation [129], which was required for simulating Hamiltonian dynamics, its just-in-time compilation into highly optimized computational XLA (Accelerated Linear Algebra) kernels, and automatic parallelism, both at *instruction and processor level* - that is, the SIMD (vectorization) and SPMD paradigms [130, 131] - which we have made extensive use of thanks to the fact that sampling is in general *embarrassingly parallel*.

Regarding our methodology, we motivate combining the QHMC and RMHMC methods as follows. The Born probability distribution is highly multimodal and infested with very constrained regions similar to those which we extensively studied on the funnel benchmark. Although QHMC could seemingly deal with both of these issues alone, we have seen in the previous section that particles of different masses require on average orders of magnitude different step sizes if we wish to maintain desired acceptance rates, c.f. table 3.2. This observation was only possible when the step size was admitted as another hyperparameter and optimized stochastically online, during the sampling process. This, conveys an insight about the "preferred" regions of the negative log probability landscape where particles of various masses tend to reside. When mass becomes a random variate itself, however, even though dual averaging can be used just as before, it can only account for the behaviour of an averagely massive particle. It becomes apparent, that even though QHMC can lead to improved exploration of both, the flat as well as the constrained regions of the wave function, due to this permeability, it is even more reliant on spatial adaptivity than the baseline HMC. The combined approach we postulate should therefore lead to more accurate sampling near the singularities and nodes of the wave function - so important for capturing the elusive correlation energy - and in effect, assignment of adequate resources for their representation with the highly flexible neural network ansatz during variational optimization.

Although we have also undertaken initial research into replacing the internal electron interaction network of PauliNet with an adaptation of DimeNet - a *directional* message passing graph neural network conceived by Klicpera et al. [46, 98] - this goal has been ultimately abandoned for the following reason, which we include here with a hopefully pedagogic effect. FermiNet [56, 74] builds its wave function ansatz on the information contained within electron-nuclei and electron-electron interactions but it largely ignores the spatiality of this information with the only exception of including it in the initial embeddings. PauliNet [55] used continuous convolution adapted from SchNet [49] which enabled construction of electronic features respecting their radial environment. Later, Gerard et al. [81] successfully adapted continuous convolutions from PauliNet into a FermiNet like ansatz, their modifications did improve performance but have unfortunately occluded the spatial dependence of the filter generating function due to the usage of distance embeddings instead of distances themselves as its inputs. We hypothesised, based on the discussion on electronic correlations so far, that further incorporation of directionality of information exchange should lead to gains in accuracy. Concretely, including angular convolutional filters in addition to radial ones with a successor of SchNet - the DimeNet network - seemed promising in capturing even more correlation energy which in the end, is purely spatial. Explicit feature conditioning on the angles between triplets of electrons should have not only improved their interpretability and provided useful implicit bias, but propagating this information through subsequent layers of the network should have had in effect a gradual encapsulation of information beyond pairwise correlations. This approach shared therefore design philosophy with the CI or CC-expansions which explicitly extend the basis set with doubly-, triply-, and higher excited configurations [13] but the choice of how many electrons to consider at each particular instance would have been guided by the spatial extent of the graph convolution and in totality, by the overall number of interaction layers. An already excellent performance of PauliNet, time restrictions and the realization that when considered all together, just the pairwise, inter-electronic distances already do contain all the spatial information present in the system, prevented further investigations.

Considering all of the above, for our final experiments, we have used the PauliNet network with multiple-determinant ansatz and pretrained it for 500 iterations using its original Metropolis-adjusted Langevin sampler (MALA) before switching to our Quantum Riemannian Hamiltonian Monte Carlo (QRHMC) sampler. All default hyperparameters provided by the authors have been used and the optimization was performed with the second-order KFAC optimizer[42]. The training curves for a totality of 10,000 iterations can be seen in figure 3.10. The reason for pretraining was that considering curvature of the wave function as represented by a barely trained network could have been hardly justified and we have found it to provide very unreliable information.

The basic building block of our algorithm has been the HMC kernel, which in the spirit of functional programming, has been then decorated with the random mass and riemannian metric functionalities using the factory design pattern. This enabled an end-to-end just-in-time compilation of the entire sampling routine using the `jax.lax.scan` command, without any need for explicit for loops and therefore allowing the XLA compiler to perform low level optimizations which alone, has lead to thousandfold speedups over our initial naive implementation. Furthermore, vectorization of our kernel with `jax.vmap` on a GPU for 100 parallel chains brought additional 5x speedup and to ensure proper parallel random number generation, the JAX implementation of Threefry hash function [132] has been used. Naturally, step size jittering and dual averaging with target acceptance rate of $A^* = 0.8$ have been used, we have also decided to use only the diagonal versions of the QHMC and RMHMC, mostly for keeping the runtime manageable and stability considerations, for the same reason the number of leapfrog steps has been kept low at only 5 with appropriate NaN checks to avoid wasting computation. Our complete implementation is available online on github.

Commenting on the results we were able to obtain, it is hard to confirm whether our modifications indeed led to an improvement. Compared to the runs we have performed with

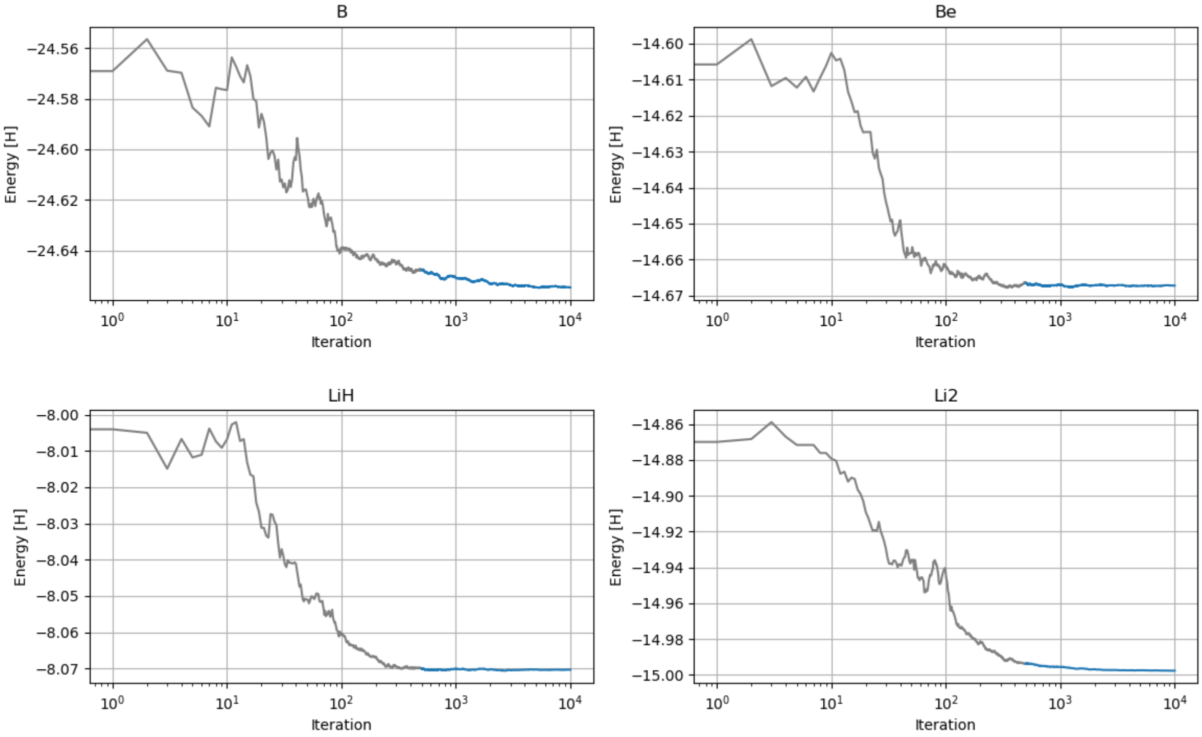

Figure 3.10.: Training curves for Boron, Beryllium, Lithium hydride and Dilithium. Depicted in gray are the pretraining steps using MALA sampler, blue are the iterations performed with our sampling algorithm. Notice, the horizontal axes have been represented in a logarithmic scale.

the MALA sampler, noticeable differences on the final energy could have been observed only for the case of Boron and Dilithium, although even those, oscillated randomly about not more than several mH from the baseline. The reasons may be plenitude, conceivably, the need for pretraining could have biased the network in an unrecoverable manner, also the performance of this particular ansatz might have not been limited by the sampling inefficiencies in the first place, of which there are some empirically supported reasons too [81]. In the end, we believe it has been the overall system interaction, which however, could not have possibly been resolved any better in the time available, which prevented observing significant improvements. It is always disappointing to report underperformance, although truthfully, our results match those published by Hermann et. al [55] which considering the novelty of our sampling methodology, should be consider a success in itself. Indeed, to the best of our knowledge it is one of the first works investigating impacts of sampling on variational optimization of wave function to date. We remain positive that further investigations along this line of research will eventually lead to meaningful progress and hopefully tighter collaboration at the intersection of statistics, artificial intelligence and the rich family of Quantum Monte Carlo methods.

4. Conclusions and outlook

In this work we have covered the topic of Quantum Monte Carlo with physics-aware, deep learning surrogates for the study of fermionic systems. Having provided an extensive state-of-the-art review of electronic structure methods, both traditional and those using machine learning, we hope to have conveyed not only the immense depth and intricacy of computational quantum chemistry and physics but also have left the reader convinced that as much as the field of deep learning can benefit the field of scientific computing, sufficient care must be undertaken and the prevailing knowledge may not simply be disregarded.

