

Working Title  
University of Oslo



Andreas Godø Lefdalsnes

June 19, 2019

# Abstract

Abstract goes here [**einstein**].

# Contents

<b>1</b>	<b>Introduction</b>	<b>3</b>
1.0.1	Electronic structure calculations . . . . .	3
1.0.2	Atom-centered descriptors . . . . .	4
1.0.3	Goals . . . . .	5
1.0.4	Contributions . . . . .	5
1.0.5	Structure . . . . .	5
<b>I</b>	<b>Theory</b>	<b>6</b>
<b>2</b>	<b>Quantum Mechanics</b>	<b>7</b>
2.0.1	Kets and bras . . . . .	7
2.0.2	Operators . . . . .	9
2.0.3	Time evolution . . . . .	11
2.0.4	The Schrödinger equation . . . . .	13
<b>3</b>	<b>Manybody Quantum Mechanics</b>	<b>14</b>
3.0.1	Hartree-Fock . . . . .	16
3.0.2	Density-functional theory . . . . .	19
<b>4</b>	<b>Molecular Dynamics</b>	<b>26</b>
4.0.1	From quantum mechanics to molecular dynamics . . . .	26
4.0.2	Molecular dynamics simulations . . . . .	30
4.0.3	Molecular dynamics potentials . . . . .	35
<b>5</b>	<b>Machine learning</b>	<b>38</b>
5.0.1	Basics of statistical learning . . . . .	39
5.0.2	Bias-variance decomposition . . . . .	40

5.0.3	Neural networks . . . . .	43
5.0.4	Backpropagation . . . . .	45
5.0.5	Optimization . . . . .	48
<b>6</b>	<b>Atom-centered descriptors</b>	<b>55</b>
6.0.1	Gaussian descriptors . . . . .	58
6.0.2	Deep Potential Molecular Dynamics . . . . .	59
<b>II</b>	<b>Implementation</b>	<b>62</b>
<b>7</b>	<b>Atomic Simulation Environment</b>	<b>63</b>
7.0.1	Installation . . . . .	64
7.0.2	Molecular Dynamics . . . . .	64
7.0.3	Calculators . . . . .	65
<b>8</b>	<b>Atomistic Machine-learning Package</b>	<b>67</b>
8.0.1	Theory . . . . .	68
8.0.2	Installation . . . . .	69
8.0.3	Training example . . . . .	69
8.0.4	Descriptors . . . . .	71
<b>9</b>	<b>Fitting to the Lennard-Jones potential</b>	<b>73</b>
9.0.1	Tensorflow implementation . . . . .	73
9.0.2	Comparison and relative error . . . . .	74
<b>III</b>	<b>Results</b>	<b>78</b>
<b>10</b>	<b>Empirical potentials</b>	<b>79</b>
10.0.1	Lennard-Jones . . . . .	79
10.0.2	Stillinger-Weber . . . . .	83
<b>11</b>	<b>Ab-Initio Molecular Dynamics</b>	<b>85</b>
<b>12</b>	<b>Conclusion</b>	<b>87</b>
	<b>Appendices</b>	<b>89</b>
<b>A</b>	<b>Title 1</b>	<b>90</b>



# Chapter 1

## Introduction

In recent years the field of artificial intelligence has made great strides, particularly in applications such as computer vision, natural language processing, speech recognition and user behaviour analytics. The exponential growth of available and high quality data has brought about a large demand for accurate, efficient and semi-automated algorithms, capable of building complicated mathematical models and making predictions without human supervision. Artificial neural networks in particular, partially inspired by simple models of human cognition, have found success in a large variety of task, largely owing to their ability to scale efficiently as the size of the dataset increases.

Data analysis and machine learning have become important parts of many subfields of physics, such as observational astronomy, condensed matter, biophysics and subatomic particle physics. Electronic structure calculations are well suited for the adoption of machine learning methods, as large datasets can be produced on demand and relatively free of noise which obscures patterns.

### 1.0.1 Electronic structure calculations

Electronic structure calculations - or numerical solutions to the manybody Schrödinger equation - are methods of modelling physical systems based on the principles of quantum mechanics. They involve very few parameters, and the results are found to be very accurate compared to laboratory measurements. However, exponential scaling greatly limits their applications, and we are typically limited to very small and simple systems. Approximations can be made which reduces the scaling to polynomial, but this limits the accuracy

of the method. The most successful and widely applied methods are the Density Functional Theory (DFT) method and the Hartree-Fock (HF) family of methods. The Hartree-Fock scales nominally as  $\mathcal{O}(N^4)$ , where  $N$  is the number of particles in the system (i.e. electrons), but typically as  $\mathcal{O}(N^3)$ . What does this mean, specifically? Maybe give an example of how  $N^3$  scaling is achieved, c.f. *density fitting, orbital localization, etc* The Density Functional Theory method scales in a similar manner, but with a larger proportionality term. However, Hartree-Fock requires us to keep track of the coordinates and spin of every particle, while DFT offers similar accuracy and scaling and is a function of only three spatial coordinates.

Molecular Dynamics (MD) freezes out the degrees of freedom of the electrons, and treats the atoms as point particles centered on the nucleus using a mean-field approach. The dynamics of the particles are treated as Newtonian, meaning the position and velocity can be integrated using a symplectic integrator such as Verlet integration. This treatment allows for systems much larger than ab-initio methods, but at the cost of neglecting both nuclear and electronic phenomena. However, the development of molecular dynamics potentials involves a large amount of parameters for any realistic system, and determining their functional form is more of an art than an exact science.

An application of machine learning which has shown some promise is the development of molecular dynamics potentials or Potential Energy Surfaces (PES). Using for example artificial neural networks which are universal function approximators the parameters are fully determined from the training data. This offers the potential of developing potential energy surfaces directly from ab-initio data or from well-developed and tested empirical potentials.

## 1.0.2 Atom-centered descriptors

For electronic structure calculations the potential energy, forces and other derived properties are determined by the cartesian coordinates. However, it is not sufficient to feed neural networks cartesian coordinates labeled with the energy. From physical theory and intuition any molecular dynamics potential must respect all of translation, rotation and permutation invariance. In order to apply machine learning methods we require a mapping from the cartesian coordinates to a 1D feature vector which conserves all of these properties.

The idea is to approximate the potential energy with a regression model:

$$\{\mathbf{R}\} \xrightarrow{\text{regression}} E = E(\{\mathbf{R}\}), \quad (1.1)$$

where  $\{\mathbf{R}\}$  represents the cartesian coordinates of our system. A mapping that satisfies the constraints of translation, rotation and permutation invariance we will refer to as a *descriptor*. The descriptor is a multidimensional function  $\mathbf{G}$  that serves as input to the regression method:

$$\{\mathbf{R}\} \rightarrow \mathbf{G}(\{\mathbf{R}\}) \xrightarrow{\text{regression}} E = E(\mathbf{G}(\{\mathbf{R}\})) \quad (1.2)$$

Once we have a descriptor and a regression model the dynamics can be readily obtained by taking derivatives:

$$\begin{aligned} \mathbf{F}_i &= -\nabla_i E \\ &= -\nabla_i E(\mathbf{G}(\{\mathbf{R}\})) \\ &= -\sum_j \frac{\partial E}{\partial G_j} \frac{\partial G_j}{\partial \mathbf{R}_i}, \end{aligned} \quad (1.3)$$

where we have applied the chain rule to break the gradient into derivatives with respect to the network inputs (obtained through backpropagation) and derivatives of the network inputs with respect to the coordinates of atom  $i$ . The potential energy is typically also broken into atomic contributions:

$$E = \sum_{i=1}^N E_{\text{atom}}(\mathbf{R}_i, \{\mathbf{R}\}), \quad (1.4)$$

where each atomic contribution is determined by the atom's local environment. This allows us to treat systems with a varying amount of particles without retraining the machine learning method each time.

### 1.0.3 Goals

### 1.0.4 Contributions

### 1.0.5 Structure



# Part I

## Theory

# Chapter 2

## Quantum Mechanics

In order to proceed to electronic structure calculations we require a solid foundation in the principles of quantum mechanics. This chapter will give a brief overview of the basic tenets of quantum mechanics and describe briefly how these rules lead to the Schrödinger equation, which is the equation governing all non-relativistic quantum mechanics. We will assume an undergraduate understanding of calculus and linear algebra, and some knowledge of mechanics is also helpful. The discussion in this chapter follows closely the discussion in [sakurai1995modern], and the reader is referred here for more details.

### 2.0.1 Kets and bras

In quantum mechanics, the state of a quantum system is represented by a *state vector* in a complex vector space. Such a vector is called a *ket*, denoted by  $|\alpha\rangle$ , following the notation of Paul Dirac. The state ket is postulated to contain all information about the state of the quantum system, such as energy, angular momentum, mass and so on. Two kets can be added to produce a new ket:

$$|\alpha\rangle + |\beta\rangle = |\gamma\rangle. \quad (2.1)$$

They can also be multiplied by a complex number:

$$c|\alpha\rangle = |\alpha\rangle c = |\delta\rangle. \quad (2.2)$$

If  $c$  is zero the resulting ket is called a *null ket*. If  $c$  is non-zero it is

postulated that the resulting ket contains the same information as the initial ket.

Observables such as momentum and spin are represented by operators acting on the vector space in question. Operators act on a ket from the left to produce a new ket:

$$A|\alpha\rangle = |\delta\rangle. \quad (2.3)$$

Of particular importance is when the action of an operator on a ket is the same as multiplication:

$$A|\alpha\rangle = c|\alpha\rangle = |\delta\rangle. \quad (2.4)$$

These kets are known as *eigenkets* and the corresponding complex numbers are known as *eigenvalues*. The physical state represented by an eigenket is known as an *eigenstate*. The eigenvalues of an operator  $A$  represent the only possible values of a measurement of the observable. For observables such as position and momentum, the operators will have a continuous spectrum of eigenvalues, whereas operators such as energy and spin have a discrete or *quantized* spectrum, whereby the term *quantum* mechanics is derived. The eigenkets of a physical observable form a complete orthogonal set, meaning any ket can be written as an expansion of eigenkets  $|a'\rangle$ :

$$|\alpha\rangle = \sum_{a'} c_{a'} |a'\rangle, \quad (2.5)$$

where  $c_{a'}$  is a complex coefficient. In principle there are infinitely many linearly independent eigenkets, depending on the dimensionality of the vector space.

A *bra space* is a vector space "dual" to the ket space. We postulate that for every ket  $|\alpha\rangle$  there exists a bra  $\langle\alpha|$ . The bra space is spanned by eigenbras  $\langle a'|$  corresponding to the eigenkets  $|a'\rangle$ . The ket and bra spaces have a dual correspondence:

$$\begin{aligned} |\alpha\rangle &\leftrightarrow \langle\alpha| \\ |a'\rangle, |a''\rangle, \dots &\leftrightarrow \langle a'|, \langle a''|, \dots \\ |\alpha\rangle + |\beta\rangle &\leftrightarrow \langle\alpha| + \langle\beta|. \end{aligned} \quad (2.6)$$

The bra dual to  $c|\alpha\rangle$  is postulated to be  $c^* \langle\alpha|$ , and more generally:

$$c_\alpha |\alpha\rangle + c_\beta |\beta\rangle \leftrightarrow c_\alpha^* \langle\alpha| + c_\beta^* \langle\beta|. \quad (2.7)$$

The *inner product* of a bra and a ket is a complex number written as a bra on the left and a ket on the right. It has the fundamental property:

$$\langle\alpha|\beta\rangle = \langle\beta|\alpha\rangle^*, \quad (2.8)$$

meaning they are complex conjugates. For this to satisfy the requirements of an inner product we must have

$$\langle\alpha|\alpha\rangle \geq 0, \quad (2.9)$$

with equality if and only if  $|\alpha\rangle$  is a null ket. We define the *norm* of a ket as

$$\sqrt{\langle\alpha|\alpha\rangle}, \quad (2.10)$$

which can be used to form normalized kets

$$|\tilde{\alpha}\rangle = \frac{1}{\sqrt{\langle\alpha|\alpha\rangle}} |\alpha\rangle, \quad (2.11)$$

with the property

$$\langle\tilde{\alpha}|\tilde{\alpha}\rangle = 1. \quad (2.12)$$

Two kets are said to be *orthogonal* if

$$\langle\alpha|\beta\rangle = 0. \quad (2.13)$$

## 2.0.2 Operators

As we mentioned briefly above, operators act on kets from the left to produce a new ket. Two operators  $A$  and  $B$  are equal

$$A = B, \quad (2.14)$$

if

$$A|\alpha\rangle = B|\alpha\rangle, \quad (2.15)$$

for an arbitrary ket in the relevant ket space. An operator  $A$  is said to be the *null operator* if

$$A|\alpha\rangle = 0. \quad (2.16)$$

Operators can be added, and addition operations are commutative and associative.

$$X + Y = Y + X, \quad (2.17)$$

$$(X + Y) + Z = X + (Y + Z). \quad (2.18)$$

Operators act on bras from the right to produce a new bra

$$\langle\alpha| A = \langle\beta|. \quad (2.19)$$

The ket  $A|\alpha\rangle$  and the bra  $\langle\alpha| A$  are in general not dual to each other. We define the *hermitian adjoint*  $A^\dagger$  through the dual correspondence:

$$A|\alpha\rangle \leftrightarrow \langle\alpha| A^\dagger. \quad (2.20)$$

An operator is said to be *hermitian* if

$$A = A^\dagger. \quad (2.21)$$

Hermitian operators have real eigenvalues, and since the result of any measurement must be a real number any operator that represents a physical observable must be Hermitian.

Operators can be multiplied. Multiplication is associative, but non-commutative:

$$XY \neq YX, \quad (2.22)$$

$$X(YZ) = (XY)Z. \quad (2.23)$$

The left product of a ket and a bra is known as the *outer product*:

$$|\alpha\rangle \langle\beta|. \quad (2.24)$$

The outer product should be treated as an operator, while the inner product  $\langle\alpha|\beta\rangle$  is a complex number. If an operator is to the left of a ket  $|\alpha\rangle A$  or to the right of a bra  $A\langle\beta|$  these are illegal products, in other words not

defined within the ruleset of quantum mechanics. The associative properties of operators are postulated to hold true as long as we are dealing with legal multiplications among kets, bras and operators. As an example, the outer product acting on a ket:

$$(|\alpha\rangle \langle\beta|) |\gamma\rangle, \quad (2.25)$$

can be equivalently regarded as scalar multiplication

$$|\alpha\rangle (\langle\alpha|\gamma\rangle) = |\alpha\rangle c = c |\alpha\rangle, \quad (2.26)$$

where  $c = \langle\alpha|\gamma\rangle$  is just a complex number.

### 2.0.3 Time evolution

In quantum mechanics, time is treated not as an observable, but as a parameter. Relativistic quantum mechanics treats space and time on the same footing, but only by demoting position to a parameter.

Suppose we have a physical system  $|\alpha\rangle$  at a time  $t_0$ . Denote the ket at a later time  $t > t_0$  by

$$|\alpha, t; t_0\rangle. \quad (2.27)$$

Time evolution is assumed to be continuous and symmetric, meaning that if we evolve the system backwards in time we should arrive at the initial state:

$$\lim_{t \rightarrow t_0} |\alpha, t; t_0\rangle = |\alpha\rangle. \quad (2.28)$$

The kets separated by a time  $\Delta t = t - t_0$  are related by the *time-evolution operator*  $\mathcal{U}$ :

$$|\alpha, t; t_0\rangle = \mathcal{U}(t, t_0) |\alpha, t_0\rangle. \quad (2.29)$$

If the state ket is normalized to unity at a time  $t_0$ , it must remain normalized at a later time:

$$\langle\alpha, t_0|\alpha, t_0\rangle = \langle\alpha, t; t_0|\alpha, t; t_0\rangle = 1. \quad (2.30)$$

This is guaranteed if the time evolution operator  $\mathcal{U}$  is a *unitary* operator:

$$\mathcal{U}^\dagger \mathcal{U} = 1. \quad (2.31)$$

We also require that the time evolution operator exhibits a composition property:

$$\mathcal{U}(t_2, t_0) = \mathcal{U}(t_2, t_1)\mathcal{U}(t_1, t_0), \quad (t_2 > t_1 > t_0), \quad (2.32)$$

meaning that the time evolution between two points  $t_0$  and  $t_2$  remains the same if we first evolve the system to an intermediate time  $t_1$ .

If we consider an infinitesimal time-evolution operator

$$|\alpha, t_0 + dt; t_0\rangle = \mathcal{U}(t_0 + dt, t_0) |\alpha, t_0\rangle, \quad (2.33)$$

it must reduce to the identity operator as the infinitesimal time interval  $dt$  goes to zero:

$$\lim_{dt \rightarrow 0} \mathcal{U}(t_0 + dt, t_0) = 1, \quad (2.34)$$

and we expect the difference between the operators to be of first order in  $dt$ .