Our main contributions were restricted to the topic of sampling from the variationally optimized wave function, the importance of which becomes apparent when one considers the flexibility of this ansatz and how inappropriate coverage of the narrow valleys of the Born probability density, characteristic of strong correlations, will inevitably lead to sub optimal allocation of the network's limited representational power. Upon through investigations of the Hamiltonian Monte Carlo family of sampling algorithms, especially from the analytic mechanics point of view, we have considered a number of improvements and extensions that have made it better suited for the task at hand. Concretely, we focused on a geometrically-motivated Riemannian Hamiltonian Monte Carlo [105, 125] we have shown to offer improved performance for ill-conditioned and highly-correlated probability density functions thanks its locally-standardising effect, which in practical terms prevents catastrophic instabilities in the sympletic integration of Hamiltonian dynamics. We have further extended our algorithm by introducing randomized mass tensor in accordance with the Quantum-Inspired Hamiltonian Monte Carlo [104], which we have found not only to stabilize the geodesic flow i.e. the trajectories of the aforementioned method but also enabled better treatment of multimodalities of any particular probability density. Finally, with an efficient implementation in JAX [128], we have employed all considered improvements for sampling the PauliNet wave function ansatz [55].

Despite the ambiguity of our final results, we believe this line of research should not be disregarded and definitely deserves more attention. The reasons for the poor performance we have observed may undoubtedly lie in the particular way we have chosen to convey our numerical experiments, but may just as well point to deeper problems. In particular, we believe the AI community should pay more attention to the efforts of statisticians, especially when it comes to variational neural network training. More principled studies of how to improve the resolution of the nodal hypersurface of the wave function are definitely a topic which will benefit not only deep learning ansätze but all Quantum Monte Carlo methods in general. Detailed investigations of how the electronic correlations can be better extracted from the sampling approach itself are just as well needed, the first steps toward this goal has already been taken when we consider how Riemannian Hamiltonian Monte Carlo utilizes local correlations i.e. the curvature of the negative log probability density to guide exploration of the configuration space. Surely, a geometric quantum theory [118, 119], with an explicit treatment of the uncertainty relations between state and momenta should provide beneficial insights too.

If one considers our work in totality, what we have attempted was obtaining samples through ergodicity of classical evolution which were nonetheless meant to represent quantum reality. Although stating it this way is not completely right, it would be definitely interesting to explore whether Hamiltonian Monte Carlo algorithms could indeed reproduce quantum expectations through some sort of quantum ergodicity. Indeed, little care is taken nowadays in considering the origins of Schrodinger's wave equation from Hamilton treatment of analytical mechanics and the ideas of *stochastic mechanics* [133] - that is, treating the Born probability density as if coming from an underlying statistical process, not an underlying quality of nature - are generally not considered sound. Although very controversial, with the Copenhagen interpretation of Quantum Mechanics firmly entrenched in academic communities, many researchers have nonetheless undertaken purely computationally oriented theories of quanta of this sort [134, 135].

Lastly, let us touch on the possible benefits of wider adoption of other deep learning methods. Indeed, much success have been seen recently with autoregressive, diffusion or energy based models for generation of highly-dimensional and intricate content, if only could it be adapted to respect the antisymmetry of the wave function, orders of magnitude improvements regarding the size of the systems we can study are conceivable. Surely, attention mechanism, also seems alluring but is at the same time immensely expensive, nonetheless its ability to infer the topology of the input by itself could enable predictions way beyond mean-field approximation, capturing even all-to-all electronic interactions when needed. In effect, performance exceeding chemical accuracy should be attainable, opening up the realms of computational superconductivity research where the increased computational effort would be well justified. Meanwhile, the initial research into the regularity of wave functions and hence, the transferability of deep QMC models across quantum systems [136, 137] has already opened up a promising avenue for expanding the impact of deep learning in quantum chemistry and solid-state physics. To this end, the concepts and challenges covered in this thesis will continue to extend its impact, particularly, the tighter marriage of physical priors and symmetries with flexible neural network ansätze or efficient and scalable sampling and optimization algorithms.

Undoubtedly, the very large scale implementations of the new algorithms, sufficiently optimized and integrated into existing HPC standards and other areas of research such as *ab-initio* molecular dynamics [15, 138] might prove as much, if not more influential. Indeed, what can be gained through intelligent modelling usually corresponds to a technical improvement, either in HPC hardware or compiler optimization, of which, let the excellent performance gains we have obtained using just-in-time compilation be the best account. One can also expect quantum computation to eventually enter the arena of electronic structure calculations, provided it can overcome its accuracy and scalability issues [139]. It's conceivable, the improvements we see today in the methodology of computational quantum physics on classical computers will directly translate into better design of quantum devices which in turn will increase their utility and adoption. Ultimately, the exponential advantage promised by quantum computing may lead to a singularity-like event after which the entirety of nanotechnology, chemistry, and material science will stand wide open in front of the new generations of engineers. In the words of Feynman, "there is plenty of room at the bottom" [140], and by unlocking new insights into the behavior of quantum systems, we may gain a better understanding of the fundamental laws of nature and be able to design and engineer new materials and technologies that were previously thought to be impossible.

A. Hamiltonian mechanics

Analytical mechanics is a completely mathematical science, taking place in an abstract realm of quantities called *generalized coordinates* q^i of the so called *configuration space* $\mathcal{Q}$. This space is a in a "well behaved", bijective correspondence to the three dimensional space in which the evolution of any physical system occurs. Although any analytical approach to mechanics is, mathematically, merely a restatement of Newton's laws of motion postulated in 17th century, physically it conveys a truly *philosophical significance* almost at odds with causality. Two fundamental concepts of analytical mechanics are those of *work function* (negative potential energy) and *kinetic energy* - two *scalar* quantities fully determining motion and replacing Newton's force and momentum respectively. As proposed in the revolutionary work of Lagrange, *Mecanique Analytique*, the evolution of the system then follows from the so called *principle of stationary action* which asserts that the path taken by nature is the one which minimizes a functional of energy - so called *action* S:

$$S := \int_{t_1}^{t_2} dt\ L(q, \dot{q}), \quad where \quad L = T - U \tag{A.1}$$

where L is a *Lagrangian*, defined as the difference between the kinetic and potential energy terms for a system at hand. In Lagrange's formalism there is therefore no need for arbitrary postulates like "action equals reaction", nothing is caused by anything, everything just follows minimization principle - perhaps the closest physics has gotten to the theory of everything.

Lagrange and Euler developed a completely new branch of calculus to treat problems of this nature - the *calculus of variations* - in which a trial path can be varied between points P_1 and P_2 till a stationary path is found. Later Hamilton expanded on it allowing the variation to not be restricted to paths between points P_1 and P_2 but rather between times t_1 and t_2 extending the principle to non-conservative systems. Also the action functional has not initially been defined in the above manner, for a detailed explanation as well as historical perspective, inquisitive reader is encouraged to refer to [3].

Using the above analytical formalism it is particularly convenient to deal with any constraints, any mechanical system with N degrees of freedom constrained by m additional conditions, will translate to a study of motion of a single, free particle in an abstract, non-euclidean and $n = N - m$ dimensional space. It also provides freedom in the choice of coordinate system in which the equations of motion are represented - a recognition which eventually led physics to study natural phenomena in terms of their co- or equivariance under various transformations, indifferent to any special reference frame - of crucial importance in both, quantum mechanics as well as Einstein's relativity [120] [3]. It should be mentioned however that, such a variational treatment of physics does restrict the nature of force to only those which can be derived from a scalar work function, leaving e.g. frictional forces outside its realm capabilities.

Hamilton's canonical equations of motion can be obtained by applying *Legendre transformation* $\mathcal{TQ} \to \mathcal{T}^*\mathcal{Q}$ to the Lagrangian function L considered as the function of generalized coordinates q^i and their time derivatives $\dot{q}^i$. Such transformation introduces new variables, dual to $\dot{q}$ which we shall call *generalized momenta*, with components:

$$p_i = \frac{\partial L}{\partial \dot{q}^i} \tag{A.2}$$

For the purpose of this introduction we shall, however, follow a different path, one centered more around the geometric structure it carries, which will prove itself useful when considering the Hamiltonian Monte Carlo sampling algorithm.

As already mentioned, Hamiltonian mechanics introduces new set of conjugate variables p_j, the direct consequence thereof will be the decomposition of the second order ordinary differential equations describing motion in some potential field into a set of first order ones instead. Mathematically, to each point on the configuration manifold $q^i \in \mathcal{Q}$ we associate a covector p^i and therefore, in the language of differential geometry, the relevant geometric structure $\mathcal{M}$ of Hamiltonian mechanics is the cotangent bundle $\mathcal{T}^*\mathcal{Q}$, which is itself a manifold of twice the dimension of $\mathcal{Q}$.

Every point on any cotangent bundle $x^i \in \mathcal{M}$, by construction, can be decomposed into a direct sum of local coordinates $\{q^i, p_i\}$, to this fact, cotangent bundles attribute their sympletic structure which is characterised by existence of a closed, non-degenerate 2-form - to be defined shortly - everywhere on $\mathcal{M}$. "Sympletic", from Greek *sym-plektikos* stands for "braided together" and is a calque of its Latin equivalent *co-plexus*, meant to highlight the dual nature of $\{q^i, p_i\}$.

A.1. Sympletic geometry

A closed, non-degenerate 2-form also known as *sympletic form* is an invertible, bilinear transformation ω - or equivalently an anti-symmetric $(0,2)$-tensor - whose exterior derivative $d\omega = 0$. It arises naturally in Hamiltonian mechanics by the following consideration [124].