These requirements are all satisfied by the operator

$$\mathcal{U}(t_0 + dt, t_0) = 1 - i\Omega dt, \quad (2.35)$$

where  $\Omega$  is a Hermitian operator:

$$\Omega^\dagger = \Omega. \quad (2.36)$$

The operator  $\Omega$  has the dimension inverse time. Frequency or inverse time is related to energy through the Planck-Einstein relation:

$$E = \hbar\omega. \quad (2.37)$$

In classical mechanics the Hamiltonian is the generator of time evolution, so we postulate that  $\Omega$  is related to the Hamiltonian operator  $H$ :

$$\Omega = \frac{H}{\hbar}. \quad (2.38)$$

The Hamiltonian operator represents the energy of our system, which is a physical observable and must therefore be Hermitian.

### 2.0.4 The Schrödinger equation

The Schrödinger equation is the fundamental equation governing non-relativistic quantum mechanics. It can be assumed as a postulate, but is usually derived from more fundamental principles. By exploiting the composition property of the time-evolution operator we find that:

$$\mathcal{U}(t + dt, t_0) = \mathcal{U}(t + dt, t)\mathcal{U}(t, t_0) = (1 - \frac{iHdt}{\hbar})\mathcal{U}(t, t_0), \quad (2.39)$$

where the time difference  $t - t_0$  is not required to be infinitesimal. Subtracting from both sides of this equation:

$$\mathcal{U}(t + dt, t_0) - \mathcal{U}(t, t_0) = -\frac{iHdt}{\hbar}\mathcal{U}(t, t_0). \quad (2.40)$$

Rearranging this equation and taking the limit  $dt \rightarrow 0$  leads to the equation:

$$i\hbar\frac{\partial}{\partial t}\mathcal{U}(t, t_0) = H\mathcal{U}(t, t_0). \quad (2.41)$$

This is known as the Schrödinger equation for the time-evolution operator. We multiply both sides by a ket  $|\alpha, t_0\rangle$ :

$$i\hbar\frac{\partial}{\partial t}\mathcal{U}(t, t_0) |\alpha, t_0\rangle = H\mathcal{U}(t, t_0) |\alpha, t_0\rangle. \quad (2.42)$$

This ket does not depend on  $t$ , leading us to the famous equation:

$$i\hbar\frac{\partial}{\partial t} |\alpha, t; t_0\rangle = H |\alpha, t; t_0\rangle. \quad (2.43)$$

This is known as the time-dependent Schrödinger equation, and gives the description for how a quantum system evolves with time. It is possible to show that in the classical limit  $\hbar \rightarrow 0$ , the expectation value of the operator  $H$  takes on the role of the energy in classical mechanics. The Schrödinger equation takes on the role of Newton's laws in quantum mechanics. However, it is not the only way to study quantum systems, as it has been shown to be an equivalent interpretation to the matrix mechanics of Werner Heisenberg and the path-integral formulation developed by Richard Feynman.



## Chapter 3

# Manybody Quantum Mechanics

The Schrödinger equation only offers exact solutions for very small and simple systems, such as the hydrogen atom or a system of non-interacting harmonic oscillators. If we want to study most systems of interest we must turn instead to numerical methods and solutions. The most prominent methods in the field of manybody quantum mechanics are the Hartree-Fock (HF) and the Density Functional Theory families of solvers. This chapter will give a brief overview of both methods, focused more on understanding and intuition than rigor. Eventually we will take these methods to be black boxes of electronic structure calculations, given some cartesian coordinates and outputting energies and forces. This section will give a brief overview of the discussion of the electronic Hamiltonian in [szabo1996modern], which also covers Hartree-Fock theory which will be expanded upon in the next section.

We want to find solutions to the non-relativistic time-independent Schrödinger equation:

$$\hat{H} |\Psi\rangle = E |\Psi\rangle, \quad (3.1)$$

with the Hamiltonian  $\hat{H}$  describing a system of nuclei and electrons with cartesian coordinates  $\mathbf{R}_a$ ,  $a = 1, 2, \dots, A$  and  $\mathbf{r}_i$ ,  $i = 1, 2, \dots, N$  respectively. The distance between nuclei  $a$  and electron  $i$  is given as the euclidean distance  $R_{ai} = |\mathbf{R}_a - \mathbf{r}_i|$  and correspondingly for the nuclei-nuclei and electron-electron distances. The full Hamiltonian for a set of  $N$  electrons and  $A$  nuclei in atomic units is:

$$\begin{aligned}
\hat{H} = & - \sum_{i=1}^N \frac{1}{2} \nabla_i^2 - \sum_{a=1}^A \frac{1}{M_a} \nabla_a^2 - \sum_{i=1}^N \sum_{a=1}^A \frac{Z_a}{R_{ia}} \\
& + \sum_{i=1}^N \sum_{j=i+1}^N \frac{1}{r_{ij}} + \sum_{a=1}^A \sum_{b=a+1}^A \frac{Z_a Z_b}{R_{ab}}.
\end{aligned} \tag{3.2}$$

The first two terms describe the kinetic energy operators of the electrons and nuclei, with  $M_a$  the ratio of the mass of nuclei  $a$  to the electron mass. The third term describes the coulomb attraction between electrons and nuclei, while the fourth and fifth terms describe the repulsion between electrons and nuclei respectively.

Since the nuclei are approximately 2000 times heavier than the electrons, the electrons can to a good approximation be described as moving in the field of fixed nuclei. In practice we neglect the kinetic energy terms of the nuclei, while considering an averaged effect from the nuclei-nuclei repulsion. The nuclei-nuclei repulsion energy averaged over adds a constant to the energy eigenvalues, but has no effect on the energy eigenfunctions. The remaining terms are known as the electronic Hamiltonian:

$$\hat{H}_e = - \sum_{i=1}^N \frac{1}{2} \nabla_i^2 - \sum_{i=1}^N \sum_{a=1}^A \frac{Z_a}{R_{ia}} + \sum_{i=1}^N \sum_{j=i+1}^N \frac{1}{r_{ij}}. \tag{3.3}$$

The electronic wavefunction  $\Psi_e = \Psi_e(\{r_i\}; \{R_a\})$  is a function of the electronic coordinates with a parametric dependence on the fixed nucleic coordinates. The electronic energy is obtained in the usual way  $E_e = \langle \Psi_e | \hat{H}_e | \Psi_e \rangle$ . The total energy of our system must now include the constant nuclear repulsion:

$$E_{tot} = E_e + \sum_{a=1}^A \sum_{b=a+1}^A \frac{Z_a Z_b}{R_{ab}}. \tag{3.4}$$

If one has solved the Schrödinger equation for the electronic Hamiltonian, one can subsequently solve for the nucleic motion using the same trick, i.e. substituting the electronic coordinates for their average values, averaged over the electronic wave function. We are then left with a nuclear Hamiltonian  $\hat{H}_n$ :

$$\begin{aligned}
\hat{H}_n &= - \sum_{a=1}^A \frac{1}{2M_a} \nabla_a^2 + \left\langle - \sum_{i=1}^N \frac{1}{2} \nabla_i^2 - \sum_{i=1}^N \sum_{a=1}^A \frac{Z_a}{R_{ia}} + \sum_{i=1}^N \sum_{j=i+1}^N \frac{1}{r_{ij}} \right\rangle \\
&\quad + \sum_{a=1}^A \sum_{b=a+1}^A \frac{Z_a Z_b}{R_{ab}} \\
&= - \sum_{a=1}^A \frac{1}{2M_a} \nabla_a^2 + E_{tot}.
\end{aligned} \tag{3.5}$$

Under this approximation the nuclei move on a potential energy surface obtained by solving the electronic Hamiltonian. We will however remain focused on the electronic structure problem. For the electronic Hamiltonian we would like to make one more approximation. The electronic Hamiltonian further simplifies if instead of considering the relative positions of the nuclei, we fix the coordinate system in the center of mass of the nucleus. This gives us the final expression we want for the electronic Hamiltonian:

$$\hat{H} = - \sum_{i=1}^N \frac{1}{2} \nabla_i^2 - \sum_{i=1}^N \frac{Z}{r_i} + \sum_{i=1}^N \sum_{j=i+1}^N \frac{1}{r_{ij}} \tag{3.6}$$

### 3.0.1 Hartree-Fock

The Hartree-Fock method is a method for finding solutions to the electronic Hamiltonian assuming the electron-electron repulsion can be approximated with a set of single-particle functions or *orbitals*, moving in a mean field generated by the presence of other electrons. The theory in this section is based upon the [sherrill2000] lecture notes from the Georgia Institute of Technology and the lecture notes on Computational Physics II [hjensen2019] from the University of Oslo. Assuming that the electrons do not interact the Hamiltonian is separable and the wavefunction is simply a product of orbitals  $\psi$  which are solutions to a onebody Hamiltonian. This gives us an ansatz for the manybody wavefunction  $\Psi$  known as the *Hartree product*:

$$\Psi(\mathbf{r}_1, \dots, \mathbf{r}_N) = \psi(\mathbf{r}_1) \cdots \psi(\mathbf{r}_N). \tag{3.7}$$

Since we are dealing with fermions this ansatz fails to satisfy the antisymmetry principle, i.e. the wavefunction is not antisymmetric with respect to the

interchange of any two particles. Fermions in addition to three spatial degrees of freedom also have a spin degree of freedom  $\sigma$  which means the fermion can be described by the space-spin coordinate  $\mathbf{x} = (\mathbf{r}, \sigma)$  with  $\mathbf{x} \in \mathbb{R}^3 \otimes \sigma$ . The problem of antisymmetry in a system of  $N$  fermions is satisfied by the introduction of *Slater determinants*

$$\Psi(\mathbf{x}_1, \dots, \mathbf{x}_N) = \frac{1}{\sqrt{N!}} \begin{vmatrix} \chi_1(\mathbf{x}_1) & \chi_2(\mathbf{x}_1) & \dots & \chi_N(\mathbf{x}_1) \\ \chi_1(\mathbf{x}_2) & \chi_2(\mathbf{x}_2) & \dots & \chi_N(\mathbf{x}_2) \\ \vdots & \vdots & \ddots & \vdots \\ \chi_1(\mathbf{x}_N) & \chi_2(\mathbf{x}_N) & \dots & \chi_N(\mathbf{x}_N) \end{vmatrix}, \quad (3.8)$$

with  $\chi(\mathbf{x})$  spin orbitals and a normalization factor  $(N!)^{-1/2}$ . The introduction of this ansatz is equivalent to assuming that all electrons move independently of each other in a mean field generated by the electron-electron repulsion. Define the one-electron operator of the electronic Hamiltonian as:

$$\hat{h}_1(\mathbf{x}_i) = -\frac{1}{2}\nabla_i^2 - \frac{Z}{r_i}, \quad (3.9)$$

with a twobody interaction term

$$\hat{v}(\mathbf{x}_i, \mathbf{x}_j) = \frac{1}{r_{ij}}, \quad (3.10)$$

which allows us to write the electronic Hamiltonian more compactly as:

$$\hat{H} = \sum_i \hat{h}_1(\mathbf{x}_i) + \sum_{i < j} \hat{v}(\mathbf{x}_i, \mathbf{x}_j). \quad (3.11)$$

The expectation value of the energy is given as the usual inner product:

$$\langle \Psi | \hat{H} | \Psi \rangle. \quad (3.12)$$

The *variational theorem* states that the expectation value of any normalized wavefunction with respect to the energy represents an upper bound to the ground state energy. This suggests a procedure wherein we vary the parameters of a set of approximate wave functions  $\Psi_T$  until an energy minimum is reached. The Hartree-Fock energy can be written in terms of integrals over the onebody and interaction terms:

$$E_{HF} = \sum_i \langle i | \hat{h} | i \rangle + \sum_{i < j} \langle ij | \hat{v} | ij \rangle_{AS}, \quad (3.13)$$

where we have introduced an antisymmetrized matrix element:

$$\langle ij|\hat{v}|ij\rangle_{AS} = \langle ij|\hat{v}|ij\rangle - \langle ij|\hat{v}|ji\rangle, \quad (3.14)$$

and the shorthand integrals:

$$\langle i|\hat{h}_1|i\rangle = \int d\mathbf{r} \chi_i^*(\mathbf{r}) \hat{h}_1 \chi_i(\mathbf{r}), \quad (3.15)$$

and

$$\langle ij|\hat{v}|ij\rangle = \int d\mathbf{r}_i d\mathbf{r}_j \chi_i^*(\mathbf{r}_i) \chi_j^*(\mathbf{r}_j) \hat{v} \chi_i(\mathbf{r}_i) \chi_j(\mathbf{r}_j). \quad (3.16)$$

In order to derive the Hartree-Fock equations we perform a linear expansion of the spin orbitals  $\chi$  in terms of a fixed orthogonal basis  $\phi$ :

$$\chi_i = \sum_{\nu} C_{i\nu} \phi_{\nu}, \quad (3.17)$$

in principle an infinite sum, but in practice truncated. This basis is usually obtained as the eigenfunctions of parts of the electronic Hamiltonian, e.g. solutions to the Schrödinger equation with a harmonic oscillator or Coulomb potential. If the coefficients belong to an orthogonal or unitary matrix, the resulting basis will preserve orthogonality. This expansion allows us to rewrite the Hartree-Fock energy as:

$$E_{HF} = \sum_i \sum_{\alpha\beta} C_{i\alpha}^* C_{i\beta} \langle \alpha|\hat{h}_1|\beta\rangle + \sum_{i<j} \sum_{\alpha\beta\delta\eta} C_{i\alpha}^* C_{j\beta}^* C_{i\delta} C_{j\eta} \langle \alpha\beta|\hat{v}|\delta\eta\rangle. \quad (3.18)$$

Using the method of Lagrangian multipliers we can define a functional to be minimized:

$$\begin{aligned} F[\Psi] &= E_{HF}[\Psi] - \sum_i \epsilon_i \langle i|j\rangle \\ &= E_{HF}[\Psi] - \sum_i \epsilon_i \sum_{\alpha} C_{i\alpha}^* C_{i\alpha}, \end{aligned} \quad (3.19)$$

with Lagrange multipliers  $\epsilon_i$ , where we have exploited the orthogonality of the basis functions to introduce the coefficients. The Lagrange multipliers are identified with the energy eigenvalues of the single-particle orbitals. Minimizing with respect to  $C_{i\alpha}^*$  yields the eigenvalue equation:

$$\sum_{\beta} \hat{h}_{\alpha\beta}^{HF} C_{i\beta} = \epsilon_i C_{i\alpha}, \quad (3.20)$$

where we have introduced the Fock matrix elements:

$$\hat{h}_{\alpha\beta}^{HF} = \langle \alpha | \hat{h}_1 | \beta \rangle + \sum_j^N \sum_{\delta\eta} C_{j\delta}^* C_{j\eta} \langle \alpha \delta | \hat{v} | \beta \eta \rangle_{AS}. \quad (3.21)$$

The single-particle integrals are usually tabulated in advance, and depend upon the choice of basis functions. Often the single-particle integrals  $\langle \alpha | \hat{h}_1 | \beta \rangle$  have analytical expressions, while the antisymmetric matrix elements must be evaluated using numerical integration. The eigenvalue problem can be written more compactly as:

$$FC = C\epsilon, \quad (3.22)$$

where  $F$  is the Fock matrix defined above,  $C$  is the matrix of coefficients and  $\epsilon$  is now a vector of single-particle energies. The Hartree-Fock equations are solved in an iterative way, starting with an initial guess for the coefficients  $C$ . Solving the eigenvalue problem yields new eigenvectors and eigenvalues. The process continues until the change in eigenvalues is within some tolerance  $\nu$ :

$$\frac{\sum_p |\epsilon_i^n - \epsilon_i^{n-1}|}{m} \leq \nu, \quad (3.23)$$

where  $p$  runs over all the single-particle energies and  $m$  is the number of single-particle states.

### 3.0.2 Density-functional theory

Density-functional theory (DFT) is a method for investigating the electronic structure of a manybody system by finding approximations to the ground state density  $n(\mathbf{r})$ . It holds many similarities to the Hartree-Fock method, since the usual method of obtaining the ground-state density involves constructing a single-determinant wave function from a set of orthonormal single-particle states, and expanding these single-particle states in terms of a known basis. However, DFT methods offer the benefit of dealing with a function of only

three spatial coordinates, while Hartree-Fock methods require you to keep track of the spatial and spin coordinates of every single-particle state.

This section is a brief summary of the material covered in the [toulouse2017] lecture notes from the Université Pierre et Marie Curie, which the reader is encouraged to check out for more detail. Our starting point is again the electronic Hamiltonian:

$$\hat{H} = - \sum_{i=1}^N \frac{1}{2} \nabla_i^2 - \sum_{i=1}^N \frac{Z}{r_i} + \sum_{i=1}^N \sum_{j=i+1}^N \frac{1}{r_{ij}}. \quad (3.24)$$

Any electronic wavefunction  $\Psi$  which solves this equation is in principle a function of  $4N$  coordinates  $\mathbf{x}_i = (\mathbf{r}_i, \sigma_i)$ ,  $i = 1, \dots, N$  where  $N$  is the number of electrons. Once we have obtained a solution to the Schrödinger equation we can obtain the one-electron density  $n(\mathbf{r})$  as:

$$n(\mathbf{r}) = N \int |\Psi(\mathbf{r}, \sigma, \mathbf{x}_2, \dots, \mathbf{x}_N)|^2 d\sigma d\mathbf{x}_2 \dots d\mathbf{x}_N. \quad (3.25)$$

Since the wavefunction is a unique functional of the Hamiltonian  $\hat{H}$ , the one-electron density is uniquely determined by the Hamiltonian. Hohenberg and Kohn showed in 1964 [hohenberg1964inhomogeneous] in their first theorem that this mapping can be inverted, i.e. that the one-electron density uniquely determines the Hamiltonian of our system (up to an arbitrary constant). Taken altogether, this means that all properties of our system, including the Hamiltonian and the manybody wavefunction are fixed by a one-electron density carrying a dependency on only 3 spatial coordinates. The electronic Hamiltonian can be rewritten as:

$$\hat{H} = \hat{F} + \hat{V}_{ne}, \quad (3.26)$$

where  $\hat{F}$  is an operator consisting of the kinetic energy and electron-electron operators and  $\hat{V}_{ne}$  is the electron-nuclei interaction. For their second theorem, Hohenberg and Kohn defined the universal density-functional:

$$F[n] = \langle \Psi[n] | \hat{F} | \Psi[n] \rangle, \quad (3.27)$$

and the total electronic energy functional:

$$E[n] = F[n] + \int \hat{V}_{ne} n(\mathbf{r}) d\mathbf{r}. \quad (3.28)$$

Hohenberg and Kohn showed that the energy functional with respect to one-electron densities  $n(\mathbf{r})$  is an upper bound to the ground state energy:

$$E_0 \leq F[n] + \int \hat{V}_{ne} n(\mathbf{r}) d\mathbf{r}, \quad (3.29)$$

with equality if and only if the one-electron density is the one-electron density corresponding to the Hamiltonian  $\hat{H}$ . This suggests a variational procedure, wherein we vary the total electronic energy functional until we reach an energy minimum:

$$E_{min} = \min_n E[n], \quad (3.30)$$

which serves as our best estimate for the ground state one-electron density  $n(\mathbf{r})$ . Levy and Lieb [**levy1979universal**, **perdew1982density**, **perdew1983physical**] proposed to redefine the universal density-functional in terms of normalized antisymmetric wavefunctions  $\Psi$  which yield a fixed density  $n$ :

$$F[n] = \min_{\Psi \rightarrow n} \langle \Psi | \hat{F} | \Psi \rangle = \langle \Psi[n] | \hat{F} | \Psi[n] \rangle, \quad (3.31)$$

wherein the minima search is performed over wavefunctions which yield the fixed density  $n$ . A search is then performed over densities  $n$  until we reach an energy minimum. This method is known as the *constrained search formulation*. Kohn and Sham proposed to decompose  $F[n]$  as:

$$F[n] = T_s[n] + E_{Hxc}[n], \quad (3.32)$$

where  $T_s[n]$  is a non-interacting kinetic-energy functional which can be defined through the constrained-search formulation:

$$T_s[n] = \min_{\Phi \rightarrow n} \langle \Phi | \hat{T} | \Phi \rangle = \langle \Phi[n] | \hat{T} | \Phi[n] \rangle, \quad (3.33)$$

wherein the minima search is now performed over normalized single-determinant wavefunctions  $\Phi$  which yield the fixed density  $n$ . The functional  $E_{Hxc}[n]$  is known as the Hartree-exchange-correlation functional. The variational procedure is now performed over single-determinant wavefunctions which yield a fixed density  $n$  and then minimized over densities:



$$\begin{aligned}
E_0 &= \min_n \left\{ F[n] + \int \hat{V}_{ne} n(\mathbf{r}) d\mathbf{r} \right\} \\
&= \min_n \left\{ \min_{\Phi \rightarrow n} \langle \Phi | \hat{T} | \Phi \rangle + E_{\text{Hxc}}[n] + \int \hat{V}_{ne} n(\mathbf{r}) d\mathbf{r} \right\} \\
&= \min_n \min_{\Phi \rightarrow n} \left\{ \langle \Phi | \hat{T} + \hat{V}_{ne} | \Phi \rangle + E_{\text{Hxc}}[n_\Phi] \right\} \\
&= \min_{\Phi} \left\{ \langle \Phi | \hat{T} + \hat{V}_{ne} | \Phi \rangle + E_{\text{Hxc}}[n_\Phi] \right\}.
\end{aligned} \tag{3.34}$$

These equations now involve only a single-determinant wave function, which is a large simplification over a variational method involving multi-determinant wave functions. Now a major part of the kinetic energy contribution can be treated through the single-determinant wave function, while only the Hartree-exchange-correlation needs to be approximated as a functional of the density. As with the Hartree-Fock method, the single determinant wavefunctions are constructed from an orthonormal basis of spin orbitals  $\chi_i(\mathbf{x})$ ,  $i = 1, \dots, M$  with spin-spatial coordinates  $\mathbf{x}_i = (\mathbf{r}_i, \sigma_i)$ . The total electronic energy can be expressed in terms of spatial orbitals  $\phi_i(\mathbf{r})$  after integrating out the spin variables:

$$E[\{\phi_i\}] = \sum_i \int \phi_i^*(\mathbf{r}) \left( -\frac{1}{2} \nabla^2 + \hat{V}_{ne} \right) \phi_i(\mathbf{r}) d\mathbf{r} + E_{\text{Hxc}}[n], \tag{3.35}$$

with the density expressed as:

$$n(\mathbf{r}) = \sum_i |\phi_i(\mathbf{r})|^2. \tag{3.36}$$

The energy minimum is obtained using the method of Lagrangian multipliers, with the constraint that the spatial orbitals be normalized we introduce the following Lagrangian:

$$\mathcal{L}[\{\phi_i\}] = E[\{\phi_i\}] - \sum_i \epsilon_i \left( \int \phi_i^*(\mathbf{r}) \phi_i(\mathbf{r}) d\mathbf{r} - 1 \right), \tag{3.37}$$

with  $\epsilon_i$  the associated Lagrangian multiplier. The energy minimum is where the Lagrangian is stationary with respect to the spatial orbitals:

$$\frac{\partial \mathcal{L}}{\partial \phi_i^*(\mathbf{r})} = 0, \quad (3.38)$$

which gives us the functional derivative:

$$\left(-\frac{1}{2}\nabla^2 + \hat{V}_{ne}\right) \phi_i(\mathbf{r}) + \frac{\partial E_{\text{Hxc}}[n]}{\partial \phi_i^*(\mathbf{r})} = \epsilon_i \phi_i(\mathbf{r}). \quad (3.39)$$

The second term can be expressed through the chain rule as:

$$\frac{\partial E_{\text{Hxc}}[n]}{\partial \phi_i^*(\mathbf{r})} = \int \frac{\partial E_{\text{Hxc}}}{\partial n(\mathbf{r}')}\frac{\partial n(\mathbf{r}')}{\partial \phi_i^*(\mathbf{r})} d\mathbf{r}'. \quad (3.40)$$

The second factor can be expressed as:

$$\frac{\partial n(\mathbf{r}')}{\partial \phi_i^*(\mathbf{r})} = \phi_i(\mathbf{r}) \delta(\mathbf{r} - \mathbf{r}'). \quad (3.41)$$

Defining the Hartree-exchange-correlation potential as the functional derivative:

$$\hat{V}_{\text{Hxc}} = \frac{\partial E_{\text{Hxc}}}{\partial \phi_i^*(\mathbf{r})}, \quad (3.42)$$

we arrive at the Kohn-Sham eigenvalue equations:

$$\left(-\frac{1}{2}\nabla^2 + \hat{V}_{ne} + \hat{V}_{\text{Hxc}}\right) \phi_i(\mathbf{r}) = \epsilon_i \phi_i(\mathbf{r}). \quad (3.43)$$

The eigenfunctions satisfying these equations are known as the Kohn-Sham orbitals, and are eigenfunctions of the Kohn-Sham one-electron Hamiltonian:

$$\hat{h}_{KS} = -\frac{1}{2}\nabla^2 + \hat{V}_{KS}, \quad (3.44)$$

with the Kohn-Sham potential defined as:

$$\hat{V}_{KS} = \hat{V}_{ne} + \hat{V}_{\text{Hxc}}. \quad (3.45)$$

The Kohn-Sham one-electron Hamiltonian defines a system of  $N$  non-interacting electrons in an effective external potential  $\hat{V}_{KS}$  ensuring that the one-electron density is the same as that of the ground-state one-electron density of a system of interacting electrons. The Hartree-exchange-correlation potential is further decomposed as:

$$\hat{V}_{\text{Hxc}} = \hat{V}_{\text{H}} + \hat{V}_{\text{XC}}, \quad (3.46)$$

where  $\hat{V}_{\text{H}}$  is the Hartree potential and  $\hat{V}_{\text{XC}}$  is the exchange-correlation potential. The Hartree potential can be expressed through the density:

$$\hat{V}_{\text{H}} = \int \frac{n(\mathbf{r}')}{|\mathbf{r} - \mathbf{r}'|} d\mathbf{r}'. \quad (3.47)$$

The exchange-correlation potential is decomposed into the exchange potential and the correlation potential:

$$\hat{V}_{\text{XC}} = \hat{V}_{\text{X}} + \hat{V}_{\text{C}}. \quad (3.48)$$

For practical calculation the spatial orbitals are expanded as a linear combination of a known basis, such as hydrogen-like functions or Gaussian-type orbitals:

$$\phi_i(\mathbf{r}) = \sum_{\nu} C_{\nu i} \chi_{\nu}(\mathbf{r}). \quad (3.49)$$

We insert these into the Kohn-Sham equations, multiply by  $\chi_i^*(\mathbf{r})$  and integrate over  $\mathbf{r}$  to arrive at the eigenvalue equations:

$$\sum_{\nu} F_{\mu\nu} C_{\nu i} = \epsilon_i \sum_{\nu} S_{\mu\nu} C_{\nu i}, \quad (3.50)$$

where  $F_{\mu\nu} = \int \chi_{\mu}^*(\mathbf{r}) \hat{h}_{KS} \chi_{\nu}(\mathbf{r}) d\mathbf{r}$  are the elements of the Kohn-Sham Fock matrix and  $S_{\mu\nu} = \int \chi_{\mu}^*(\mathbf{r}) \chi_{\nu}(\mathbf{r}) d\mathbf{r}$  are the elements of the overlap matrix of basis elements. The Fock matrix is decomposed into its constituent parts:

$$F_{\mu\nu} = H_{\mu\nu} J_{\mu\nu} + V_{XC,\mu\nu}, \quad (3.51)$$

where  $H_{\mu\nu}$  are the one-electron integrals:

$$H_{\mu\nu} = \int \chi_{\mu}^*(\mathbf{r}) \left( -\frac{1}{2} \nabla^2 + \hat{V}_{ne}(\mathbf{r}) \right) \chi_{\nu}(\mathbf{r}) d\mathbf{r}, \quad (3.52)$$

$J_{\mu\nu}$  is the Hartree-potential contribution:

$$J_{\mu\nu} = \sum_{\lambda} \sum_{\gamma} P_{\lambda\gamma} (\chi_{\mu} \chi_{\nu} | \chi_{\lambda} \chi_{\gamma}), \quad (3.53)$$

where we have defined the density matrix  $P_{\lambda\gamma}$  as:

$$P_{\gamma\lambda} = \sum_i C_{\gamma i} C_{\lambda i}^*, \quad (3.54)$$

and the two-electron integrals are defined as:

$$(\chi_\mu \chi_\nu | \chi_\lambda \chi_\gamma) = \int \int \frac{\chi_\mu^*(\mathbf{r}_1) \chi_\nu(\mathbf{r}_1) \chi_\lambda^*(\mathbf{r}_2) \chi_\gamma(\mathbf{r}_2)}{|\mathbf{r}_1 - \mathbf{r}_2|} d\mathbf{r}_1 d\mathbf{r}_2. \quad (3.55)$$

and finally the exchange-correlation potential contribution:

$$\hat{V}_{\text{XC},\mu\nu} = \int \chi_\mu^*(\mathbf{r}) \hat{V}_{\text{XC}}(\mathbf{r}) \chi_\nu(\mathbf{r}) d\mathbf{r}. \quad (3.56)$$

The Kohn-Sham equations can be written more compactly as the matrix equation:

$$FC = SC\epsilon, \quad (3.57)$$

where  $\epsilon$  is a vector containing the energy eigenvalues of the basis elements. The equations are then solved iteratively through diagonalization to obtain the matrix of coefficients  $C$ . Once we have obtained the coefficients the density  $n(\mathbf{r})$  is calculated as:

$$n(\mathbf{r}) = \sum_\gamma \sum_\lambda P_{\gamma\lambda} \chi_\gamma(\mathbf{r}) \chi_\lambda^*(\mathbf{r}), \quad (3.58)$$

where the summation is over the number of basis functions  $M$ . The total electronic energy can now be expressed as:

$$E = \sum_\mu \sum_\nu P_{\nu\mu} H_{\mu\nu} + \frac{1}{2} \sum_\mu \sum_\nu P_{\nu\mu} J_{\mu\nu} + E_{\text{XC}}. \quad (3.59)$$

Typically the energy contribution from the exchange-correlation  $E_{\text{XC}}$  has a complicated non-linear dependence on the density and must therefore be evaluated through numerical integration.

# Chapter 4

## Molecular Dynamics

The ab-initio methods discussed in the previous chapter are appropriate for systems with a relatively small amount of electrons, but suffer from cubic scaling as the number of electrons grows. If we would like to study systems of molecules, nanoscale structures or calculate transport coefficients we have to make further approximations limiting the degrees of freedom of our system. Instead of solving Schrödinger's equation to obtain the wave function, classical approximations are made to treat the atoms as point particles, with their interactions governed by a classical potential. This eliminates both the nucleonic and electronic degrees of freedom and allows for much larger systems to be simulated, though at the cost of quantum effects exhibited by the electrons. In addition, the construction of molecular dynamics potentials involves guessing at a functional form, and typically a large amount of parameters to be determined from experiments and simulations. In this chapter we will show how quantum mechanics and classical mechanics can be bridged semi-rigorously, and also give an introduction to central concepts of molecular dynamics, such as initialization, integration and construction and evaluation of potentials.