Consider a point transformation of the configuration space $\mathcal{Q}$ (not $\mathcal{M}$), that is, a bijective, continuous transformation of a space onto itself $f : \mathbb{R}^n \to \mathbb{R}^n$. The new coordinates Q^i obtained in that process are given by:

$$Q^i = f_i(q^1, q^2, ..., q^n) \qquad with \qquad \det\big(J_f(\mathbf{q})\big) \neq 0 \quad \forall \mathbf{q} \tag{A.3}$$

where $J_f(\mathbf{q})$ is the Jacobian matrix of f evaluated at point $\mathbf{q}$

Furthermore, consider any generic covector field, assigning a covector $p \in \mathcal{T}^*\mathcal{Q}|_q$ to every point $q \in \mathcal{Q}$:

$$p = p_i dq^i \tag{A.4}$$

where p_i are the components and the 1-forms dq_i should be though of as a basis. We've used the Einstein's summation formula over repeated indices and we shall continue using it throughout this entire chapter. By the rules of tensor calculus, p_i will transform co-variantly whereas it's basis 1-forms dq_i, contra-variantly:

$$P_i = \frac{\partial q^j}{\partial Q^i} p_j \qquad dQ^i = \frac{\partial Q^i}{\partial q^j} dq^j \tag{A.5}$$

that in turn trivially implies existence of a natural object, independent of the manifold coordinates - the *tautological one-form θ*, known also as *sympletic potential* [118]:

$$\theta = p_i dq^i \overset{q^i \to Q^i}{\longrightarrow} P_i dQ^i = \frac{\partial q^j}{\partial Q^i} \frac{\partial Q^i}{\partial q^k} p_j dq^k = \delta^j_k p_j dq^k = p_k dq^k \tag{A.6}$$

and by taking an exterior derivative of θ, defined using the wedge product $\wedge$, we automatically obtain an *invariant 2-form $\omega = d\theta$* which is also closed since $dd\theta = 0$ [141]:

$$\omega := d\theta = dp_i \wedge dq^i = \frac{\partial p_i}{\partial q^j} dq^j \wedge dq^i \tag{A.7}$$

A.2. Canonical equations

Equipped with the knowledge of the sympletic geometry of the phase space $\mathcal{M} = \mathcal{T}^* \mathcal{Q}$, consider any function $H(\mathbf{q}, \mathbf{p})$ on $\mathcal{M}$. Given the duality of the differential forms and vector fields, in the sense that members of the former serve as linear transformation on elements of the latter and vice versa, we can implicitly define a *vector field $\vec{X}_H$* on $\mathcal{TM}$ corresponding to H

$$dH(\vec{v}) = \omega(\vec{X}_H, \vec{v}) \tag{A.8}$$

where dH is the differential of H and $\vec{v}$ is an arbitrary vector field, in essence, an argument of dH. Since ω is a 2-form it could act on two vector fields to produce a scalar field, but it can just as well act on a single vector field, to produce a 1-form - a mechanism similar to rising and lowering indices with a metric tensor [26].

The integral curves of such a vector field, referred to as *hamiltonian flow $\vec{X}_H$* cover the entire phase space without intersection and uniquely determine the evolution of a classical system. In local coordinates [1], the vector field can be represented as [118]:

$$\vec{X}_H = \frac{\partial H}{\partial p_i} \frac{\partial}{\partial q^i} - \frac{\partial H}{\partial q^i} \frac{\partial}{\partial p_i} \tag{A.9}$$

Since a vector field is an operator, we can compute the rate of change of any tensor field on $\mathcal{M}$ using the Lie derivative formalism [5], for example for a scalar function $f : \mathcal{M} \to \mathbb{R}$:

$$\frac{df}{dt} := \mathcal{L}_{\vec{X}_H}(f) := \vec{X}_H(f) = \frac{\partial H}{\partial p_i} \frac{\partial f}{\partial q^i} - \frac{\partial H}{\partial q^i} \frac{\partial f}{\partial p_i} \tag{A.10}$$

[1] Notice some ambiguity of indexing of the basis vectors $\frac{\partial}{\partial x^i}$, normally, local coordinates of a manifold would be denoted with an upper index but since we have to do with a cotangent bundle and $x^i := \{q^i, p_j\}$, half of the coordinates are also the covector components and shall therefore be indexed with a lower index

notice that $\vec{X}_H(f) \equiv df(\vec{X}_H) \equiv \omega(\vec{X}_f, \vec{X}_H) := \{f, H\}$ defines the *Poisson bracket* and if $\{f, H\} = 0$ the function f is said to be *invariant* under H.

When applied to the coordinate curves, the Hamiltonian vector field acts as an *infinitesimal generator* [3, 5] for the system's evolution in time and yields the much celebrated *canonical differential equations of Hamilton*:

$$\frac{dq^i}{dt} = \vec{X}_H(q^i) = \frac{\partial H}{\partial p_i} \qquad \frac{dp_i}{dt} = \vec{X}_H(p_i) = -\frac{\partial H}{\partial q_i} \tag{A.11}$$

A.3. Properties of Hamiltonian dynamics

Time reversibility: Since Hamilton's canonical equations of motion define really an infinitesimal, bijective change of coordinates in agreement with the time evolution of a time-homogenous system, nothing prevents us from application thereof in reverse direction.

Energy conservation: Let's investigate the time rate of change of a *time-homogenous* Hamiltonian function itself:

$$\frac{dH}{dt} = \frac{\partial H}{\partial q^i}\frac{dq^i}{dt} + \frac{\partial H}{\partial p^i}\frac{dp_i}{dt} = \frac{\partial H}{\partial q^i}\frac{\partial H}{\partial p_i} - \frac{\partial H}{\partial p_i}\frac{\partial H}{\partial q^i} = 0 \tag{A.12}$$

where the first equality is simply the total derivative formula and the latter follows from the substitution of eq. A.11 for $\frac{dq^i}{dt}$ and $\frac{dp_i}{dt}$.

Sympleticity: Let's consider two, $2n$ dimensional, arbitrary vector fields $\vec{v}, \vec{u}$ at the tangent bundle of the phase manifold $\mathcal{TM}$. Expressed in local coordinates we have:

$$\vec{v} = \nu_q^i \frac{\partial}{\partial q^i} + \nu_p^i \frac{\partial}{\partial p_i}, \qquad \vec{u} = \mu_q^i \frac{\partial}{\partial q^i} + \mu_p^i \frac{\partial}{\partial p_i} \tag{A.13}$$

The sympletic form ω defined in eq. A.7, as any 2-form is a bilinear map, it takes two vector fields and produces a scalar, here we obtain:

$$\begin{gathered}\omega(\vec{v}, \vec{u}) := \left(dp_i \wedge dq^i\right)(\vec{v}, \vec{u}) = \sum_i^n dp_i(\vec{v})dq^i(\vec{u}) - dq^i(\vec{v})dp_i(\vec{u}) = \sum_i^n \nu_p^i \mu_q^i - \nu_q^i \mu_p^i \\ \omega(\vec{v}, \vec{u}) = \sum_i^n \det\begin{pmatrix} \nu_p^i & \mu_q^i \\ \nu_q^i & \mu_p^i \end{pmatrix}\end{gathered} \tag{A.14}$$

from which, the matrix representation of a sympletic form ω, usually denoted J, follows as:

$$\omega(\vec{v}, \vec{u}) = \vec{v}^T J \vec{u}, \qquad J = \begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix} \tag{A.15}$$

where I_n is $n \times n$ identity matrix. Verbosely, the above results represent the sum of areas of the parallelograms defined by projections of vectors $\vec{v}|_{\{\mathbf{q},\mathbf{p}\}}, \vec{u}|_{\{\mathbf{q},\mathbf{p}\}}$ [2] onto the coordinate

[2] We follow the common differential geometric notation [5, 26] with a vertical bar $\vec{v}|_{\mathbf{x}}$ to denote evaluation of the vector field at $\mathbf{x}$ instead of $\vec{v}(\mathbf{x})$ common in vector calculus. Since tensor fields are operators we reserve the latter for the *action* of the vector field on any other tensor.

planes $\{q^i, p_i\}$, $i = 1, ..., n$. Sympletic form can be therefore regarded as a *measure of volume* and by definition, it's a globally invariant one on $\mathcal{M}$.

Now, consider a coordinate transformation on $\mathcal{M}$, that is, a diffeomorphism which maps the phase space back to itself $F(\mathbf{q}, \mathbf{p}) : \mathcal{M} \to \mathcal{M}$. A *pushforward* or *differential* dF it induces, at any given point $\{\mathbf{q}, \mathbf{p}\} \in \mathcal{M}$, is a mapping between the tangent spaces $\mathcal{TM}|_{\{\mathbf{q},\mathbf{p}\}} \to \mathcal{TM}|_{\{F(\mathbf{q},\mathbf{p})\}}$, and it simply follows as the Jacobian of the transformation evaluated at that point [26]:

$$dF(\vec{v}|_{\{\mathbf{q},\mathbf{p}\}}) := \mathcal{J}_{F(\mathbf{q},\mathbf{p})}\vec{v}|_{\{\mathbf{q},\mathbf{p}\}} \overset{abbr.}{\equiv} \mathcal{J}_F\vec{v} \tag{A.16}$$

The transformation T is said to be *sympletic* if it leaves the value of the sympletic form $\omega(\vec{v}, \vec{u})$ unchanged, notice, we do not mean the sympletic form $\omega(\cdot, \cdot)$ itself - it is guaranteed to be invariant by construction. Considering all of the above, the condition of symplecity reads:

$$(\mathcal{J}_F\vec{v})^T \begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix} (\mathcal{J}_F\vec{u}) = \vec{v}^T \begin{bmatrix} 0 & -I_n \\ I_n & 0 \end{bmatrix} \vec{u} \qquad \forall \vec{v}, \vec{u} \tag{A.17}$$

It remains to be shown that the infinitesimal coordinate transformation corresponding to the time evolution, as defined by the canonical equations of motion eq. A.11, is sympletic. We might actually get away without computing the Jacobian if we consider an equivalent condition that the Lie derivative of $\omega(\vec{v}, \vec{u})$, $\forall \vec{v}, \vec{u}$ under the Hamiltonian vector field $\vec{X}_H$ should vanish:

$$\mathcal{L}_{\vec{X}_H}(\omega) = 0 \tag{A.18}$$

To show that, we first make use of the Cartan formula [26]:

$$\mathcal{L}_{\vec{X}_H}(\omega) = d(i_{\vec{X}_H}(\omega)) + i_{\vec{X}_H}(d\omega) \tag{A.19}$$

By construction $d\omega = 0$, thus what remains to be shown is that the exterior derivative d of the interior product $i_{\vec{X}_H}(\omega)$ equals zero or conversly, that it is *locally exact*, i.e. it can be written as the exterior derivative of some 0-form. That actually holds trivially by the definition of the Hamiltonian vector field in eq. A.8 and the definition of interior product of a 2-form with a vector field [26]:

$$\left(i_{\vec{X}_H}(\omega)\right)(\vec{v}) := \omega(\vec{X}_H, \vec{v}) := dH(\vec{v}) \qquad \forall \vec{v} \tag{A.20}$$

which concludes the proof.

Volume preservation: Although symplecity is more general, the volume preservation it implies can be also shown by a more straightforward computation, namely, that the Hamiltonian vector field $\vec{X}_H$ is *divergence free*:

$$\nabla \cdot \vec{X}_H = \frac{\partial^2 H}{\partial p_i \partial q^i} - \frac{\partial^2 H}{\partial q^i \partial p_i} = 0 \tag{A.21}$$

In statistical mechanics, this result is also known as *Liouville's theorem* [111]

A.4. Sympletic integrators

Hamiltonian vector field $\vec{X}_H$ defines an infinitesimal coordinate transformation on the sympletic manifold which corresponds to the evolution of the system in time. These dynamics, as was shown in the previous section, have the property of leaving the sympletic form $\omega(\vec{v}, \vec{u})$ invariant and as a result, conserve the phase space volume. For the purpose of numerical solution of the canonical equations of motions we require time discretization, thus it becomes important to determine what other transformations $\mathbf{q}_n, \mathbf{p}_n \rightarrow \mathbf{q}_{n+1}, \mathbf{p}_{n+1}$ - which are by necessity *not* infinitesimal - also maintain the desirable geometric properties of Hamilton's equations.

In particular, we will require that the numerical integrators for the solution of Hamiltonian equations of motion meet the condition A.17 and if that's the case we, shall call them *sympletic integrators* [142]. Nonetheless, even if the sympletic and time reversibility conditions can be met, the energy conservation property can be guaranteed only to an extend governed by the order of the method. This will lead to a slight, although permanent, energy drift and for obvious reasons will play a particularly important role in simulations over long time intervals.

First order integrators: Implementing naively the explicit Euler scheme for Hamiltonian dynamics would not lead to desired results as it's known to exhibit only conditional stability. Better results can be obtained by utilizing an implicit scheme which defines the *sympletic Euler* method [142]:

$$\begin{aligned} \mathbf{q}_{n+1} &= \mathbf{q}_n + \varepsilon \frac{\partial H}{\partial \mathbf{p}}(\mathbf{q}_n, \mathbf{p}_{n+1}) \overset{abbr.}{\equiv} \mathbf{p}_n + \varepsilon H_p \\ \mathbf{p}_{n+1} &= \mathbf{p}_n - \varepsilon \frac{\partial H}{\partial \mathbf{q}}(\mathbf{q}_n, \mathbf{p}_{n+1}) \overset{abbr.}{\equiv} \mathbf{p}_n - \varepsilon H_q \end{aligned} \tag{A.22}$$

with ε being the step size. The sympletic Euler method is really nothing more than the Taylor expansion of the Hamiltonian equations up to first order (just evaluated at particular points), it's therefore obvious it's a first order method i.e. we incur error of the order $\mathcal{O}(\varepsilon^2)$. The Jacobian of this transformation $\mathcal{J}_{SE} = \frac{\partial(\mathbf{q}_{n+1}, \mathbf{p}_{n+1})}{\partial(\mathbf{q}_n, \mathbf{p}_n)}$ follows as [142]:

$$\begin{bmatrix} I & -\varepsilon H_{pp} \\ 0 & I + \varepsilon H_{qp} \end{bmatrix} \frac{\partial(\mathbf{q}_{n+1}, \mathbf{p}_{n+1})}{\partial(\mathbf{q}_n, \mathbf{p}_n)} = \begin{bmatrix} I + \varepsilon H_{qp} & 0 \\ -\varepsilon H_{qq} & I \end{bmatrix} \tag{A.23}$$

from which the Jacobian can be evaluated and the sympletic condition A.17 verified to hold, refer to [142] for more details.

Second order integrators: Composition of two sympletic Euler methods, by their symmetry leads to cancellation of certain terms and defines a *time-reversible*, second order method, so called *Störmer-Verlet method* [142]:

$$\begin{aligned}
\mathbf{p}_{n+\frac{1}{2}} &= \mathbf{p}_n - \frac{\varepsilon}{2}\frac{\partial H}{\partial \mathbf{q}}(\mathbf{q}_n, \mathbf{p}_{n+\frac{1}{2}}) \\
\mathbf{q}_{n+1} &= \mathbf{q}_n + \frac{\varepsilon}{2}\left(\frac{\partial H}{\partial \mathbf{p}}(\mathbf{q}_n, \mathbf{p}_{n+\frac{1}{2}}) + \frac{\partial H}{\partial \mathbf{p}}(\mathbf{q}_{n+1}, \mathbf{p}_{n+\frac{1}{2}})\right) \\
\mathbf{p}_{n+1} &= \mathbf{p}_{n+\frac{1}{2}} - \frac{\varepsilon}{2}\frac{\partial H}{\partial \mathbf{q}}(\mathbf{q}_{n+1}, \mathbf{p}_{n+\frac{1}{2}})
\end{aligned} \tag{A.24}$$

Depending on the field, the method goes under the name of *Störmer* in astronomy, *Verlet* in molecular dynamics or *leapfrog* in the context of partial differential equations, also notice that for separable Hamiltonians i.e. $H(\mathbf{q}, \mathbf{p}) = U(\mathbf{q}) + T(\mathbf{q})$ the method becomes explicit:

$$\begin{aligned}
\mathbf{p}_{n+\frac{1}{2}} &= \mathbf{p}_n - \frac{\varepsilon}{2}\frac{\partial U}{\partial \mathbf{q}}(\mathbf{q}_n) \\
\mathbf{q}_{n+1} &= \mathbf{q}_n + \varepsilon\frac{\partial T}{\partial \mathbf{p}}(\mathbf{p}_{n+\frac{1}{2}}) \\
\mathbf{p}_{n+1} &= \mathbf{p}_{n+\frac{1}{2}} - \frac{\varepsilon}{2}\frac{\partial H}{\partial \mathbf{q}}(\mathbf{q}_{n+1})
\end{aligned} \tag{A.25}$$

Moreover, in actual implementation, if one is not interested in momentum at time n, the first and last equations in the above definition can be combined after an initial half-step is taken and from then on, its updates will follow for the half-integer time only $\mathbf{p}_{n+\frac{1}{2}}$, making the updates in $\mathbf{q}$ and $\mathbf{p}$ "leaping" over each other.

Higher order schemes can be obtained within the Runge-Kuta family of methods by imposing certain conditions on the slope coefficients, these conditions read [142]:

$$b_i a_{ij} + b_j a_{ji} = b_i b_j \qquad \forall\ i, j = 1, ..., s \tag{A.26}$$

The principles guiding the design of higher order methods and from which the above condition ultimately stems rely fundamentally on direct numerical treatment of the principles of analytical mechanics, in particular the *generating functions* and *variational principles*. The former, are solutions to the Hamilton-Jacobi partial differential equation and high order sympletic integration schemes follow from their solution using particular ansatz of the numerical flow. In the case of variational principles the method requires numerical approximation of the action integral (eq. A.1) from which discrete Euler-Lagrange conditions follow. Depending on the quadrature scheme used we then end up with various sympletic integrators, in particular, the trapezoidal rule leads to the already covered Störmer-Verlet method [142].

List of Figures

List of Tables

Bibliography

[1] S. H. Simon. *The Oxford Solid State Basics*. 1st ed. Oxford: Oxford University Press, 2013. ISBN: 978-0-19-968077-1 978-0-19-968076-4.

[2] T. F. Jordan. *Quantum Mechanics in Simple Matrix Form*. Mineola, N.Y: Dover Publications, 2006. ISBN: 978-0-486-44530-4.

[3] L. Cornellius. *Variational Principles of Mechanics*. First.

[4] D. Bohm. *Quantum Theory*. New York: Dover Publications, 1989. ISBN: 978-0-486-65969-5.