### 4.0.1 From quantum mechanics to molecular dynamics

We will follow the route described in the lecture notes [marx2000] given at the Winterschool 2000 at the John von Neuman Institute for Computing, Jölich. Our starting point is the full Hamiltonian for a set of  $N$  electrons and  $A$  nuclei:

$$\begin{aligned} \hat{H} = & - \sum_{i=1}^N \frac{\hbar^2}{2m_e} \nabla_i^2 - \sum_{a=1}^A \frac{\hbar^2}{2M_a} \nabla_a^2 - \sum_{i=1}^N \sum_{a=1}^A \frac{e^2 Z_a}{R_{ia}} \\ & + \sum_{i=1}^N \sum_{j=i+1}^N \frac{e^2}{r_{ij}} + \sum_{a=1}^A \sum_{b=a+1}^A \frac{e^2 Z_a Z_b}{R_{ab}}. \end{aligned} \quad (4.1)$$

We want to find solutions to the time-dependent non-relativistic Schrodinger equation:

$$i\hbar \frac{\partial}{\partial t} \Psi = \hat{H} \Psi. \quad (4.2)$$

The wave function is separated in terms of the electronic and nuclear coordinates with the ansatz:

$$\Psi(\{\mathbf{r}\}_i, \{\mathbf{R}\}_a, t) \approx \Phi(\{\mathbf{r}\}_i) \chi(\{\mathbf{R}\}_a) \exp \left[ \frac{i}{\hbar} \int_{t_0}^t dt' E_e(t') \right], \quad (4.3)$$

with the electronic and nuclear wave functions normalized to unity at every instance of time. A phase factor is introduced to make the equations look nice:

$$E_e = \int d\mathbf{r} d\mathbf{R} \Phi^* \chi^* \hat{H} \Phi \chi, \quad (4.4)$$

where the integration occurs over all spatial coordinates  $\{\mathbf{r}\}_i, \{\mathbf{R}\}_a$ . This is a single determinant ansatz which must lead to a mean-field description of the dynamics. Inserting this ansatz into the Schrodinger equation reveals the following set of equations:

$$i\hbar \frac{\partial \Phi}{\partial t} = - \sum_i \frac{\hbar^2}{2m_e} \nabla_i^2 \Phi + \left\{ \int d\mathbf{R} \chi^* V_{ne} \chi \right\} \Phi, \quad (4.5)$$

$$i\hbar \frac{\partial \chi}{\partial t} = - \sum_a \frac{\hbar^2}{2M_a} \nabla_a^2 \chi + \left\{ \int d\mathbf{r} \Phi^* \hat{H} \Phi \right\} \chi. \quad (4.6)$$

These coupled equations form the framework for the time-dependent self-consistent field (TDSCF) method. The electrons and nucleons move on a potential energy surface obtained from averages over the opposite class of

degrees of freedom (the nuclear and electronic wave functions respectively). In the framework of classical molecular dynamics we approximate the nuclei as classical point particles. This can be done by rewriting the nuclear wave function as

$$\chi = A \exp[iS/\hbar], \quad (4.7)$$

with an amplitude factor  $A$  and a phase  $S$  which are both considered to be real. The TDSCF equations are rewritten in terms of these variables

$$\frac{\partial S}{\partial t} + \sum_a \frac{1}{2M_a} (\nabla_a S)^2 + \int d\mathbf{r} \Phi^* \hat{H} \Phi = \hbar^2 \sum_a \frac{1}{2M_a} \frac{\nabla_a^2 A}{A}, \quad (4.8)$$

$$\frac{\partial A}{\partial t} + \sum_a \frac{1}{M_a} (\nabla_a A)(\nabla_a S) + \sum_a \frac{1}{2M_a} A (\nabla_a^2 S) = 0. \quad (4.9)$$

This set of equations is known as the "quantum fluid dynamical representation". The term for  $S$  contains a term for  $\hbar$  which vanishes in the classical limit  $\hbar \rightarrow 0$ :

$$\frac{\partial S}{\partial t} + \sum_a \frac{1}{2M_a} (\nabla_a S)^2 + \int d\mathbf{r} \Phi^* \hat{H} \Phi = 0. \quad (4.10)$$

This formulation of the nuclear dynamics is isomorphic to the Hamilton-Jacobi formulation:

$$\frac{\partial S}{\partial t} + \hat{H} = 0, \quad (4.11)$$

with the classical Hamilton function

$$\hat{H} = T(\{P_a\}) + V(\{R_a\}), \quad (4.12)$$

with coordinates  $\{R_a\}$  and conjugate momenta  $\{P_a\}$ . If we identify the conjugate momenta with the phase  $S$  as:

$$\mathbf{P}_a = \nabla_a S, \quad (4.13)$$

we obtain the following Newtonian equations of motion:

$$\begin{aligned}
\frac{d\mathbf{P}_a}{dt} &= -\nabla_a V = -\nabla_a \int d\mathbf{r} \Phi^* \hat{H} \Phi \quad \text{or} \\
M_a \frac{d^2 \mathbf{R}_I}{dt^2} &= -\nabla_a \int d\mathbf{r} \Phi^* \hat{H} \Phi \\
&= -\nabla_a V_e^E(\{\mathbf{R}_a(t)\}).
\end{aligned} \tag{4.14}$$

Under this formulation of nuclear dynamics the nuclei move according to the laws of classical mechanics in an effective potential  $V_e^E$  generated by the electrons. After averaging out the electronic degrees of freedom this potential is now only a function of the nuclear coordinates. the electronic degrees of freedom. For consistency the nuclear wave function appearing in the TDSCF equation for the electronic degrees of freedom has to be replaced by the positions of the nuclei. This is accomplished by replacing the nuclear density  $|\chi|^2$  in the limit  $\hbar \rightarrow 0$  by a product of delta functions  $\prod_a \delta(\mathbf{R}_a - \mathbf{R}_a(t))$  centered at the instantaneous positions  $\{\mathbf{R}_a(t)\}$  of the classical nuclei. This leads to a time-dependent wave equation for the electrons:

$$i\hbar \frac{\partial \Phi}{\partial t} = -\sum_i \frac{\hbar}{2m_e} \nabla_i^2 \Phi + \hat{V}_{ne} \Phi, \tag{4.15}$$

which evolve quantum mechanically as the nuclei propagate classically. This mixed approach is commonly referred to as *Ehrenfest molecular dynamics*. Under this formulation of nuclear dynamics the nuclei evolve classically while the electrons evolve according to the laws of quantum mechanics. Although the underlying equations describe a mean-field theory, the Ehrenfest approach includes transitions between electronic states. In order to arrive at a purely classical description of the dynamics of both the nuclei and the electrons we need to make further simplifications. Firstly we restrict the electronic wave function  $\Phi$  to the ground state wave function  $\Phi_0$  at every instant of time. This means the nuclei move on a single potential energy surface:

$$V_e^E = \int d\mathbf{r} \Phi_0^* \hat{H} \Phi_0 = E_0(\{\mathbf{R}_I\}), \tag{4.16}$$

that is obtained by solving the Schrodinger equation for the ground state electron wave function:

$$\hat{H} \Phi_0 = E_0 \Phi_0. \tag{4.17}$$



Since we are now dealing with a single potential energy surface, the problem of computing the energy surface can be decoupled from computing the expectation values from the electronic wave function. First one produces an appropriate set of nuclear configurations by solving the time-independent Schrodinger equation. Second, these configurations are fitted to an analytical functional form to produce a global potential energy surface. Finally the Newtonian equations of motions are solved on this energy surface, producing a set of classical trajectories. To deal with the large number of degrees of freedom as the number of nuclei in the system increases, the global potential energy surface is approximated as an expansion of manybody contributions:

$$V_e^E \approx V_e^{\text{approx}} = \sum_a v_1(R_a) + \sum_{a<b} v_2(R_a, R_b) + \sum_{a<b<c} v_3(R_a, R_b, R_c), \quad (4.18)$$

and is typically truncated at 2, 3 or 4-body interactions depending on the complexity of the atoms and molecules in the system. This renders the problem of computing dynamics purely classical:

$$M_a \frac{d^2 R_I}{dt^2} = -\nabla_a V_e^{\text{approx}}. \quad (4.19)$$

This reduction in the number of degrees of freedom is a huge simplification which allows us to study much larger and more complex systems than ab-initio methods. However, many approximations have to be made to get to this formulation of atomic/molecular mechanics, and neglecting the electronic degrees of freedom effectively precludes chemical transformations from appearing in the simulations. In addition, the analytical functional forms of the potentials usually include many parameters to be determined, and they often have to be tailored to the quantities one is trying to compute.

## 4.0.2 Molecular dynamics simulations

The theory in this and the following sections is based partly on [frenkel2001understanding], which explains the physics behind many popular methods for computer calculation and simulation. Classical molecular dynamics is a method for computing equilibrium and transport properties of manybody systems obeying classical laws of motion. While a large number of simplifications have to be made in order to describe quantum mechanical systems classically, the approximation works surprisingly well except for atoms which are quite light (He, H<sup>2</sup>) or

for atoms with a vibrational energy which is substantially larger than the thermal energy of the system ( $h\nu > k_B T$ ).

In order to calculate properties of the system they have to be expressed in terms of the positions and velocities of the constituent nuclei. For instance the temperature can be related to the average kinetic energy of the system:

$$\langle \frac{1}{2}mv^2 \rangle = \frac{N_f}{2}k_B T, \quad (4.20)$$

where  $N_f$  is the number of degrees of freedom in our system. At every instant of time the total kinetic energy of our system defines an instantaneous temperature, which has to be averaged over a large number of timesteps in order to produce the equilibrium property. In practice, one is satisfied when the fluctuations in the instantaneous temperature appear reasonably small.

To run a molecular dynamics simulation one requires a set of initial conditions, i.e. a set of initial positions and velocities for every atom in the system. Typically the atoms are placed by replicating a unit cell a number of times in every dimension. A unit cell consists of a set of lattice vectors which define the placement of every atom in the unit cell. For instance the face-centered cubic cell (FCC) contains 4 atoms:

$$\begin{aligned} \mathbf{r}_1 &= (0, 0, 0) \\ \mathbf{r}_2 &= \left(\frac{b}{2}, \frac{b}{2}, 0\right) \\ \mathbf{r}_3 &= \left(0, \frac{b}{2}, \frac{b}{2}\right) \\ \mathbf{r}_4 &= \left(\frac{b}{2}, 0, \frac{b}{2}\right), \end{aligned} \quad (4.21)$$

where  $b$  is known as the lattice constant and defines the size of the unit cells.

The velocities are typically initialized with a random uniform distribution or the Maxwell-Boltzmann distribution. The Maxwell-Boltzmann distribution is the one most often used, since the equilibrium distribution tends towards this distribution. The exact form however will differ from the one we started with.

Given these initial conditions, the system will not be in an equilibrium state at  $t = 0$ . To evolve the system to an equilibrium state one most commonly evolves the system for a number of timesteps until fluctuations in

dynamic properties such as the total potential energy or the temperature settle down. Once we are in equilibrium we can start calculating thermodynamic averages.

As we mentioned before, the global energy surface as a function of nuclear coordinates  $\{\mathbf{r}\}$  is approximated as an expansion of manybody contributions:

$$V_e^E(\{\mathbf{r}\}) = \sum_i v_1(r_i) + \sum_{i<j} v_2(r_i, r_j) + \sum_{i<j<k} v_3(r_i, r_j, r_k), \quad (4.22)$$

wherein each  $n$ -body term is an analytical function of  $n$  coordinates. As an atom moves on the energy surface it feels a force which is the gradient of the potential energy surface. This means atom  $i$  feels an acceleration:

$$\mathbf{F}_i = m_i \frac{d^2 \mathbf{r}_i}{dt^2} = -\nabla_i V_e^E(\{\mathbf{r}\}). \quad (4.23)$$

For a system of  $N$  atoms with only pairwise interactions this means the forces must be calculated  $N(N-1)/2$  times for every timestep which means we have a time complexity of order  $\mathcal{O}(N^2)$ . The force calculation is by far the most important part of any molecular dynamics simulation, and the most time consuming. A number of techniques are employed in order to reduce the time usage, perhaps the most common is the use of neighbor lists. Using neighbor lists, each atom carries a list of neighbors within a cut-off radius  $r_{cut}$  and interactions beyond this cut-off are neglected. This reduces the time-complexity to merely  $\mathcal{O}(N)$ , with a proportionality constant dependent upon the average number of neighbors in the system within a cutoff  $r_{cut}$ . For a large system this can be a huge reduction in complexity, but the choice of cut-off can obviously massively impact the dynamics of the system.

In order to simulate the dynamics of a system governed by a conservative force  $\mathbf{F} = -\nabla V_e^E$  we need to integrate the Newtonian equations of motion. The equations of motion are typically not solvable analytically, which means we require an effective numerical method for integration. Some important considerations for molecular dynamics are conservation of energy and accuracy for large time steps. The most common method used is the Velocity-Verlet algorithm. At any given time step  $t$ , the position  $\mathbf{r}(t + \Delta t)$  and velocity  $\mathbf{v}(t + \Delta t)$  at the next time step  $t + \Delta t$  is calculated as:

$$\begin{aligned}
\mathbf{r}(t + \Delta t) &= \mathbf{r}(t) + \mathbf{v}(t)\Delta t + \frac{1}{2}\mathbf{a}(t)\Delta t^2, \\
\mathbf{a}(t + \Delta t) &= -\frac{1}{m}\nabla V_e^E(\mathbf{r}(t + \Delta t)), \\
\mathbf{v}(t + \Delta t) &= \mathbf{v}(t) + \frac{1}{2}(\mathbf{a}(t) + \mathbf{a}(t + \Delta t))\Delta t.
\end{aligned} \tag{4.24}$$

The error in the Velocity-Verlet method is of order  $\mathcal{O}(\Delta t^2)$ , which means that it is not particularly accurate for large time steps over a long time. However, the long term energy drift of the method is small, which is very desirable. It is also not very memory-intensive, which matters for simulating very large systems.

Molecular dynamics is usually performed within a cubic box of fixed volume  $V = L_x \cdot L_y \cdot L_z$ , where  $L_i$  is the length of the box in direction  $i$ . Molecular dynamics is typically limited by the number of particles we are able to simulate, of the order  $10^6 - 10^8$ , which means the size of the box is often decided by the desired density  $\rho = N/V$ . Since the number of particles is always much smaller than the number of particles in realistic systems, approximations are required. Periodic boundary conditions can be applied to the box in order to approximate an infinite system. Typically particle coordinates are restricted to the simulation box, which can be expressed in pseudocode as:

---

**Algorithm 1** Continuity

---

```

if  $x < -L_x/2$  then
     $x += L_x$ 
end if
if  $x > L_x/2$  then
     $x -= L_x$ 
end if

```

---

Distance and distance vectors between particles should also obey the minimum image convention:

---

**Algorithm 2** Minimum image

---

```
dx = xj - xi
if dx < -Lx/2 then
    dx += Lx
end if
if dx > Lx/2 then
    dx -= Lx
end if
```

---

These conditions should be applied in every dimension. This approach runs the risk of introducing nonphysical artifacts of the simulation, such as a macromolecule interacting with its own image, and for coulombic interactions the system must be charge neutral to avoid summing to an infinite charge. The optimal system size with periodic boundary conditions will therefore depend on the intended simulation length, the desired accuracy and the dynamics which are being studied.

Thus far we have discussed molecular dynamics for a system of  $N$  particles within a fixed volume  $V$  and constant energy  $E$ , i.e. in the microcanonical ensemble  $NVE$ . Other ensembles are also impossible, such as the canonical ensemble  $NVT$  and the isothermal-isobaric ensemble  $NPT$ .

Simulations in the canonical ensemble can be achieved by modifying the verlet integration algorithm. The simplest thermostat possible follows from the equipartition theorem:

$$T \propto \langle mv^2 \rangle, \quad (4.25)$$

meaning some amount of kinetic energy i.e. velocity can be added or subtracted to every atom in order to maintain a constant temperature at every timestep. Multiplying every velocity by a factor  $\lambda = \sqrt{T_0/T(t)}$  where  $T(t)$  is the instantaneous temperature and  $T_0$  is the desired temperature will achieve desired effect. This approach significantly alters the trajectories of the system however, which means this thermostat should only be applied for adjusting the temperature of the system and not for taking ensemble averages in equilibrium.

The most common thermostat used is the Nosé-Hoover thermostat, which is one of the most accurate and efficient algorithms for achieving realistic constant-temperature conditions (see the lecture notes [che210D2012]).

Nosé introduced an extended Hamiltonian with two additional degrees of freedom:

- $s$  - the position of an imaginary coupled heat reservoir
- $p_s$  - the conjugate momentum of the heat reservoir

In addition it introduces an effective mass  $Q$  such that  $p_s = Q \frac{ds}{dt}$ .

Hoover modified Nosé's approach by introducing the Hamiltonian:

$$H = \frac{1}{2} \sum m_i |\mathbf{p}_i|^2 + U(\mathbf{r}) + \frac{\xi Q}{2} + 3Nk_B T \ln s, \quad (4.26)$$

where  $\xi$  is a friction coefficient and  $\mathbf{p}_i = m_i \mathbf{v}_i \times s$  are the particle momenta. This leads to a new set of Newtonian equations of motions with an additional force that is proportional to the velocity:

$$\begin{aligned} \frac{d\mathbf{r}_i}{dt} &= \mathbf{v}_i \\ \frac{d\mathbf{v}_i}{dt} &= -\frac{1}{m_i} \frac{\partial U(\mathbf{r})}{\partial \mathbf{r}_i} - \xi \mathbf{v}_i \\ \frac{d\xi}{dt} &= \left( \sum m_i |\mathbf{v}_i|^2 - 3Nk_B T \right) / Q \\ \frac{d \ln s}{dt} &= \xi. \end{aligned} \quad (4.27)$$

These can then be solved with a numerical integration scheme such as the velocity-Verlet algorithm.

### 4.0.3 Molecular dynamics potentials

The dynamics of an ensemble of particles is governed by their interactions. In molecular dynamics we stipulate that the interactions are decided only by the relative positions of the particles, i.e. only conservative forces act on the atoms. This means that the force on atom  $i$  is fully described by a potential energy  $U$ :

$$\mathbf{F}_i = -\nabla_i U(\{\mathbf{r}\}), \quad (4.28)$$

which in principle depends on the position of atom  $i$  and all other atoms in the system. Finding an appropriate potential for the system of atoms

which you intend to study can be arduous work, and usually involves fitting an analytical functional form with a large set of parameters to a potential energy surface from ab initio quantum mechanical calculations.

Potentials can be classified as either bonded or non-bonded. Bonded potentials compute the interactions for a predefined set of atoms and molecules in the simulations, while non-bonded potentials compute the interactions between *all* pairs, triplets etc. of atoms (usually within a certain radius). Larger, more complex systems typically contain a mix of bonded and non-bonded potentials, for example a system of rigid water molecules interacting with a surface of carbon atoms.

One of the simplest potentials meant to simulate a realistic system is the Lennard-Jones potential:

$$U(r_{ij}) = 4\epsilon \left( \left( \frac{\sigma}{r_{ij}} \right)^{12} - \left( \frac{\sigma}{r_{ij}} \right)^6 \right), \quad (4.29)$$

with  $r_{ij} = |\mathbf{r}_j - \mathbf{r}_i|$ . For the Lennard-Jones potential there are only two parameters to be decided, a characteristic energy  $\epsilon$  and a characteristic length  $\sigma$ . This potential is meant to emulate the relatively weak interactions between noble gas atoms such as Argon. It can be separated into two terms:

- $-\left(\frac{\sigma}{r}\right)^6$  - owing to the long-term attraction from van der Waals interactions
- $+\left(\frac{\sigma}{r}\right)^{12}$  - owing to the short-term repulsion from the Pauli principle

While the van der Waals term is justified by theory, the repulsion term is justified by numerical efficiency - as it contains the square of the van der Waals term - and because it models the Pauli repulsion accurately.