[5] P. J. Olver. *Applications of Lie Groups to Differential Equations*. Vol. 107. Graduate Texts in Mathematics. New York, NY: Springer New York, 1986. ISBN: 978-1-4684-0276-6 978-1-4684-0274-2. DOI: `10.1007/978-1-4684-0274-2`. (Visited on 10/07/2022).

[6] M. Kumar. *Quantum: Einstein, Bohr and the Great Debate about the Nature of Reality*. New York, 2011.

[7] D. Bohm, B. J. Hiley, B. J. Hiley, and D. Bohm. *The Undivided Universe: An Ontological Interpretation of Quantum Theory.* 2006. ISBN: 978-0-203-98038-5. (Visited on 06/09/2022).

[8] J. Von Neumann and J. Von Neumann. *Mathematical Foundations of Quantum Mechanics*. Princeton Landmarks in Mathematics and Physics. Princeton Chichester: Princeton University Press, 1996. ISBN: 978-0-691-02893-4 978-0-691-08003-1.

[9] A. Einstein, B. Podolsky, and N. Rosen. "Can Quantum-Mechanical Description of Physical Reality Be Considered Complete?" In: *Physical Review* 47.10 (May 1935), pp. 777–780. ISSN: 0031-899X. DOI: `10.1103/PhysRev.47.777`. (Visited on 09/19/2021).

[10] P. A. M. Dirac. *The Principles of Quantum Mechanics*. New York: Snowball Publishing, 2013. ISBN: 978-1-60796-560-2.

[11] J. Schwichtenberg. *No-Nonsense Quantum Field Theory*. First printing. Karlsruhe: No-Nonsense Books, 2020. ISBN: 978-3-948763-01-5.

[12] F. Giustino. *Materials Modelling Using Density Functional Theory: Properties and Predictions*. 1st ed. Oxford: Oxford University Press, 2014. ISBN: 978-0-19-966243-2 978-0-19-966244-9.

[13] W. Kutzelnigg and P. von Herigonte. "Electron Correlation at the Dawn of the 21st Century". In: *Advances in Quantum Chemistry*. Vol. 36. Elsevier, 2000, pp. 185–229. ISBN: 978-0-12-034836-7. DOI: `10.1016/S0065-3276(08)60484-0`. (Visited on 02/20/2023).

[14] D. P. Tew, W. Klopper, and T. Helgaker. "Electron Correlation: The Many-Body Problem at the Heart of Chemistry: Electron Correlation: A Many-Body Problem". In: *Journal of Computational Chemistry* 28.8 (June 2007), pp. 1307–1320. ISSN: 01928651. DOI: `10.1002/jcc.20581`. (Visited on 02/20/2023).

[15] D. Marx and J. Hutter. *Ab Initio Molecular Dynamics*.

[16] C. J. Cramer. *Essentials of Computational Chemistry: Theories and Models*. 2nd ed. Chichester, West Sussex, England ; Hoboken, NJ: Wiley, 2004. ISBN: 978-0-470-09182-1 978-0-470-09181-4.

[17] S. Sorella and L. Capriotti. "Green Function Monte Carlo with Stochastic Reconfiguration: An Effective Remedy for the Sign Problem". In: *Physical Review B* 61.4 (Jan. 2000), pp. 2599–2612. ISSN: 0163-1829, 1095-3795. DOI: `10.1103/PhysRevB.61.2599`. (Visited on 01/26/2023).

[18] P. W. Anderson. "More Is Different: Broken Symmetry and the Nature of the Hierarchical Structure of Science." In: *Science* 177.4047 (Aug. 1972), pp. 393–396. ISSN: 0036-8075, 1095-9203. DOI: `10.1126/science.177.4047.393`. (Visited on 02/19/2023).

[19] A. Szabo and N. S. Ostlund. *Modern Quantum Chemistry: Introduction to Advanced Electronic Structure Theory*. Mineola, N.Y: Dover Publications, 1996. ISBN: 978-0-486-69186-2.

[20] K. E. Drexler. *Engines of Creation: The Coming Era of Nanotechnology*. 6. [print.] Anchor Books. New York: Doubleday, 1990. ISBN: 978-0-385-19973-5.

[21] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal. "Quantum Monte Carlo Simulations of Solids". In: *Reviews of Modern Physics* 73.1 (Jan. 2001), pp. 33–83. ISSN: 0034-6861, 1539-0756. DOI: `10.1103/RevModPhys.73.33`. (Visited on 11/07/2022).

[22] N. Marzari, A. Ferretti, and C. Wolverton. "Electronic-Structure Methods for Materials Design". In: *Nature Materials* 20.6 (June 2021), pp. 736–749. ISSN: 1476-1122, 1476-4660. DOI: `10.1038/s41563-021-01013-3`. (Visited on 05/25/2022).

[23] J. Schmidt, M. R. G. Marques, S. Botti, and M. A. L. Marques. "Recent Advances and Applications of Machine Learning in Solid-State Materials Science". In: *npj Computational Materials* 5.1 (Dec. 2019), p. 83. ISSN: 2057-3960. DOI: `10.1038/s41524-019-0221-0`. (Visited on 03/07/2021).

[24] F. Becca and S. Sorella. *Quantum Monte Carlo Approaches for Correlated Systems*. First. Cambridge University Press, Nov. 2017. ISBN: 978-1-107-12993-1 978-1-316-41704-1. DOI: `10.1017/9781316417041`. (Visited on 09/16/2022).

[25] J. M. Zhang and N. J. Mauser. "Optimal Slater-determinant Approximation of Fermionic Wave Functions". In: *Physical Review A* 94.3 (Sept. 2016), p. 032513. ISSN: 2469-9926, 2469-9934. DOI: `10.1103/PhysRevA.94.032513`. arXiv: `1510.05634 [quant-ph]`. (Visited on 11/26/2022).

[26] L. W. Tu. *An Introduction to Manifolds*. Universitext. New York, NY: Springer New York, 2011. ISBN: 978-1-4419-7399-3 978-1-4419-7400-6. DOI: `10.1007/978-1-4419-7400-6`. (Visited on 08/06/2022).

[27] M. Motta, D. M. Ceperley, G. K.-L. Chan, J. A. Gomez, E. Gull, S. Guo, C. A. Jiménez-Hoyos, T. N. Lan, J. Li, F. Ma, A. J. Millis, N. V. Prokof'ev, U. Ray, G. E. Scuseria, S. Sorella, E. M. Stoudenmire, Q. Sun, I. S. Tupitsyn, S. R. White, D. Zgid, S. Zhang, and Simons Collaboration on the Many-Electron Problem. "Towards the Solution of the Many-Electron Problem in Real Materials: Equation of State of the Hydrogen Chain with State-of-the-Art Many-Body Methods". In: *Physical Review X* 7.3 (Sept. 2017), p. 031059. ISSN: 2160-3308. DOI: `10.1103/PhysRevX.7.031059`. (Visited on 11/12/2022).

[28] R. J. Bartlett and M. Musiał. "Coupled-Cluster Theory in Quantum Chemistry". In: *Reviews of Modern Physics* 79.1 (Feb. 2007), pp. 291–352. ISSN: 0034-6861, 1539-0756. DOI: `10.1103/RevModPhys.79.291`. (Visited on 10/17/2022).

[29] X. Zhang, J. Zhu, Z. Wen, and A. Zhou. *Gradient Type Optimization Methods for Electronic Structure Calculations*. Aug. 2013. arXiv: `arXiv:1308.2864`. (Visited on 11/07/2022).

[30] N. T. Hung and A. R. T. Nugraha. "Quantum Espresso Hands-on Tutorial". In: (), p. 76.

[31] J. Toulouse. *Review of Approximations for the Exchange-Correlation Energy in Density-Functional Theory*. Sept. 2022. arXiv: `arXiv:2103.02645`. (Visited on 02/09/2023).

[32] M. F. Kasim and S. M. Vinko. "Learning the Exchange-Correlation Functional from Nature with Fully Differentiable Density Functional Theory". In: *Physical Review Letters* 127.12 (Sept. 2021), p. 126403. ISSN: 0031-9007, 1079-7114. DOI: `10.1103/PhysRevLett.127.126403`. arXiv: `2102.04229 [physics]`. (Visited on 02/10/2023).

[33] R. Jastrow. "Many-Body Problem with Strong Forces". In: *Physical Review* 98.5 (June 1955), pp. 1479–1484. ISSN: 0031-899X. DOI: `10.1103/PhysRev.98.1479`. (Visited on 02/03/2023).

[34] R. P. Feynman and M. Cohen. "Energy Spectrum of the Excitations in Liquid Helium". In: *Physical Review* 102.5 (June 1956), pp. 1189–1204. ISSN: 0031-899X. DOI: `10.1103/PhysRev.102.1189`. (Visited on 02/23/2023).

[35] L. F. Tocchio, F. Becca, A. Parola, and S. Sorella. "Role of Backflow Correlations for the Non-Magnetic Phase of the t-t' Hubbard Model". In: *Physical Review B* 78.4 (July 2008), p. 041101. ISSN: 1098-0121, 1550-235X. DOI: `10.1103/PhysRevB.78.041101`. arXiv: `0805.1476 [cond-mat]`. (Visited on 02/20/2023).

[36] M. Casula and S. Sorella. "Geminal Wave Functions with Jastrow Correlation: A First Application to Atoms". In: *The Journal of Chemical Physics* 119.13 (Oct. 2003), pp. 6500–6511. ISSN: 0021-9606, 1089-7690. DOI: `10.1063/1.1604379`. (Visited on 02/19/2023).

[37] C. J. Geyer. "Practical Markov Chain Monte Carlo". In: *Statistical Science* 7.4 (Nov. 1992). ISSN: 0883-4237. DOI: `10.1214/ss/1177011137`. (Visited on 03/12/2023).

[38] C. M. Bishop. *Pattern Recognition and Machine Learning*. Information Science and Statistics. New York: Springer, 2006. ISBN: 978-0-387-31073-2.

[39] K. Choo, A. Mezzacapo, and G. Carleo. "Fermionic Neural-Network States for Ab-Initio Electronic Structure". In: *Nature Communications* 11.1 (May 2020), pp. 1–7. ISSN: 2041-1723. DOI: `10.1038/s41467-020-15724-9`. (Visited on 09/12/2022).

[40] H. Robbins and S. Monro. "A Stochastic Approximation Method". In: *The Annals of Mathematical Statistics* 22.3 (Sept. 1951), pp. 400–407. ISSN: 0003-4851, 2168-8990. DOI: `10.1214/aoms/1177729586`. (Visited on 03/14/2023).

[41] S. Sorella. "Generalized Lanczos Algorithm for Variational Quantum Monte Carlo". In: *Physical Review B* 64.2 (June 2001), p. 024512. ISSN: 0163-1829, 1095-3795. DOI: `10.1103/PhysRevB.64.024512`. arXiv: `cond-mat/0009149`. (Visited on 10/20/2022).

[42] J. Martens and R. Grosse. "Optimizing Neural Networks with Kronecker-factored Approximate Curvature". In: (), p. 10.

[43] S.-i. Amari. "Natural Gradient Works Efficiently in Learning". In: (), p. 26.

[44] S. Amari, H. Nagaoka, S. Amari, and S. Amari. *Methods of Information Geometry*. Trans. by D. Harada. Nachdruck. Translations of Mathematical Monographs 191. Providence, Rhode Island: American Mathematical Society, 2007. ISBN: 978-0-8218-4302-4 978-0-8218-0531-2.

[45] M. M. Bronstein, J. Bruna, T. Cohen, and P. Veličković. "Geometric Deep Learning: Grids, Groups, Graphs, Geodesics, and Gauges". In: *arXiv:2104.13478 [cs, stat]* (May 2021). arXiv: `2104.13478 [cs, stat]`. (Visited on 12/11/2021).

[46] J. Klicpera, J. Groß, and S. Günnemann. "Directional Message Passing for Molecular Graphs". In: *arXiv:2003.03123 [physics, stat]* (Mar. 2020). arXiv: `2003.03123 [physics, stat]`. (Visited on 11/09/2021).

[47] J. Klicpera, F. Becker, and S. Günnemann. "GemNet: Universal Directional Graph Neural Networks for Molecules". In: *arXiv:2106.08903 [physics, stat]* (Oct. 2021). arXiv: `2106.08903 [physics, stat]`. (Visited on 11/09/2021).

[48] N. Thomas, T. Smidt, S. Kearnes, L. Yang, L. Li, K. Kohlhoff, and P. Riley. "Tensor Field Networks: Rotation- and Translation-Equivariant Neural Networks for 3D Point Clouds". In: *arXiv:1802.08219 [cs]* (May 2018). arXiv: `1802.08219 [cs]`. (Visited on 11/07/2021).

[49] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller. "SchNet: A Continuous-Filter Convolutional Neural Network for Modeling Quantum Interactions". In: *arXiv:1706.08566 [physics, stat]* (Dec. 2017). arXiv: `1706.08566 [physics, stat]`. (Visited on 01/15/2022).

[50] Ł. Maziarka, A. Pocha, J. Kaczmarczyk, K. Rataj, T. Danel, and M. Warchoł. "Mol-CycleGAN: A Generative Model for Molecular Optimization". In: *Journal of Cheminformatics* 12.1 (Dec. 2020), p. 2. ISSN: 1758-2946. DOI: `10.1186/s13321-019-0404-1`. (Visited on 02/06/2022).

[51] "Graph Convolutional Policy Network for Goal-Directed Molecular Graph Generation". In: (). (Visited on 12/18/2021).

[52] C.-T. Chen and G. X. Gu. "Generative Deep Neural Networks for Inverse Materials Design Using Backpropagation and Active Learning". In: *Advanced Science* 7.5 (Mar. 2020), p. 1902607. ISSN: 2198-3844, 2198-3844. DOI: `10.1002/advs.201902607`. (Visited on 05/04/2022).

[53] B. Sanchez-Lengeling and A. Aspuru-Guzik. "Inverse Molecular Design Using Machine Learning: Generative Models for Matter Engineering". In: *Science* 361.6400 (July 2018), pp. 360–365. ISSN: 0036-8075, 1095-9203. DOI: `10.1126/science.aat2663`. (Visited on 01/17/2022).

[54] G. Carleo and M. Troyer. "Solving the Quantum Many-Body Problem with Artificial Neural Networks". In: *Science* 355.6325 (Feb. 2017), pp. 602–606. DOI: `10.1126/science.aag2302`. (Visited on 09/15/2022).

[55] J. Hermann, Z. Schätzle, and F. Noé. "Deep-Neural-Network Solution of the Electronic Schrödinger Equation". In: *Nature Chemistry* 12.10 (Oct. 2020), pp. 891–897. ISSN: 1755-4349. DOI: `10.1038/s41557-020-0544-y`. (Visited on 09/12/2022).

[56] D. Pfau, J. S. Spencer, A. G. D. G. Matthews, and W. M. C. Foulkes. "*Ab Initio* Solution of the Many-Electron Schrödinger Equation with Deep Neural Networks". In: *Physical Review Research* 2.3 (Sept. 2020), p. 033429. ISSN: 2643-1564. DOI: `10.1103/PhysRevResearch.2.033429`. (Visited on 09/10/2022).

[57] Z.-A. Jia, B. Yi, R. Zhai, Y.-C. Wu, G.-C. Guo, and G.-P. Guo. "Quantum Neural Network States: A Brief Review of Methods and Applications". In: *Advanced Quantum Technologies* 2.7-8 (Aug. 2019), p. 1800077. ISSN: 2511-9044, 2511-9044. DOI: `10.1002/qute.201800077`. arXiv: `1808.10601 [physics, physics:quant-ph]`. (Visited on 01/29/2023).

[58] S. Battaglia. *Machine Learning Wavefunction*. Feb. 2022. arXiv: `arXiv:2202.13916`. (Visited on 09/05/2022).

[59] Y. LeCun, S. Chopra, R. Hadsell, M. Ranzato, and F. J. Huang. "A Tutorial on Energy-Based Learning". In: (), p. 59.

[60] N. Le Roux and Y. Bengio. "Representational Power of Restricted Boltzmann Machines and Deep Belief Networks". In: *Neural Computation* 20.6 (June 2008), pp. 1631–1649. ISSN: 0899-7667, 1530-888X. DOI: `10.1162/neco.2008.04-07-510`. (Visited on 02/21/2023).

[61] Y. Nomura, A. S. Darmawan, Y. Yamaji, and M. Imada. "Restricted Boltzmann Machine Learning for Solving Strongly Correlated Quantum Systems". In: *Physical Review B* 96.20 (Nov. 2017), p. 205152. ISSN: 2469-9950, 2469-9969. DOI: `10.1103/PhysRevB.96.205152`. (Visited on 10/17/2022).

[62] A. Valenti, E. Greplova, N. H. Lindner, and S. D. Huber. "Correlation-Enhanced Neural Networks as Interpretable Variational Quantum States". In: *Physical Review Research* 4.1 (Jan. 2022), p. L012010. ISSN: 2643-1564. DOI: `10.1103/PhysRevResearch.4.L012010`. (Visited on 01/29/2023).

[63] N. Yoshioka, W. Mizukami, and F. Nori. “Solving Quasiparticle Band Spectra of Real Solids Using Neural-Network Quantum States”. In: *Communications Physics* 4.1 (May 2021), p. 106. ISSN: 2399-3650. DOI: `10.1038/s42005-021-00609-0`. (Visited on 02/20/2023).

[64] K. Choo, G. Carleo, N. Regnault, and T. Neupert. “Symmetries and Many-Body Excited States with Neural-Network Quantum States”. In: *Physical Review Letters* 121.16 (Oct. 2018), p. 167204. ISSN: 0031-9007, 1079-7114. DOI: `10.1103/PhysRevLett.121.167204`. arXiv: `1807.03325 [cond-mat, physics:quant-ph]`. (Visited on 02/21/2023).

[65] D.-L. Deng, X. Li, and S. D. Sarma. “Quantum Entanglement in Neural Network States”. In: *Physical Review X* 7.2 (May 2017), p. 021021. ISSN: 2160-3308. DOI: `10.1103/PhysRevX.7.021021`. arXiv: `1701.04844 [cond-mat, physics:quant-ph]`. (Visited on 02/01/2023).

[66] L. Amico, R. Fazio, A. Osterloh, and V. Vedral. “Entanglement in Many-Body Systems”. In: *Reviews of Modern Physics* 80.2 (May 2008), pp. 517–576. ISSN: 0034-6861, 1539-0756. DOI: `10.1103/RevModPhys.80.517`. (Visited on 02/19/2023).

[67] C.-Y. Park and M. J. Kastoryano. “Geometry of Learning Neural Quantum States”. In: *Physical Review Research* 2.2 (May 2020), p. 023232. ISSN: 2643-1564. DOI: `10.1103/PhysRevResearch.2.023232`. arXiv: `1910.11163 [cond-mat, physics:quant-ph, stat]`. (Visited on 10/18/2022).