While the Lennard-Jones potential is very simple, requiring only two parameters to be determined, it has shown to be effective at modelling noble gas atoms and is commonly used as a building block for more complicated interactions.

Another simple and common potential is the Stillinger-Weber potential, which is meant to model the interactions between silicon atoms (see the lecture notes [**che8002013**]). Silicon forms tetrahedral bonded structures as well as pairwise interactions which means the potential includes a twobody and a threebody interaction:

$$U = \sum_{i < j} v_2(r_{ij}) + \sum_{i < j < k} v_3(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k). \quad (4.30)$$

The twobody interaction models the pairwise interaction:

$$v_2(r) = \begin{cases} \epsilon A (Br^{-p} - r^{-q}) \exp[(r - a)^{-1}], & r < a \\ 0 & r \geq a \end{cases} \quad (4.31)$$

This twobody term resembles the Lennard-Jones potential, but with an exponential cutoff. The threebody term models the tetrahedral angles, and is a sum over three triplets:

$$v_3(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) = h_{jik} + h_{ijk} + h_{ikj}, \quad (4.32)$$

with the angular interaction  $h_{jik} = h(r_{ij}, r_{ik}, \theta_{jik})$  and

$$h_{jik} = \begin{cases} \epsilon \lambda \exp \left[ \frac{\gamma}{r_{ij} - a} + \frac{\gamma}{r_{ik} - a} \right] (\cos \theta_{jik} - \cos \theta_{jik}^0)^2 & r_{ij} < a \\ 0 & r_{ij} \geq a \end{cases} \quad (4.33)$$

where  $\theta_{jik}^0$  is an "equilibrium" angle. The terms  $\epsilon, A, B, p, q, \lambda, \gamma$  are parameters to be decided and  $a$  is a cutoff radius. The inclusion of the threebody terms prove to be quite important in silicon maintaining its equilibrium crystal structure. The Stillinger-Weber potential agrees quite well with experiment, and its relatively simple form makes it well suited for the testing and evaluation of new potentials.



# Chapter 5

## Machine learning

Machine learning is the study of algorithms and statistical models employed by computing systems capable of performing tasks without explicit instruction. While traditional algorithms rely on some specified input and a ruleset for determining the output, machine learning is instead concerned with a set of generic algorithms which can find patterns in a broad class of data sets. This section will give a brief overview of machine learning, and more specifically the class of algorithms known as neural networks, and will follow closely the review by [mehta2019high] which the reader is encouraged to seek out for further information.

Examples of machine learning problems include identifying objects in images, transcribing text from audio and making film recommendations to viewers based on their watch history. Machine learning problems are often subdivided into estimation and prediction problems. In both cases, we choose some observable  $\mathbf{x}$  (e.g. the period of a pendulum) related to some parameters  $\boldsymbol{\theta}$  (e.g. the length and the gravitational constant) through a model  $p(\mathbf{x}|\boldsymbol{\theta})$  that describes the probability of observing  $\mathbf{x}$  given  $\boldsymbol{\theta}$ . Subsequently we perform an experiment to obtain a dataset  $\mathbf{X}$  and use these data to fit the model. Fitting the model means finding the parameters  $\hat{\boldsymbol{\theta}}$  that provide the best explanation for the data. *Estimation* problems are concerned with the accuracy of  $\hat{\boldsymbol{\theta}}$ , whereas prediction problems are concerned with the ability of the model  $p(\mathbf{x}|\boldsymbol{\theta})$  to make new predictions. Physics has traditionally been more concerned with the estimation of model parameters, while in this thesis we will be focused on the accuracy of the model.

Many problems in machine learning are defined by the same set of ingredients. The first is the dataset  $\mathcal{D} = (\mathbf{X}, \mathbf{Y})$ , where  $\mathbf{X}$  is a matrix containing

observations of the independent variables  $\mathbf{x}$ , and  $\mathbf{Y}$  is a matrix containing observations of dependent variables. Second is a model  $\mathbf{F} : \mathbf{x} \rightarrow \mathbf{y}$  which is a function of the parameters  $\boldsymbol{\theta}$ . Finally we have a cost function  $\mathcal{C}(\mathbf{Y}, \mathbf{F}(\mathbf{X}; \boldsymbol{\theta}))$  that judges the performance of our model at generating predictions.

In the case of linear regression we consider a set of independent observations  $\mathbf{X} = [\mathbf{x}_1 \ \mathbf{x}_2 \ \dots \ \mathbf{x}_N]$  related to a set of dependent observations  $\mathbf{y} = (y_1, y_2, \dots, y_N)$  through a linear model  $f(\mathbf{x}; \boldsymbol{\theta}) = x_1 \cdot w_1 + x_2 \cdot w_2 + \dots + x_P \cdot w_P$ , with parameters  $\boldsymbol{\theta} = (w_1, w_2, \dots, w_P)$ . The cost function is the well known sum of least squares  $\mathcal{C}(\mathbf{y}, f(\mathbf{X}; \boldsymbol{\theta})) = \sum_i^N (y_i - f(\mathbf{x}_i; \boldsymbol{\theta}))^2$  and the best fit is chosen as the set of parameters which minimize this cost function:  $\hat{\boldsymbol{\theta}} = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \mathcal{C}(\mathbf{Y}, f(\mathbf{X}; \boldsymbol{\theta}))$ .

### 5.0.1 Basics of statistical learning

Statistical learning theory is a field of statistics dealing with the problem of making predictions from data. We start with an unknown function  $y = f(x)$  and our goal is to develop a function  $h(x)$  such that  $h \sim f$ . We fix a hypothesis set  $\mathcal{H}$  that the algorithm is willing to consider. The expected error of a particular  $h$  over all possible inputs  $x$  and outputs  $y$  is:

$$E[h] = \int_{X \times Y} \mathcal{C}(h(x), y) \rho(x, y) dx dy, \quad (5.1)$$

where  $\mathcal{C}$  is a cost function and  $\rho(x, y)$  is the joint probability distribution for  $x$  and  $y$ . This is known as the *expected error*. Since this is impossible to compute without knowledge of the probability distribution  $\rho$ , we instead turn to the *empirical error*. Given  $n$  data points the empirical error is given as:

$$E_S[h] = \frac{1}{n} \sum_i^n \mathcal{C}(h(x_i), y_i). \quad (5.2)$$

The *generalization error* is defined as the difference between the expected and empirical errors:

$$G = E[h] - E_S[h]. \quad (5.3)$$

We say an algorithm is able to learn from data or *generalize* if

$$\lim_{n \rightarrow \infty} G = 0. \quad (5.4)$$

We are in general unable to compute the expected error, and therefore unable to compute the generalization error. The most common approach known as *cross-validation* is to estimate the generalization error by subdividing our dataset into a *training* set and a *test* set. The value of the cost function on the training set is called the *in-sample* error and the value of the cost function on the test set the *out-of-sample* error. Assuming the dataset is sufficiently large and representative of  $f$ , and the subsampling into train and test datasets is unbiased, the in-sample error can serve as an appropriate proxy for the generalization error.

In figure 5.1 we show the typical evolution of the errors as the number of data points increase. It is assumed that the function being learned is sufficiently complicated that we cannot learn it exactly, and that we have a sizeable number of data points available. The in-sample error will decrease monotonically, as our model is not able to learn the underlying data exactly. In contrast, the out-of-sample error will decrease, as the sampling noise decreases and the training data set becomes more representative of the underlying probability distribution. In the limit, these errors both approach same value, which is known the model *bias*. The bias represents the best our model could do in the infinite data limit. The out-of-sample error produced from the sampling noise is known as *variance*, and will vanish completely given an infinite representative data set.

In figure 5.2 we show the typical evolution of the out-of-sample error as the model *complexity* increases. Model complexity is a measure of the degrees of freedom in the model space, for example the number of coefficients in a polynomial regression. In the figure we can see that bias decreases monotonically as model complexity increases, as the model is able to fit a larger space of functions. However, the variance will also increase as the model becomes more susceptible to sampling noise. In general the lowest out-of-sample error, and therefore generalization error, is achieved at an intermediate model complexity. We also find that as model complexity increases, a larger amount of data points is required to be able to reasonably fit the true function.

## 5.0.2 Bias-variance decomposition

Consider a dataset  $\mathcal{D}(\mathbf{X}, \mathbf{y})$  of  $n$  pairs of independent and dependent variables. Assume the true data is generated from a noisy model:

$$y = f(\mathbf{x}) + \epsilon, \tag{5.5}$$

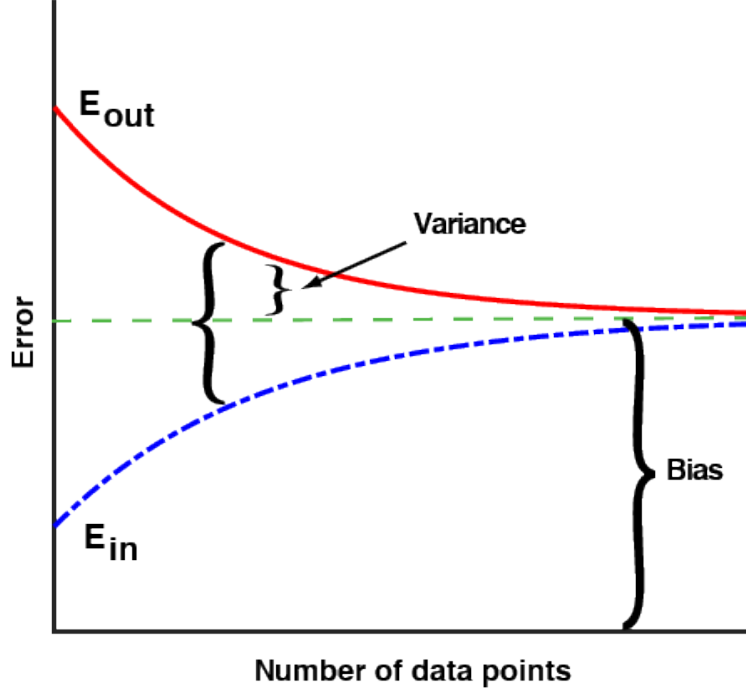


Figure 5.1: Typical in-sample and out-of-sample error as a function of the number of data points. It is assumed that the number of data points is not small, and that the true function cannot be exactly fit.

where  $\epsilon$  is normally distributed with mean  $\mu$  and standard deviation  $\sigma$ . Assume that we have an estimator  $h(\mathbf{x}; \boldsymbol{\theta})$  trained by minimizing a cost function  $\mathcal{C}(\mathbf{y}, h(\mathbf{x}))$  which we take to be the sum of squared errors:

$$\mathcal{C}(\mathbf{y}, h(\mathbf{x})) = \sum_i^n (y_i - h(\mathbf{x}_i; \boldsymbol{\theta}))^2. \quad (5.6)$$

Our best estimate for the model parameters:

$$\boldsymbol{\theta}_{\mathcal{D}} = \arg \min_{\boldsymbol{\theta}} \mathcal{C}(\mathbf{y}, h(\mathbf{x}; \boldsymbol{\theta})), \quad (5.7)$$

is a function of the dataset  $\mathcal{D}$ . If we imagine we have a set of datasets  $\mathcal{D}_j = (\mathbf{y}_j, \mathbf{X}_j)$ , each with  $n$  samples, we would like to calculate the expectation value of the cost function over all these datasets  $E_{\mathcal{D}, \epsilon}$ . We would also like to calculate the expectation value over different instances of the noise  $\epsilon$ . The

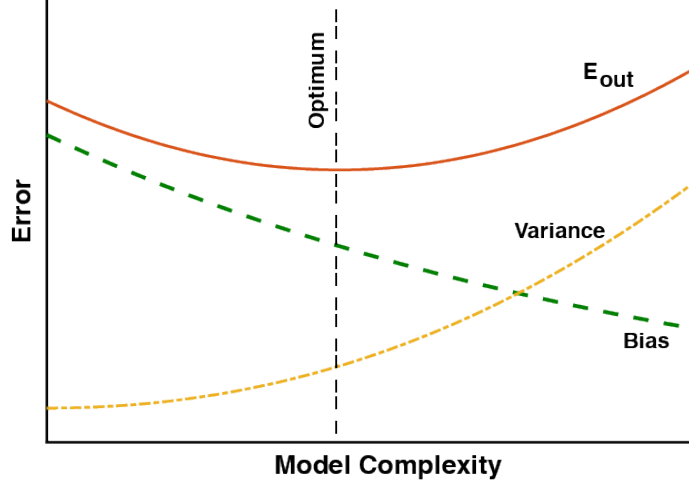


Figure 5.2: Typical out-of-sample error as a function of model complexity for a fixed dataset. Bias decreases monotonically with model complexity, while variance increases as a result of sampling noise.

expected generalization error can be decomposed as:

$$\begin{aligned} E_{\mathcal{D}, \epsilon}[\mathcal{C}(\mathbf{y}, h(\mathbf{X}; \boldsymbol{\theta}_{\mathcal{D}}))] &= E \left[ \sum_i (y_i - h(\mathbf{x}_i; \boldsymbol{\theta}_{\mathcal{D}}))^2 \right] \\ &= \sum_i \sigma_{\epsilon}^2 + E_{\mathcal{D}}[(f(\mathbf{x}_i) - f(\mathbf{x}_i; \boldsymbol{\theta}_{\mathcal{D}}))^2]. \end{aligned} \quad (5.8)$$

The second term can be further decomposed as

$$\begin{aligned} E_{\mathcal{D}}[(f(\mathbf{x}_i) - f(\mathbf{x}_i; \boldsymbol{\theta}_{\mathcal{D}}))^2] \\ = (f(\mathbf{x}_i) - E_{\mathcal{D}}[h(\mathbf{x}_i; \boldsymbol{\theta}_{\mathcal{D}})])^2 + E[(h(\mathbf{x}_i; \boldsymbol{\theta}_{\mathcal{D}}) - E[h(\mathbf{x}_i; \boldsymbol{\theta}_{\mathcal{D}})])^2] \end{aligned} \quad (5.9)$$

The first term is what we have referred to as the bias:

$$\text{Bias}^2 = \sum_i (f(\mathbf{x}_i) - E_{\mathcal{D}}[h(\mathbf{x}_i; \boldsymbol{\theta}_{\mathcal{D}})])^2. \quad (5.10)$$

The bias measures the expectation value of the deviation of our model from the true function, i.e. the best we can do in the infinite data limit.

The second term is what we have referred to as the variance:

$$\text{Var} = \sum_i E[(h(\mathbf{x}_i; \boldsymbol{\theta}_{\mathcal{D}}) - E[h(\mathbf{x}_i; \boldsymbol{\theta}_{\mathcal{D}})])^2] \quad (5.11)$$

The variance measures the deviation of our model due to finite-sampling effects. Combining these effects we can decompose the out-of-sample error into:

$$E_{\text{out}} = \text{Bias}^2 + \text{Var} + \text{Noise}, \quad (5.12)$$

with  $\text{Noise} = \sum_i \sigma_{\epsilon}^2$ .

In general it can be much more difficult to obtain sufficient good data than to train a very complex model. Therefore it is often useful in practice to use a less complex model with higher bias, because it is less susceptible to finite-sampling effects.

### 5.0.3 Neural networks

Artificial Neural Networks (ANN) or Deep Neural Networks (DNN) are supervised learning models vaguely inspired by biological neural networks. The building blocks of neural networks are neurons that take a vector input of  $d$  features  $\mathbf{x} = (x_1, \dots, x_d)$  and produce a scalar output  $a(\mathbf{x})$ . A neural networks consists of layers of these neurons stacked together with the output of one layer serving as input for another. The first layer is typically known as the *input layer*, the middle layers as *hidden layers* and the final layer the *output layer*. The basic architecture is shown in figure 5.3. In almost all cases the output  $a_i(\mathbf{x})$  of neuron  $i$  can be decomposed into a linear operation on the inputs passed through a non-linear activation function:

$$a_i(\mathbf{x}) = \sigma_i(z_i), \quad (5.13)$$

where  $\sigma_i$  is a non-linear function and  $z_i$  is the dot product between the inputs  $\mathbf{x}$  and a set of neuron-specific weights  $\mathbf{w}_i$ :

$$z_i = \mathbf{x}^T \mathbf{w}_i + b_i. \quad (5.14)$$

The term  $b_i$  is a neuron-specific re-centering of the input. Typical choices of non-linearities/activation functions include the sigmoid and hyperbolic tangent functions, and Rectified Linear Units (ReLU). When the activation function is non-linear, the neural network with a single hidden layer

can be proven to be a *universal function approximator*, given an arbitrarily large number of neurons. We typically also want functions that are monotonic and smooth with a monotonic derivative.

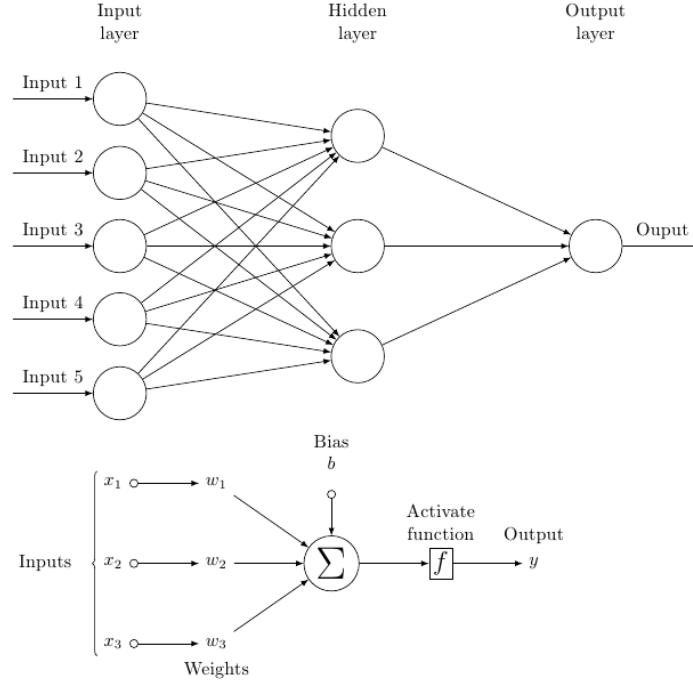


Figure 5.3: Text.

The simplest neural networks are known as *feed-forward* neural networks (FNN). The input layer is the vector  $\mathbf{x}$  of inputs, while each neuron in the first hidden layer performs a dot product between its weights  $\mathbf{w}_i$  and the inputs and passes it through a non-linearity  $\sigma_i$ . The activation function is typically shared across one or multiple layers  $\sigma_i = \sigma$ . The vector of neuron outputs  $\mathbf{a}_i$  serves as input to the next hidden layer until we reach the final layer. In the final layer the choice of activation function is dependent on the problem we are trying to solve. If we are performing non-linear regression the final activation function is often the identity  $\sigma_i(z) = z$ , or if we are doing classification the soft-max function is often employed.

Let  $\mathbf{x}$  be a vector of  $d = 1, \dots, D$  inputs or *features*. Let  $a_i^{(h)}$  denote the output of neuron  $i = 1, \dots, N_l$  in layer  $l = 1, \dots, L$ . The output of neuron  $i$  in the first hidden layer  $a_i^{(1)}$  is thus:

$$\begin{aligned} z_i^{(1)} &= \mathbf{x}^T \mathbf{w}_i^{(1)} + b_i^{(1)}, \\ a_i^{(1)} &= \sigma_i^{(1)}(z_i^{(1)}). \end{aligned} \tag{5.15}$$

The inputs are iterated through each hidden layer until we reach the final layer. Denote the vector of outputs  $\mathbf{o} = (o_1, \dots, o_O)$ :

$$\begin{aligned} z_i^{(L)} &= (\mathbf{a}_{L-1})^T \mathbf{w}_i^{(L)} + b_i^{(L)}, \\ a_i^{(L)} &= o_i \\ &= \sigma_i^{(L)}(z_i^{(L)}) \\ &= \sigma_i^{(L)} \left( (\mathbf{a}_{L-1})^T \mathbf{w}_i^{(L)} + b_i^{(L)} \right) \\ &= \sigma_i^{(L)} \left( \left( \sigma_1^{(L-1)}, \dots, \sigma_{N_{L-1}}^{(L-1)} \right)^T \mathbf{w}_i^{(L)} + b_i^{(L)} \right). \end{aligned} \tag{5.16}$$

This allows us to compose a complicated function  $\mathbf{F} : \mathbb{R}^D \rightarrow \mathbb{R}^O$ , with  $D$  the number of inputs and  $O$  the number of outputs. The *universal approximation theorem* tells us that this simple architecture can approximate any of a large set of continuous functions given appropriate choice of weights  $\mathbf{w}_i^h$  and mild assumptions on the activation functions. The theorem requires only a single hidden layer, where the strength of the approximation relies on the number of neurons. In practice it has been found that adding more layers produces faster convergence and higher accuracy, which has given rise to the field of *deep learning*.

#### 5.0.4 Backpropagation

Given a set of datapoints  $(\mathbf{x}_i, y_i)$ ,  $i = 1, \dots, n$ , the value of the cost function is entirely determined by the weights and biases of each neuron in the network. We define learning narrowly as adjusting the parameters of the network in order to minimize the cost function.

*Gradient descent* is a simple, but powerful method of finding the minima of



differentiable functions. Given a function  $F : \mathbb{R}^d \rightarrow \mathbb{R}$ , and an initial value  $\mathbf{x}_0$  we define an iterative procedure:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \eta \nabla F(\mathbf{x}_n), \quad (5.17)$$

where  $\eta$  is known as the *learning rate*. The procedure terminates when the norm  $|\nabla F(\mathbf{x}_n)|$  or  $|\mathbf{x}_{n+1} - \mathbf{x}_n|$  is appropriately small.

The learning rate is not necessarily fixed throughout the procedure, and proves crucial to the convergence of the method. If  $f$  is convex, and  $\eta$  is reasonably small, convergence is guaranteed. Convergence may be very slow however, and if  $f$  is not convex you are only guaranteed to find local minima, and this makes the method very sensitive to initial conditions.

In order to train the model, we need to calculate the derivative of the cost function with respect to a very large number of parameters multiple times. However, numerical calculation of gradients is very time consuming. The *backpropagation* algorithm is a clever use of the chain rule that allows us to calculate gradients efficiently.

Assume that there are  $L$  layers in our network with  $l = 1, 2, \dots, L$  indexing the layers, including the output layer and all the hidden layers. Let  $w_{ij}^l$  denote the weight for the connection from the  $i$ -th neuron in layer  $l - 1$  to the  $j$ -th neuron in layer  $l$ . Let  $b_j^l$  denote the bias of this  $j$ -th neuron.

The activation  $a_j^l$  of the  $j$ -th neuron in the  $l$ -th layer is related to the activities of the neurons in the layer  $l - 1$  by:

$$a_j^l = f \left( \sum_i w_{ij}^l a_i^{l-1} + b_j^l \right) = f(z_j^l), \quad (5.18)$$

where  $f$  is some activation function.

The cost function  $\mathcal{C}$  depends directly on the activations in the output layer, and indirectly on the activations in all the lower layers. Define the error  $\Delta_j^L$  of the  $j$ -th neuron in the  $L$ -th (final) layer as the change in cost function with respect to the weighted input  $z_j^L$ :

$$\Delta_j^L = \frac{\partial \mathcal{C}}{\partial z_j^L}. \quad (5.19)$$

Define analogously the error  $\Delta_j^l$  of neuron  $j$  in the  $l$ -th layer as the change in cost function with respect to the weighted input  $z_j^l$ :

$$\Delta_j^l = \frac{\partial \mathcal{C}}{\partial z_j^l}. \quad (5.20)$$

This can also be interpreted as the change in cost function with respect to the bias  $b_j^l$ :

$$\Delta_j^l = \frac{\partial \mathcal{C}}{\partial z_j^l} = \frac{\partial \mathcal{C}}{\partial b_j^l} \frac{\partial b_j^l}{\partial z_j^l} = \frac{\partial \mathcal{C}}{\partial b_j^l}, \quad (5.21)$$

since  $\partial b_j^l / \partial z_j^l = 1$ .

The error depends on neurons in layer  $l$  only through the activation of neurons in layer  $l + 1$ , so using the chain rule we can write:

$$\begin{aligned} \Delta_j^l &= \frac{\partial \mathcal{C}}{\partial z_j^l} = \sum_i \frac{\partial \mathcal{C}}{\partial z_i^{l+1}} \frac{\partial z_i^{l+1}}{\partial z_j^l} \\ &= \sum_i \Delta_i^{l+1} \frac{\partial z_i^{l+1}}{\partial z_j^l} \\ &= \sum_i \Delta_i^{l+1} w_{ij}^{l+1} f'(z_j^l) \\ &= \left( \sum_i \Delta_i^{l+1} w_{ij}^{l+1} \right) f'(z_j^l). \end{aligned} \quad (5.22)$$

The sum comes from the fact that any error in neuron  $j$  in the  $l$ -th layer propagates to all the neurons in the layer  $l + 1$ , so we have to sum up these errors.

This gives us the equations we need to update the weights and biases of our network:

$$\frac{\partial \mathcal{C}}{\partial w_{ij}^l} = \frac{\partial \mathcal{C}}{\partial z_j^l} \frac{\partial z_j^l}{\partial w_{ij}^l} = \Delta_j^l a_i^{l-1}. \quad (5.23)$$

$$\frac{\partial \mathcal{C}}{\partial b_j^l} = \Delta_j^l. \quad (5.24)$$

Now, if we have the error of every neuron  $j$  at the output layer,  $\Delta_j^L$ , equation 5.22 gives us the recipe for calculating the error in the preceding

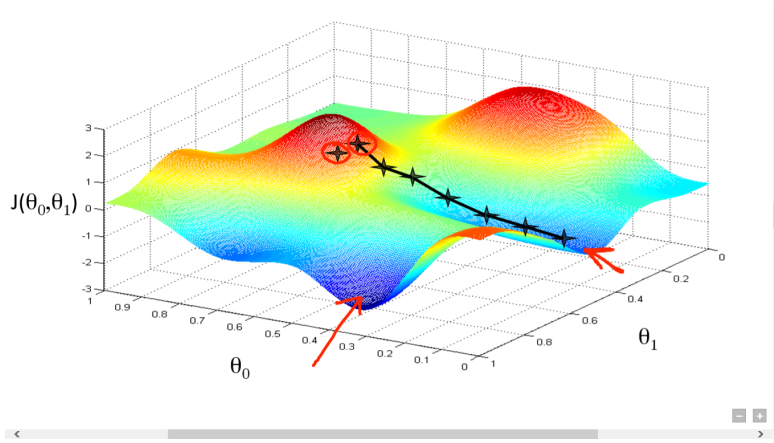


Figure 5.4: Text.

layer until we reach the first hidden layer, and we are done. All we are missing is the error at the output layer:

$$\Delta_j^L = \frac{\partial \mathcal{C}}{\partial z_j^L} = \frac{\partial}{\partial z_j^L} \left( \sum_o \frac{1}{2} (z_o^L - y_o)^2 \right) = z_j^L - y_j. \quad (5.25)$$

Now, since the derivatives for the weights depend on the activations in the preceding layer, this suggests an iterative procedure for training the network. First we feed the input data through the network and obtain activations and an output, then these outputs are backpropagated through the network to update the weights and biases. This is then repeated for some number of steps or until the network achieves acceptable accuracy.

### 5.0.5 Optimization

In order to begin the minimization procedure and find an optimal set of weights and biases for our network we first need some initial values. Often weights are initialized with small values distributed around zero, drawn from a uniform or normal distribution. The bias can be initialized to zero, but enforcing that all biases have some small value ensures that every neuron has output which can be backpropagated in the first training cycle. In the Pytorch neural network framework - which is the one we will be using - weights are initialized uniformly as

$$\begin{aligned}\sigma &= 1/N_{in} \\ w &= \mathcal{U}(-\sigma, \sigma)\end{aligned}\tag{5.26}$$

where  $N_{in}$  is the number of inputs to the weight/layer of weights and  $\mathcal{U}$  is the uniform distribution.

A critical choice for building neural networks is the choice of activation function. As we have mentioned briefly above, we have a small set of requirements in order for the neural network to be an universal function approximator, namely that the function be nonconstant, bounded, continuous and monotonically increasing. We also desire that the function be fast to evaluate, with a derivative that is simple to calculate.

A function that fulfills all of these criteria is the sigmoid function:

$$\begin{aligned}\sigma(x) &= \frac{1}{1 + e^{-x}}, \\ \sigma'(x) &= \sigma(x)(1 - \sigma(x)).\end{aligned}\tag{5.27}$$

The sigmoid function is defined on the entire real number line, and outputs a number between 0 and 1. It also has a continuous derivative which is simple to calculate. The sigmoid function is well suited for binary classifiers, since it easily creates a decision boundary between two categories 0 and 1, and is often the first goto for AI programmers.

Another simple and commonly used activation function is the hyperbolic tangent

$$\begin{aligned}\tanh(x) &= \frac{e^{2x} - 1}{e^{2x} + 1}, \\ \tanh'(x) &= (1 - \tanh(x)^2).\end{aligned}\tag{5.28}$$

The hyperbolic tangent is defined on the real number line and outputs a number between  $-1$  and  $1$ .

Both the sigmoid and the hyperbolic tangent functions exhibit the problem of vanishing gradients. In deep learning, you typically have several layers of neurons outputting some linear combination of the activation functions. Through the backpropagation algorithm, each neuron receives a weight update proportional to the partial derivative of the loss with respect to its weights.

For the sigmoid and hyperbolic tangent activations these gradients will be in the range  $(-1, 1)$ , and therefore often exceedingly small. This means that many neurons may receive effectively no weight update in any given training epoch, which severely impedes training.

The Rectified Linear Unit (ReLU) has gained popularity in recent years for its effectiveness in the field of convolutional neural networks. The ReLU function is defined as:

$$\begin{aligned}\text{ReLU}(x) &= \begin{cases} x & x \geq 0 \\ 0 & x < 0 \end{cases} \\ \text{ReLU}'(x) &= \begin{cases} 1 & x \geq 0 \\ 0 & x < 0 \end{cases}\end{aligned}\tag{5.29}$$

Since the ReLU function in principle has no upper bound, it does not suffer the effect of vanishing gradients. Instead, the ReLU function can produce exploding gradients, if one or more activations becomes very large. In practice the gradients are often "clipped", i.e. truncated above a certain upper bound. With the ReLU function one can also encounter "dying" neurons, since the input to any neuron that does not exceed zero is set to zero, which means that the number of neurons receiving an effective weight update

From the point of view of the neural network, the inputs are just dimensionless numbers, which are passed through layers until it spits out some outputs. It is therefore often useful to standardize the inputs. A common method is to shift the inputs by their mean and normalize by their standard deviation:

$$x'_i = \frac{x_i - \bar{x}}{\sigma_x}.\tag{5.30}$$

This is often useful in speeding up the training of neural networks. In theory any shifting and rescaling can be reproduced simply by updating the weights and biases of the network. In practice however, as the weights often start as small numbers, it can take quite a few training cycles before the weights are of appropriate size. Standardizing the inputs also makes it impose to impose a penalty on the weights in order to reduce overfitting.

The most common methods for training neural networks are variations on the simple gradient descent scheme as mentioned above:

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \eta \nabla F(\mathbf{x}_n). \quad (5.31)$$

Since we are interested in training neural networks,  $\mathbf{x}_n$  represents the weights and biases of the network after  $n$  training cycles,  $\eta$  is our learning rate and  $\nabla F(\mathbf{x}_n)$  is the gradient of the loss function.

The learning rate  $\eta$  controls how fast we reach a given minimum. A small learning rate will require more training cycles and a larger learning rate will require fewer. The learning rate also controls the numerical stability of the method, if it is too large the weights may diverge and if it is too small the weight may never converge. Gradient descent also tends to exhibit oscillating behaviour around the minimum. Often a schedule is imposed on the learning rate, such as reducing it by a fixed amount every few epochs or incorporating an exponentially decaying learning rate. [cite: Mehta et al] As discussed in Mehta et. al. using gradient descent to optimize our neural network imposes some limitations:

- *Gradient descent finds local minima* - if the GD algorithm converges, it will converge to a local minimum. This can lead to poor performance in complicated cost function landscapes
- *Gradients are expensive to compute* - the loss function often includes a term for each data point, which means the gradient also does.
- *GD is sensitive to learning rate* - as mentioned above, GD is very sensitive to learning rate
- *GD is sensitive to initial conditions* - gradient descent can take two different initial values and converge at two drastically different values
- *GD does not take curvature into account* - the learning rate for gradient descent is the same in all directions of parameter space, which means the maximum learning rate is set by the behaviour of the steepest direction

A common method for ameliorating some of these concerns is minibatching the inputs. Instead of calculating the gradient on the entire dataset for every epoch, we instead calculate the gradient on a subset of the data called a minibatch. If there are  $n$  datapoints and a minibatch size of  $m$  the total number of batches is  $n/m$ . If we denote each minibatch  $b_k$ ,  $k = 1, 2, \dots, n/m$  the gradient becomes:

$$\nabla \mathcal{C} = \frac{1}{n} \sum_{i=1}^N \nabla \mathcal{L}_i \rightarrow \frac{1}{m} \sum_{i \in b_k} \nabla \mathcal{L}_i, \quad (5.32)$$

i.e. instead of averaging the loss over the entire dataset we instead average over a minibatch. This significantly speeds up the calculation, since we do not use the entire dataset for every update. In addition, introducing stochasticity in the division of the dataset into minibatches introduces stochasticity to the weights update, which should help gradient descent in overcoming local minima.