[68] X. Gao and L.-M. Duan. “Efficient Representation of Quantum Many-Body States with Deep Neural Networks”. In: *Nature Communications* 8.1 (Sept. 2017), p. 662. ISSN: 2041-1723. DOI: `10.1038/s41467-017-00705-2`. (Visited on 02/21/2023).

[69] D. Kochkov, T. Pfaff, A. Sanchez-Gonzalez, P. Battaglia, and B. K. Clark. *Learning Ground States of Quantum Hamiltonians with Graph Networks*. Oct. 2021. arXiv: `arXiv:2110.06390`. (Visited on 08/15/2022).

[70] O. Sharir, Y. Levine, N. Wies, G. Carleo, and A. Shashua. “Deep Autoregressive Models for the Efficient Variational Simulation of Many-Body Quantum Systems”. In: *Physical Review Letters* 124.2 (Jan. 2020), p. 020503. ISSN: 0031-9007, 1079-7114. DOI: `10.1103/PhysRevLett.124.020503`. arXiv: `1902.04057 [cond-mat]`. (Visited on 11/17/2022).

[71] M. Hutter. *On Representing (Anti)Symmetric Functions*. July 2020. arXiv: `arXiv:2007.15298`. (Visited on 02/21/2023).

[72] H. Huang, J. M. Landsberg, and J. Lu. *Geometry of Backflow Transformation Ansatz for Quantum Many-Body Fermionic Wavefunctions*. Nov. 2021. arXiv: `arXiv:2111.10314`. (Visited on 02/20/2023).

[73] K. Hornik, M. Stinchcombe, and H. White. “Multilayer Feedforward Networks Are Universal Approximators”. In: *Neural Networks* 2.5 (Jan. 1989), pp. 359–366. ISSN: 08936080. DOI: `10.1016/0893-6080(89)90020-8`. (Visited on 10/04/2022).

[74] J. S. Spencer, D. Pfau, A. Botev, and W. M. C. Foulkes. *Better, Faster Fermionic Neural Networks*. Nov. 2020. arXiv: `arXiv:2011.07125`. (Visited on 11/11/2022).

[75] K. He, X. Zhang, S. Ren, and J. Sun. "Deep Residual Learning for Image Recognition". In: *arXiv:1512.03385 [cs]* (Dec. 2015). arXiv: `1512.03385 [cs]`. (Visited on 05/15/2020).

[76] S. S. Schoenholz and E. D. Cubuk. *JAX, M.D.: A Framework for Differentiable Physics*. Dec. 2020. arXiv: `1912.04232 [cond-mat, physics:physics, stat]`. (Visited on 05/20/2022).

[77] W. Ren, W. Fu, and J. Chen. *Towards the Ground State of Molecules via Diffusion Monte Carlo on Neural Networks*. Apr. 2022. arXiv: `arXiv:2204.13903`. (Visited on 02/27/2023).

[78] D. Luo and B. K. Clark. "Backflow Transformations via Neural Networks for Quantum Many-Body Wave Functions". In: *Physical Review Letters* 122.22 (June 2019), p. 226401. ISSN: 0031-9007, 1079-7114. DOI: `10.1103/PhysRevLett.122.226401`. (Visited on 10/17/2022).

[79] K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer. "Unifying Machine Learning and Quantum Chemistry with a Deep Neural Network for Molecular Wavefunctions". In: *Nature Communications* 10.1 (Dec. 2019), p. 5024. ISSN: 2041-1723. DOI: `10.1038/s41467-019-12875-2`. (Visited on 04/13/2021).

[80] Y. LeCun. "Object Recognition with Gradient-Based Learning". In: (). (Visited on 02/23/2023).

[81] L. Gerard, M. Scherbela, P. Marquetand, and P. Grohs. *Gold-Standard Solutions to the Schr\"odinger Equation Using Deep Learning: How Much Physics Do We Need?* Oct. 2022. arXiv: `arXiv:2205.09438`. (Visited on 03/01/2023).

[82] Y. Gal. "Uncertainty in Deep Learning". In: (), p. 174.

[83] D. Pflüger. *Spatially Adaptive Sparse Grids for High-Dimensional Problems*. 1. Aufl. München: Verl. Dr. Hut, 2010. ISBN: 978-3-86853-555-6.

[84] H.-J. Bungartz and M. Griebel. "Sparse Grids". In: *Acta Numerica* 13 (May 2004), pp. 147–269. ISSN: 0962-4929, 1474-0508. DOI: `10.1017/S0962492904000182`. (Visited on 09/10/2020).

[85] I. Daubechies. *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics 61. Philadelphia, Pa: Society for Industrial and Applied Mathematics, 1992. ISBN: 978-0-89871-274-2.

[86] Ian Goodfellow and Y. Bengio. *Deep Learning*. (Visited on 11/04/2020).

[87] M. Raissi, Paris Perdikaris, and George Em Karniadakis. "Physics Informed Deep Learning Data-driven Solutions and Discovery of Nonlinear Partial Differential Equations". In: *Physics Informed Deep Learning* (). (Visited on 11/22/2020).

[88] E. Noether and M. A. Tavel. "Invariant Variation Problems". In: *Transport Theory and Statistical Physics* 1.3 (Jan. 1971), pp. 186–207. ISSN: 0041-1450, 1532-2424. DOI: `10.1080/00411457108231446`. arXiv: `physics/0503066`. (Visited on 12/11/2021).

[89] R. C. Staudemeyer and E. R. Morris. "Understanding LSTM – a Tutorial into Long Short-Term Memory Recurrent Neural Networks". In: *arXiv:1909.09586 [cs]* (Sept. 2019). arXiv: `1909.09586 [cs]`. (Visited on 06/20/2021).

[90] J. Bruna, W. Zaremba, A. Szlam, and Y. LeCun. *Spectral Networks and Locally Connected Networks on Graphs*. May 2014. arXiv: `arXiv:1312.6203`. (Visited on 03/01/2023).

[91] M. Defferrard, X. Bresson, and P. Vandergheynst. "Convolutional Neural Networks on Graphs with Fast Localized Spectral Filtering". In: *arXiv:1606.09375 [cs, stat]* (Feb. 2017). arXiv: `1606.09375 [cs, stat]`. (Visited on 12/12/2021).

[92] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. "Attention Is All You Need". In: *arXiv:1706.03762 [cs]* (Dec. 2017). arXiv: `1706.03762 [cs]`. (Visited on 01/15/2021).

[93] T. S. Cohen and M. Welling. "Group Equivariant Convolutional Networks". In: *arXiv:1602.07576 [cs, stat]* (June 2016). arXiv: `1602.07576 [cs, stat]`. (Visited on 11/15/2021).

[94] T. S. Cohen and M. Welling. "Steerable CNNs". In: *arXiv:1612.08498 [cs, stat]* (Dec. 2016). arXiv: `1612.08498 [cs, stat]`. (Visited on 11/15/2021).

[95] D. E. Worrall, S. J. Garbin, D. Turmukhambetov, and G. J. Brostow. "Harmonic Networks: Deep Translation and Rotation Equivariance". In: *arXiv:1612.04642 [cs, stat]* (Apr. 2017). arXiv: `1612.04642 [cs, stat]`. (Visited on 12/11/2021).

[96] W. T. F. Edward H. Adelson. "The Design and Use of Steerable Filters". In: (), p. 43.

[97] M. Weiler, M. Geiger, M. Welling, W. Boomsma, and T. Cohen. "3D Steerable CNNs: Learning Rotationally Equivariant Features in Volumetric Data". In: *arXiv:1807.02547 [cs, stat]* (Oct. 2018). arXiv: `1807.02547 [cs, stat]`. (Visited on 12/11/2021).

[98] J. Gasteiger, S. Giri, J. T. Margraf, and S. Günnemann. *Fast and Uncertainty-Aware Directional Message Passing for Non-Equilibrium Molecules*. Apr. 2022. arXiv: `arXiv:2011.14115`. (Visited on 03/03/2023).

[99] S. Batzner, A. Musaelian, L. Sun, M. Geiger, J. P. Mailoa, M. Kornbluth, N. Molinari, T. E. Smidt, and B. Kozinsky. "SE(3)-Equivariant Graph Neural Networks for Data-Efficient and Accurate Interatomic Potentials". In: *arXiv:2101.03164 [cond-mat, physics:physics]* (July 2021). arXiv: `2101.03164 [cond-mat, physics:physics]`. (Visited on 11/15/2021).

[100] Z. Li, K. Meidani, P. Yadav, and A. Barati Farimani. "Graph Neural Networks Accelerated Molecular Dynamics". In: *The Journal of Chemical Physics* 156.14 (Apr. 2022), p. 144103. ISSN: 0021-9606, 1089-7690. DOI: `10.1063/5.0083060`. (Visited on 05/21/2022).

[101] H. Li, Z. Wang, N. Zou, M. Ye, R. Xu, X. Gong, W. Duan, and Y. Xu. *Deep-Learning Density Functional Theory Hamiltonian for Efficient Ab Initio Electronic-Structure Calculation*. May 2022. arXiv: `2104.03786 [cond-mat, physics:physics, physics:quant-ph]`. (Visited on 06/01/2022).

[102] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. Second edition. Adaptive Computation and Machine Learning Series. Cambridge, Massachusetts: The MIT Press, 2018. ISBN: 978-0-262-03924-6.

[103] R. M. Neal. *MCMC Using Hamiltonian Dynamics*. May 2011. DOI: 10.1201/b10905. arXiv: 1206.1901 [physics, stat]. (Visited on 08/06/2022).