It is also common to introduce regularization to the weights and biases of the networks. Regularization in the context of neural networks means adding a term proportional to the  $L_p$  norm of the weights and biases. For example using the L2-norm our cost function becomes:

$$\nabla \mathcal{C} = \frac{1}{n} \sum_i \nabla \mathcal{L}_i \rightarrow \frac{1}{n} \sum_i \nabla \mathcal{L}_i + \lambda \|\mathbf{w}\|^2 = \frac{1}{n} \sum_i \nabla \mathcal{L}_i + \lambda \sum_{ij} w_{ij}^2, \quad (5.33)$$

i.e. we sum up all the weights squared. The parameter  $\lambda$  is known as a regularization parameter. This extra term adds a penalty on the size of the weights dependant on the regularization parameter. This has been shown to reduce overfitting, as the weights cannot be adjusted to arbitrary size in order to fit the training dataset.

Gradient descent is often paired with a momentum parameter that serves as a "memory" of the direction we are moving. This can be implemented as a modification to the gradient descent update:

$$\begin{aligned} \mathbf{v}_n &= \gamma \mathbf{v}_{n-1} + \eta \nabla F(\mathbf{x}_n), \\ \mathbf{x}_{n+1} &= \mathbf{x}_n - \mathbf{v}_n, \end{aligned} \quad (5.34)$$

where  $0 < \gamma < 1$  is a memory parameter that controls the time scale for the memory. It is termed a momentum parameter because the equations for updating the parameters  $\mathbf{x}$  are analogous to the equations of motion for a particle moving in a viscous medium. Momentum helps the gradient descent algorithm gain speed in directions where the curvature is flat while dampening speed in high curvature directions.

First-order gradient descent methods differ from quasi-Newton methods in that they do not keep track of the curvature which is encoded in the so-called

Hessian matrix of second order derivatives. Second-order methods accomplish this by calculating or approximating the Hessian and normalizing the learning rate by the curvature.

The RMSprop algorithm keeps a running average of both the first and second moment of the gradient. If we term the gradient as  $\mathbf{g} = \nabla F(\mathbf{x})$  and the first and second orders as  $\mathbf{m} = \mathbb{E}[\mathbf{g}]$  and  $\mathbf{s} = \mathbb{E}[\mathbf{g}^2]$  we can write the update rule as:

$$\begin{aligned}\mathbf{g}_n &= \nabla F(\mathbf{x}_n) \\ \mathbf{s}_n &= \beta \mathbf{s}_{n-1} + (1 - \beta) \mathbf{g}_n^2 \\ \mathbf{x}_{n+1} &= \mathbf{x}_n - \eta \frac{\mathbf{g}_n}{\sqrt{\mathbf{s}_n + \epsilon}}\end{aligned}\tag{5.35}$$

where  $\beta$  is a parameter which controls the time scale of the averaging and  $\epsilon$  is a small regularization constant to prevent divergences. Operations involving vectors are understood to be element-wise. This results in a dampening of the learning rate in directions where the gradient is consistently large, and allows us to use a larger learning rate for areas of flat curvature.

Perhaps the most commonly used optimizer today is the closely related ADAM optimizer. In addition to keeping a running average of the first and second-order moments ADAM performs an additional bias correction to account for the fact that we are estimating first and second order moments using a running average. The update rule is given as:

$$\begin{aligned}\mathbf{g}_n &= \nabla F(\mathbf{x}_n) \\ \mathbf{m}_n &= \beta_1 \mathbf{m}_{n-1} + (1 - \beta_1) \mathbf{g}_n \\ \mathbf{s}_n &= \beta_2 \mathbf{s}_{n-1} + (1 - \beta_2) \mathbf{g}_n^2 \\ \hat{\mathbf{m}}_n &= \frac{\mathbf{m}_n}{1 - \beta_1^n} \\ \hat{\mathbf{s}}_n &= \frac{\mathbf{s}_n}{1 - \beta_2^n} \\ \mathbf{x}_{n+1} &= \mathbf{x}_n - \eta \frac{\hat{\mathbf{m}}_n}{\sqrt{\hat{\mathbf{s}}_n + \epsilon}},\end{aligned}\tag{5.36}$$

with  $\beta_1$  and  $\beta_2$  as memory parameters for the first and second moments respectively. This update rule effectively normalizes the learning rate by the standard deviation of the gradient, which is a natural measurement scale.



This also has the effect of cutting off directions where the gradient varies by a lot.

## Chapter 6

# Atom-centered descriptors

Electronic structure calculations are among the most computationally intensive scientific calculations, and the rapid development of modern computers have made many previously unthinkable computer simulations a central part of the scientific toolkit. Because of their computationally demanding nature, electronic structure calculations today occupy a large portion of resources on modern scientific supercomputer facilities.

For small-scale systems, we have discussed the Hartree-Fock and Density Functional Theory methods as the primary workhorses of ab-initio electronic structure calculations. However, these methods suffer from very poor scaling as the system size increases, with Hartree-Fock naively scaling as  $\mathcal{O}(N^4)$  with  $N$  the number of electrons and Density Functional Theory scaling as  $\mathcal{O}(N^4)$ , but with a larger proportionality scaling.

In many cases the exact details of the electronic structure are less important than the long-time behaviour of the atoms and molecules involved in the simulation, and classical approximations can be made as in molecular dynamics, which comes close to linear scaling. This allows us to simulate systems of up to millions or hundreds of millions of atoms, which can approximate nano- or micro-scale systems if periodic boundary conditions are applied. However, the question remains as to how you develop an accurate classical potential which can accurately reproduce fundamentally quantum systems, with a speed that allows us to enter into realistic timescales (i.e. nano or microseconds).

The most common approach to developing molecular dynamics potentials is to guess a functional form based on your physical intuition and experience with the systems and calculate appropriate parameters from data obtained

from DFT calculations. The number of parameters involved can range from two in the case of Lennard-Jones or hundreds of parameters in the case of complex, many-atom potentials such as the AMBER and CHARMM force fields. The imposition of functional forms to quantum data is an artform, and the potential must often be tailored to not only the chemical species and number of atoms in your system, but also the specific experimental quantity you are trying to extract, such as the energy, radial distribution function or transport coefficients. Notably there are dozens of MD potentials describing different models of water (H<sub>2</sub>O), each fine-tuned for a specific system structure or parameter.

Due to recent developments in the field of machine learning, the question has been raised as to how it may be possible to automate the process of developing potentials. In their article introducing the Atomistic Machine-Learning Package (AMP) [khorshidi2016amp] outline one potential way forward. The idea is to approximate the potential energy with a regression model:

$$\{\mathbf{R}\} \xrightarrow{\text{Regression}} E = E(\{\mathbf{R}\}), \quad (6.1)$$

where  $\{\mathbf{R}\}$  is the set of nuclear coordinates of our system. Most machine learning methods operate over a set of one-dimensional so-called *feature vectors*, where every vector element represents a feature of the data set. For example the amount of precipitation in a given area at a given time is a function of features such as humidity, cloud cover, air pressure etc. This is a vector of some length  $F$ , while the nuclear coordinates represent a point in  $3N$ -dimensional phase space. This difference in representation requires some way of mapping the nuclear coordinates to features which can be employed by a machine learning method.

The naive approach would be to simply feed in the nuclear coordinates as a 1D vector, and then perform a regression on the dataset in order to obtain the potential energy. However, our physical intuition imposes some constraints on the potential energy. In particular, the potential energy of a microscopic system should be translationally, rotationally and permutationally invariant.

Translational invariance implies that the addition of any three-dimensional vector to every coordinate in the system should not in any way alter the potential energy of the system. This should not be the case with a naive mapping, as for a given set of weights (or equivalent) smaller/larger coordi-

nates values would be mapped to smaller/larger activations, and therefore alter the final output. Rotational invariance implies that the potential energy of the system should not change as the system is rotated about an axis. This also should not be the case in the context of a naive mapping, as any change to any of the inputs would be mapped in a non-linear way to produce a different output. Finally, permutation invariance implies that swapping the coordinates of any two atoms of the same chemical species would produce the same potential energy. This should also not be the case, for the same reasons as we just discussed. These constraints together heavily restrict the functional form that any mapping to the potential energy could have, which means a more careful analysis should be considered.

In order for the mapping to be applicable to systems of varying size, a decomposition into atomic energy contributions is performed:

$$E(\{\mathbf{R}\}) = \sum_{i=1}^N E_{\text{atom}}(\{\mathbf{R}\}). \quad (6.2)$$

The individual energy contributions  $E_{\text{atom}}$  are then approximated by performing a regression analysis. The atomic energy contributions are usually limited to its local environment through the introduction of a cutoff radius  $R_c$ :

$$E_{\text{atom}}(\{\mathbf{R}\}) \approx E_{\text{atom}}(\mathbf{R}_i, \{\mathbf{R}\}_j \mid |\mathbf{R}_{ij}| < R_c) \quad (6.3)$$

meaning that interactions are only treated if the interatomic distance is smaller than the cutoff. This is a good approximation for a sensible choice of cutoff radius if no electrostatic interactions are involved. Long-range interactions can also be treated through the introduction of methods such as Ewald summation, but this will not be discussed here.

A mapping that satisfies the above constraints we will refer to as a *descriptor*, and is used as input to the regression method:

$$\{\mathbf{R}\} \rightarrow \mathbf{G}(\{\mathbf{R}\}) \xrightarrow{\text{regression}} E_{\text{atom}} = E_{\text{atom}}(\mathbf{G}(\{\mathbf{R}\})). \quad (6.4)$$

Once we have a descriptor and a regression model the dynamics can be readily obtained by taking derivatives:

$$\begin{aligned}
\mathbf{F}_i &= -\nabla_i E \\
&= -\nabla_i \sum_i^{\text{local}} E_{\text{atom}}(\mathbf{G}(\{\mathbf{R}\})) \\
&= -\sum_i^{\text{local}} \sum_j \frac{\partial E_{\text{atom}}}{\partial G_j} \frac{\partial G_j}{\partial \mathbf{R}_i},
\end{aligned} \tag{6.5}$$

where we have applied the chain rule to break the gradient into derivatives with respect to the network inputs (obtained through backpropagation) and derivatives of the network inputs with respect to the coordinates of atom  $i$ . Once we have the forces the system can be propagated through time using the Velocity-Verlet equations.

### 6.0.1 Gaussian descriptors

In their paper on neural-network representations of potential energy surfaces [behler2007generalized] suggested the decomposition of the mapping  $\mathbf{G}_i$  of atom  $i$  into two subvectors  $\mathbf{G}_i^I$  and  $\mathbf{G}_i^{II}$  representing pairwise and three-body interactions respectively. The components of  $\mathbf{G}_i^I$  are comprised of sums of gaussian functions of the pairwise distance  $R_{ij}$ :

$$f_i^I = \sum_{j \neq i}^{\text{local}} \exp(-\eta(R_{ij} - R_s)^2/R_c^2) f_c(R_{ij}), \tag{6.6}$$

with the sum over the local environment of atom  $i$ . The parameters  $\eta$  and  $R_s$  represent the width and center of the gaussian functions respectively. The term  $f_c$  is a cutoff function which decays smoothly to zero at the cutoff radius. Behler and Parrinello proposed the following cutoff:

$$f_c(R) = \begin{cases} \frac{1}{2} (1 + \cos(\pi R/R_c)) & R < R_c \\ 0 & R > R_c, \end{cases} \tag{6.7}$$

however other functional forms are possible. For the angular part such terms are multiplied, which means the function decays much more rapidly as we approach the cutoff. The only requirement we pose is that the function be continuous with a continuous first derivative in  $r \in [0, \infty)$ , approach one as  $R \rightarrow 0$  and zero as  $R \rightarrow R_c$ .

The components of the three-body subvector are defined incorporating the angles  $\theta_{ijk}$  between every triplet of atoms:

$$f_i^{II} = 2^{1-\zeta} \sum_{j,k \neq i}^{\text{local}} (1 + \lambda \cos \theta_{ijk})^\zeta \exp \left( -\eta (R_{ij}^2 + R_{ik}^2 + R_{jk}^2) / R_c^2 \right) \times f_c(R_{ij}) f_c(R_{ik}) f_c(R_{jk}). \quad (6.8)$$

The components for each subvector is calculated by varying the different parameters  $\eta, R_s, \zeta, \lambda$ . The choice of neither symmetry functions, cutoff, nor the parameters employed by Behler and Parrinello are unique. The guiding wisdom is that atomic environments with different potential energies should give differing energies, while remaining invariant under translation, rotation and permutation. Finally we note that the descriptors and the neural network models are not interchangeable, as each neural network is trained for a specific set of input vectors, and must be retrained if the way inputs are composed changes.

## 6.0.2 Deep Potential Molecular Dynamics

Deep Potential Molecular Dynamics (DPMD) is a method proposed by [PhysRevLett.120.143001] in response to the successes of methods such as Behler-Parrinello, Gaussian Approximation Potentials (GAP) and Gradient-Domain Machine Learning (GDML). These methods all involve some amount of handcrafting the inputs, and building these inputs for larger, more complex systems is not straightforward. The Deep Potential method assigns a local environment and reference frame to each atom. The total potential energy is a sum of atomic contributions as before:

$$E = \sum_i E_i, \quad (6.9)$$

with the atomic energy determined by its nearest neighbors:

$$E_i = E(\mathbf{R}_i, \{\mathbf{R}\}_j : |\mathbf{R}_{ij}| < R_c). \quad (6.10)$$

The position of each neighbor of atom  $i$  is described by the relative positions  $\mathbf{R}_{ij} = \mathbf{R}_j - \mathbf{R}_i$ , which preserves translational symmetry.

Rotational symmetry is conserved by constructing a local frame for each atom. Two neighboring atoms  $a$  and  $b$  are picked by a user-specified rule (default: two closest). The environment of atom  $i$  is then described by three unit vectors:

$$\begin{aligned} \mathbf{e}_{i1} &= \mathbf{e}(\mathbf{R}_{ia}), \\ \mathbf{e}_{i2} &= \mathbf{e}(\mathbf{R}_{ib} - (\mathbf{R}_{ib} \cdot \mathbf{e}_{i1})\mathbf{e}_{i1}), \\ \mathbf{e}_{i3} &= \mathbf{e}_{i1} \times \mathbf{e}_{i2}, \end{aligned} \tag{6.11}$$

where  $\mathbf{e}(\mathbf{R})$  denotes the normalized vector  $\mathbf{e}(\mathbf{R}) = \mathbf{R}/|\mathbf{R}|$ . Together these vectors form an orthonormal basis for the reference frame of atom  $i$ . The local coordinates  $\mathbf{R}_{ij}$  can then be obtained from the global coordinates  $\mathbf{R}_{ij}^0$  through the transformation:

$$\mathbf{R}_{ij} = \mathbf{R}_{ij}^0 \cdot \mathcal{R}, \tag{6.12}$$

where  $\mathcal{R} = [\mathbf{e}_{i1} \ \mathbf{e}_{i2} \ \mathbf{e}_{i3}]$  is a rotation matrix with columns given by the local basis vectors.

The neural network input vector for every atom-to-atom interaction  $\mathbf{D}_{ij}$  can be given with radial-only or full radial-angular information:

$$D_{ij}^\alpha = \begin{cases} \left( \frac{1}{R_{ij}}, \frac{\mathbf{R}_{ij}}{|\mathbf{R}_{ij}|} \right) & \text{full information,} \\ \left( \frac{1}{R_{ij}} \right) & \text{radial-only information,} \end{cases} \tag{6.13}$$

with  $\alpha = 0$  when only radial information is specified and  $\alpha = 0, 1, 2, 3$  when full information is provided. Radial information is typically sufficient for long-range interactions such as van-der-Waals forces, while covalent bonding can be modeled by including only the closest atoms. We therefore specify two separate cutoff shells, one for the radial information  $R_c$  and one for treating angular interactions  $R_a$ .

In order to preserve permutation symmetry the inputs vectors  $\mathbf{D}_{ij}$  are sorted first according to chemical species, and then within each chemical species according to their inverse distance  $1/R_{ij}$ . The vector of subvectors  $\mathbf{D}_i$  is then fed through a neural network to produce the atomic energy contribution. The network input size is fixed according to the maximum number of neighbors in the system which is being studied, with  $\mathbf{D}_{ij} = \mathbf{0}$  if there are fewer neighbors within the radial cutoff.