[104] Z. Liu and Z. Zhang. *Quantum-Inspired Hamiltonian Monte Carlo for Bayesian Sampling*. Aug. 2020. arXiv: arXiv:1912.01937. (Visited on 03/01/2023).

[105] M. Girolami, B. Calderhead, and S. A. Chin. "Riemann Manifold Langevin and Hamiltonian Monte Carlo". In: (), p. 38.

[106] S. D. Team. *Stan Modeling Language Users Guide and Reference Manual*. 2023. URL: https://mc-stan.org/users/documentation/ (visited on 03/01/2023).

[107] M. Heideman, D. Johnson, and C. Burrus. "Gauss and the History of the Fast Fourier Transform". In: *IEEE ASSP Magazine* 1.4 (Oct. 1984), pp. 14–21. ISSN: 1558-1284. DOI: 10.1109/MASSP.1984.1162257.

[108] J. W. Cooley and J. W. Tukey. "An Algorithm for the Machine Calculation of Complex Fourier Series". In: ().

[109] A. Scemama, T. Lelièvre, G. Stoltz, E. Cancès, and M. Caffarel. "An Efficient Sampling Algorithm for Variational Monte Carlo". In: *The Journal of Chemical Physics* 125.11 (Sept. 2006), p. 114105. ISSN: 0021-9606, 1089-7690. DOI: 10.1063/1.2354490. (Visited on 02/23/2023).

[110] S. Duane, A. Kennedy, B. J. Pendleton, and D. Roweth. "Hybrid Monte Carlo". In: *Physics Letters B* 195.2 (Sept. 1987), pp. 216–222. ISSN: 03702693. DOI: 10.1016/0370-2693(87)91197-X. (Visited on 09/04/2022).

[111] M. E. Tuckerman. *Statistical Mechanics: Theory and Molecular Simulation*. Oxford ; New York: Oxford University Press, 2010. ISBN: 978-0-19-852526-4.

[112] M. D. Hoffman and A. Gelman. *The No-U-Turn Sampler: Adaptively Setting Path Lengths in Hamiltonian Monte Carlo*. Nov. 2011. arXiv: arXiv:1111.4246. (Visited on 03/11/2023).

[113] F. J. Anscombe. "Graphs in Statistical Analysis". In: *The American Statistician* 27.1 (Feb. 1973), pp. 17–21. ISSN: 0003-1305, 1537-2731. DOI: 10.1080/00031305.1973.10478966. (Visited on 03/13/2023).

[114] C. Andrieu and J. Thoms. "A Tutorial on Adaptive MCMC". In: *Statistics and Computing* 18.4 (Dec. 2008), pp. 343–373. ISSN: 0960-3174, 1573-1375. DOI: 10.1007/s11222-008-9110-y. (Visited on 03/14/2023).

[115] M. D. Hoffman and P. Sountsov. "Tuning-Free Generalized Hamiltonian Monte Carlo". In: *Proceedings of The 25th International Conference on Artificial Intelligence and Statistics*. PMLR, May 2022, pp. 7799–7813. (Visited on 03/13/2023).

[116] R. M. Neal. "Slice Sampling". In: *The Annals of Statistics* 31.3 (June 2003). ISSN: 0090-5364. DOI: 10.1214/aos/1056562461. (Visited on 03/14/2023).

[117] C. H. Bennett. "Mass Tensor Molecular Dynamics". In: *Journal of Computational Physics* 19.3 (Nov. 1975), pp. 267–279. ISSN: 00219991. DOI: 10.1016/0021-9991(75)90077-7. (Visited on 03/18/2023).

[118] A. Carosso. *Geometric Quantization*. Jan. 2018. arXiv: `arXiv:1801.02307`. (Visited on 07/27/2022).

[119] N. M. J. Woodhouse. *Geometric Quantization*. 2nd ed. Oxford Mathematical Monographs. Oxford: Clarendon press, 1997. ISBN: 978-0-19-850270-8.

[120] J. Schwichtenberg. *Physics from Symmetry*. Undergraduate Lecture Notes in Physics. Cham: Springer International Publishing, 2018. ISBN: 978-3-319-66630-3 978-3-319-66631-0. DOI: `10.1007/978-3-319-66631-0`. (Visited on 04/23/2022).

[121] O. Calin and D.-c. E. Chang, eds. *Geometric Mechanics on Riemannian Manifolds: Applications to Partial Differential Equations*. Boston: Birkhäuser, 2005. ISBN: 978-0-8176-4354-6.

[122] B. F. Schutz. *Geometrical Methods of Mathematical Physics*. Cambridge ; New York: Cambridge University Press, 1980. ISBN: 978-0-521-23271-5 978-0-521-29887-2.

[123] J. Jost. *Geometry and Physics*. Heidelberg ; London ; New York: Springer, 2009. ISBN: 978-3-642-00540-4 978-3-642-00541-1.

[124] M. Betancourt and L. C. Stein. *The Geometry of Hamiltonian Monte Carlo*. Dec. 2011. arXiv: `arXiv:1112.4118`. (Visited on 08/18/2022).

[125] M. J. Betancourt. “A General Metric for Riemannian Manifold Hamiltonian Monte Carlo”. In: vol. 8085. 2013, pp. 327–334. DOI: `10.1007/978-3-642-40020-9_35`. arXiv: `1212.4693 [physics, stat]`. (Visited on 03/13/2023).

[126] S.-S. Chern, W.-h. Ch’en, and K. S. Lam. *Lectures on Differential Geometry*. Series on University Mathematics vol. 1. Singapore ; River Edge, N.J: World Scientific, 1999. ISBN: 978-981-02-3494-2.

[127] D. P. Bertsekas. *Convex Optimization Algorithms*. Nashua: Athena scientific, 2015. ISBN: 978-1-886529-28-1.

[128] J. Bradbury, R. Frostig, P. Hawkins, M. J. Johnson, C. Leary, D. Maclaurin, G. Necula, A. Paszke, J. VanderPlas, S. Wanderman-Milne, and Q. Zhang. *JAX: composable transformations of Python+NumPy programs*. Version 0.3.13. 2018. URL: `http://github.com/google/jax`.

[129] A. G. Baydin, B. A. Pearlmutter, A. A. Radul, and J. M. Siskind. *Automatic Differentiation in Machine Learning: A Survey*. Feb. 2018. arXiv: `arXiv:1502.05767`. (Visited on 10/02/2022).

[130] I. Foster. *Designing and Building Parallel Programs*.

[131] A. S. Tanenbaum and T. Austin. *Structured Computer Organization*. 6th ed. Boston: Pearson, 2013. ISBN: 978-0-13-291652-3.

[132] J. K. Salmon, M. A. Moraes, R. O. Dror, and D. E. Shaw. “Parallel Random Numbers: As Easy as 1, 2, 3”. In: *Proceedings of 2011 International Conference for High Performance Computing, Networking, Storage and Analysis*. Seattle Washington: ACM, Nov. 2011, pp. 1–12. ISBN: 978-1-4503-0771-0. DOI: `10.1145/2063384.2063405`. (Visited on 03/09/2023).

[133] E. Nelson. “Review of Stochastic Mechanics”. In: *Journal of Physics: Conference Series* 361 (May 2012), p. 012011. ISSN: 1742-6596. DOI: 10.1088/1742-6596/361/1/012011. (Visited on 08/18/2022).

[134] G. ’t Hooft. *The Cellular Automaton Interpretation of Quantum Mechanics*. Vol. 185. Fundamental Theories of Physics. Cham: Springer International Publishing, 2016. ISBN: 978-3-319-41284-9 978-3-319-41285-6. DOI: 10.1007/978-3-319-41285-6. (Visited on 11/07/2020).

[135] S. Wolfram. *A Project to Find the Fundamental Theory of Physics*. Champaign, IL: Wolfram Media, Inc, 2020. ISBN: 978-1-57955-035-6.

[136] N. Gao. “AB-INITIO POTENTIAL ENERGY SURFACES BY PAIRING GNNS WITH NEURAL WAVE FUNCTIONS”. In: (2022), p. 18.

[137] M. Scherbela, R. Reisenhofer, L. Gerard, P. Marquetand, and P. Grohs. *Solving the Electronic Schr\"odinger Equation for Multiple Nuclear Geometries with Weight-Sharing Deep Neural Networks*. Dec. 2021. arXiv: arXiv:2105.08351. (Visited on 01/04/2023).

[138] W. Jia, H. Wang, M. Chen, D. Lu, L. Lin, R. Car, W. E, and L. Zhang. *Pushing the Limit of Molecular Dynamics with Ab Initio Accuracy to 100 Million Atoms with Machine Learning*. Sept. 2020. arXiv: 2005.00223 [physics]. (Visited on 05/31/2022).

[139] S. Aaronson. *Quantum Computing since Democritus*. Cambridge: Cambridge University Press, 2013. ISBN: 978-0-521-19956-8.

[140] R. P. Feynman. “Plenty of Room at the Bottom”. In: (), p. 7.

[141] M. Betancourt. *A Geometric Theory of Higher-Order Automatic Differentiation*. Dec. 2018. arXiv: arXiv:1812.11592. (Visited on 10/06/2022).

[142] E. Hairer, C. Lubich, and G. Wanner. *Geometric Numerical Integration: Structure-Preserving Algorithms for Ordinary Differential Equations*. 2nd ed. Springer Series in Computational Mathematics 31. Berlin ; New York: Springer, 2006. ISBN: 978-3-540-30663-4.