The authors also propose a scheme of force learning, wherein the force Root Mean Square Error (RMSE) is incorporated into the Loss Function:

$$\mathcal{L} = \frac{p_\epsilon}{N} \sum_i (\Delta E_i^2) + \frac{p_f}{3N} \sum_i |\Delta \mathbf{F}_i|^2, \quad (6.14)$$

with  $p_\epsilon$  and  $p_f$  energy and force pre-factors which are adjusted throughout the learning process. The term  $\Delta E_i$  denotes the error between the sum of network outputs and the correct potential energy, while the term  $\Delta \mathbf{F}_i$  denotes the error in the force output. The pre-factors are adjusted based on the learning rate:

$$p(t) = p^{\text{limit}} \left[ 1 - \frac{r_l(t)}{r_l^0} \right] + p^{\text{start}} \left[ \frac{r_l(t)}{r_l^0} \right], \quad (6.15)$$

with  $r_l(t)$  and  $r_l^0$  the learning rate at time step  $t$  and time step 0 respectively. The learning rate decays exponentially as:

$$r_l(t) = r_l^0 \cdot d_r^{t/d_s}, \quad (6.16)$$

with  $d_r$  the decay rate and  $d_s$  the decay steps. The decay rate should be less than 1. The force error is often a magnitude or two larger than the energy error, and it is believed that incorporating the force into the loss should improve the learning rate for physics-based applications which incorporate forces. The virial information can also be treated in this manner, but this will not be discussed here.



# Part II

## Implementation

# Chapter 7

## Atomic Simulation Environment

The Atomic Simulation environment (ASE<sup>1</sup>) [**larsen2017atomic**] is a software package written in Python for the purpose of setting up, steering and analyzing atomistic simulations. Python is an interpreted, high-level general purpose language, with a powerful, concise syntax which allows one to perform very complex tasks with few lines of code. Python can also easily be extended and interfaced with fast and mature libraries. The modular interface of Python makes ASE easily extensible: in particular the calculator interface for evaluating energies, forces and much more has been implemented for software packages such as LAMMPS, VASP, Quantum Espresso and many more. The Atomic Simulation Environment is intended to be:

- Easy to use
- Flexible
- Customizable
- Pythonic
- Open to participation

The real drawback of Python is that it is an interpreted language, which results in slow execution. It can also be quite memory-intensive, which makes

---

<sup>1</sup> <https://wiki.fysik.dtu.dk/ase/index.html>

pure Python unsuitable for large scale computations and simulations. It is therefore common to write the computationally demanding tasks in a lower level compiled language, and build a Python interface for calling functions and classes.

### 7.0.1 Installation

ASE requires an installation of

- Python 2.7, 3.4-3.6
- Numpy 1.9 or newer
- Scipy 0.14 or newer

This can be easily obtained through the Anaconda or Miniconda packages<sup>2</sup>, or follow the instructions on the Python website<sup>3</sup>. Once you have the prerequisites ASE can be installed using pip:

```
pip install ase
```

### 7.0.2 Molecular Dynamics

Here we will demonstrate how to setup a simple Argon crystal, set the velocities and integrate the system using the Velocity Verlet equations. First we import some prerequisites and define the system:

```
from ase.lattice.cubic import FaceCenteredCubic
from ase import units
from ase.md.velocitydistribution import
    MaxwellBoltzmannDistribution
from ase.md.verlet import VelocityVerlet

symbol = "Ar"
size = (3, 3, 3)
atoms = FaceCenteredCubic(symbol=symbol, size=size, pbc=
    True)
MaxwellBoltzmannDistribution(atoms, 300 * units.kB)
```

---

<sup>2</sup><https://anaconda.org/>

<sup>3</sup><https://www.python.org/>

This defines a face-centered-cubic (FCC) crystal unit cell with 4 atoms, and a system size of  $3 \times 3 \times 3$  unit cells for a total of  $4 \cdot 3^3 = 108$  atoms with periodic boundary conditions. We thereafter give the atoms velocities according to the Maxwell-Boltzmann distribution such that the system temperature is approximately 300 Kelvin.

Next we give the atoms a Lennard-Jones *calculator* for calculating dynamics and create a Velocity Verlet *integrator* for operating on the atoms according to their forces. Calculators will be discussed in the next section.

```
calc = LennardJones(sigma=3.405, epsilon=1.0318e-2)
atoms.set_calculator(calc)
dyn = VelocityVerlet(atoms, 5 * units.fs)
```

The parameters  $\sigma$  and  $\epsilon$  define the well known length and energy scales for the Lennard-Jones potential. The Velocity Verlet integrator is initialized with a timestep of 5 femtoseconds, which strikes a balance between accuracy and the timescale of the simulation. Finally we run the system for 1000 steps:

```
for i in range(100):
    dyn.run(10)
```

Every 10 timesteps the system can be printed, written to file or some other form of analysis and processing.

### 7.0.3 Calculators

The calculator interface gives ASE an easy to use, flexible and customizable way of computing dynamics, and makes ASE viable for a wide range of electronic structure calculations. For ASE a calculator is a black box that receives positions, atomic numbers and so on and outputs energies, forces, stresses; all that is required to perform atomistic simulations. The basic interface is defined in the ASE source code as:

```
import numpy as np

class Calculator:
    def get_potential_energy(self, atoms=None,
                           force_consistent=False):
        return 0.0
```

```

def get_forces(self, atoms):
    return np.zeros((len(atoms), 3))

def get_stress(self, atoms):
    return np.zeros(6)

def calculation_required(self, atoms, quantities):
    return False

```

There is also an interface for DFT calculators, which also requires the implementation of spins, occupation numbers, the fermi level and so on.

In the previous section we used the Lennard-Jones calculator as an example, however this is generally not advised. Apart from the Lennard-Jones being a toy potential unsuited for real systems, the ASE implementation is also written in pure Python, which makes it very, very slow. Usually the calculator interface is used as a wrapper for much faster compiled code or software packages; for example the Asap calculator is a Python wrapper for the Effective Medium Theory potential which is implemented in Fortran, and is orders of magnitudes faster. A Python molecular dynamics code is justified since the bulk of computation code is spent on computing forces, however for large scale molecular dynamics simulations one would be advised to use fast compiled code such as LAMMPS<sup>4</sup> or GROMACS<sup>5</sup>. ASE really shines when it comes to small-scale simulations, experimentation and testing out new methods, and this is why it has been chosen as a basis for this thesis. Fortunately, ASE has an extensive code base of calculators for Molecular Dynamics and Density Functional Theory codes such as Asap, CP2K, LAMMPS, GROMACS which are all implemented in lower-level languages.

---

<sup>4</sup> <https://lammmps.sandia.gov/>

<sup>5</sup> <http://www.gromacs.org/>

## Chapter 8

# Atomistic Machine-learning Package

The Atomistic Machine-learning Package (AMP) [khorshidi2016amp] is a software package written in Python with the intent of bringing machine learning to electronic structure calculations. The software is intended to interface with ASE and the OpenKIM API for usage in LAMMPS. The interface to AMP is written purely in Python while computationally intensive tasks are outsourced to Fortran modules and the machine learning performed through a Tensorflow backend. The Python interface makes AMP flexible and easily extended, and this makes the package ideal for prototyping and testing both newer and more established machine learning methods.

A suggested workflow is as follows:

- Use AMP for training, testing and validation of novel descriptors and systems
- Use the AMP calculator for smaller scale simulation in the ASE environment
- Export the network compliant with the OpenKIM API for usage in more mature, large-scale electronic structure calculations such as LAMMPS

Unfortunately, the software is not currently fully compliant with the latest version of the OpenKIM API, which introduces artifacts in the simulation.

However, the authors have recently received a large grant from the U.S. Department of Energy<sup>1</sup>, which should facilitate further development.

### 8.0.1 Theory

AMP is intended to interface extensively with ASE, and the primary interface to AMP is the AMP calculator. The AMP calculator is an ASE compliant calculator that accepts cartesian coordinates and outputs energies and forces, just as a classical molecular dynamics calculator. The primary difference between the AMP calculator and ASE calculators such as the Lennard-Jones calculator is the train method, which takes a set of *images*, i.e. a set of snapshots of atomic configurations labeled with the potential energy and forces. The calculator is fitted to the images and can subsequently used to predict atomic energies and forces from never before seen atomic configurations.

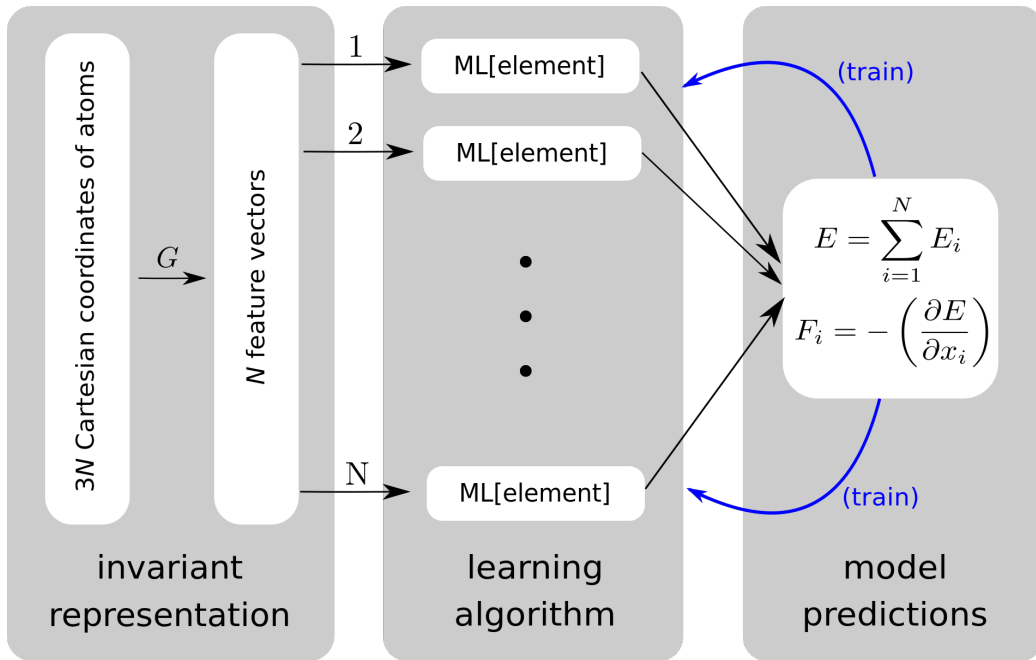


Figure 8.1: Schematics of how AMP works in atom-centered mode. Reprinted from article.

<sup>1</sup><https://www.brown.edu/news/2018-09-20/simulations>

AMP provides interfaces for both atom-centered descriptors, as we discussed in chapter 6, and image-centered descriptors, which are formed from the complete set of cartesian coordinates. A set of feature vectors are formed from an image, which are fed through a neural network to produce the potential energy. The neural network is implemented through a Tensorflow interface, which provides a full neural network graph with weights and activations. The inputs can be backpropagated through the network to produce both weight updates and derivatives with respect to the inputs.

## 8.0.2 Installation

AMP requires an installation of

- Python 2.7, 3.4-3.6, 3.6 is recommended
- Numpy 1.9 or newer
- Scipy 0.14 or newer
- ASE

Python installations can be easily obtained through the Anaconda or Miniconda packages<sup>2</sup>, or follow the instructions on the Python website<sup>3</sup>. ASE installation is described in the chapter on ASE. Once you have the prerequisites AMP can be installed using pip:

```
pip install amp-atomistics
```

## 8.0.3 Training example

In the previous chapter we showed how to run a molecular dynamics simulation using ASE. Here we will show how an AMP calculator can be fitted to molecular dynamics data with only minor modifications to the code.

First we import the prerequisites and define the system:

---

<sup>2</sup><https://anaconda.org/>

<sup>3</sup> <https://www.python.org/>



```

import ase.io
from ase.lattice.cubic import FaceCenteredCubic
from ase import units
from ase.md.velocitydistribution import
    MaxwellBoltzmannDistribution
from ase.md.verlet import VelocityVerlet

from amp import AMP
from amp.descriptor.gaussian import Gaussian
from amp.model.neuralnetwork import NeuralNetwork
from amp.model import LossFunction

symbol = "Ar"
size = (3, 3, 3)
atoms = FaceCenteredCubic(symbol=symbol, size=size, pbc=
    True)
MaxwellBoltzmannDistribution(atoms, 300 * units.kB)

```

The class AMP is the primary object of the AMP package, and is implemented through the ASE interface. We will use a neural network machine learning model and Gaussian descriptors, which are an implementation of the Behler-Parrinello method. To implement force training we require access to the LossFunction class. We will also be using the ASE Input/Output (ase.io) module to generate an ASE *Trajectory*, which is an object storing the time-evolution of a simulation and usually interpreted as a time series of atoms.

```

traj = ase.io.Trajectory("training.traj", "w")
calc = LennardJones(sigma=3.405, epsilon=1.0318e-2)
atoms.set_calculator(calc)
atoms.get_potential_energy()
atoms.get_forces()
dyn = VelocityVerlet(atoms, 5 * units.fs)
traj.write(atoms)

```

In order to write the potential energy and forces to file they must first be calculated using the ASE Atoms methods, which require a calculator and otherwise raise an error. We can then evolve the system forward in time and save the atomic configuration every 10 time steps:

```

for i in range(100):
    dyn.run(10)
    atoms.get_potential_energy()
    atoms.get_forces()
    traj.write(atoms)

```

After the data has been generated we can train the AMP calculator using the Lennard-Jones calculations as input:

```

calc = Amp(descriptor=Gaussian(),
            model=NeuralNetwork(hiddenlayers=(10, 10, 10)))
calc.model.lossfunction = LossFunction(
    convergence={"energy_rmse": 1E-2,
                 "force_rmse": 1E-2})
calc.train(images="training.traj")

```

Once the calculator has been trained, the parameters are stored in a file called "amp.amp", and can be loaded using the command:

```

calc = Amp.load("amp.amp")

```

From there the calculator can be used as any other ASE calculator, in electronic structure calculations on Atoms objects.

## 8.0.4 Descriptors

Currently AMP provides support for three different descriptors. Gaussian descriptors implement the Behler-Parrinello method of radial and angular symmetry functions. For this descriptor the derivatives are also available, which means it can be used for computing dynamics. Some physical intuition and chemical knowledge is necessary to choose various parameters for the symmetry functions - unless you wish to do it by trial and error - but AMP provides defaults for multiple chemical species.

Zernike descriptors represent a tensor product between spherical harmonics and Zernike polynomials. The local atomic environment of atom  $i$  is represented by an atomic density function  $\rho_i(\mathbf{r}_i)$ :

$$\rho_i(\mathbf{r}_i) = \sum_{j \neq i}^{\text{local}} \eta_j \delta(\mathbf{r}_i - \mathbf{r}_{ij}) f_c(\|\mathbf{r}_{ij}\|), \quad (8.1)$$

with  $f_c(r)$  a cutoff function as described in chapter 6. The 3-D Zernike basis set  $Z$  is formed by a tensor product between the Zernike polynomials basis set  $R$  and the spherical harmonics  $Y$ . The Zernike basis set is defined inside and on the surface of the  $S^2$  unit sphere as:

$$Z_{nl}^m(r, \theta, \phi) = R_n^l(r) Y_l^m(\theta, \phi), \quad (8.2)$$

where  $n \geq 0$  is an integer,  $l$  is restricted to even  $n - l \geq 0$  and  $m$  is an integer such that  $|m| \leq l$ . Functions defined inside and on the  $S^2$  sphere can be represented by the 3-D Zernike basis as:

$$f(r, \theta, \phi) = \sum_{n=0}^{\infty} \sum_l \sum_{m=-l}^l c_{nl}^m Z_{nl}^m(r, \theta, \phi), \quad (8.3)$$

with coefficients  $c_{nl}^m = \langle Z_{nl}^m, f \rangle$  computed as projections of  $f$  onto the Zernike basis. These projections form the basis of the Zernike fingerprint,  $\mathbf{G}_i$ . Centering the atomic density preserves translation invariance of the atomic environment, while the projections onto the Zernike basis set preserves rotational invariance. Permutation invariance is maintained by keeping the constant values  $\eta_j$  the same within each chemical species.

Bispectrum descriptors are computed much in the same way as the Zernike descriptors. The 4-D spherical harmonics form a complete, orthogonal basis set for the  $S^3$  4-D unit sphere, and the components of the Bispectrum fingerprints  $\mathbf{G}_i$  can be computed by projecting the atomic density function onto them. For more information the reader is encouraged to check out the AMP paper [ref here].

## Chapter 9

# Fitting to the Lennard-Jones potential

In order to demonstrate the construction of a neural network potential, we use Tensorflow to reconstruct the Lennard-Jones potential. The Lennard-Jones potential is the simplest realistic molecular dynamics potential, and since it is a function of only radial distance, symmetry functions are not required. We can thus use neural networks to perform a simple regression on a function accepting one input and producing one output. This serves to illustrate some of the intuitions and problems one runs into when using more complex methods such as atom-centered descriptors and gives a nice introduction into modern Tensorflow. We will be using the Tensorflow 2.0 beta version recently released, since it introduces a wide array of changes which will likely prove influential to the long-term direction of the Tensorflow project.

### 9.0.1 Tensorflow implementation

Text.

```
from ase.lattice.cubic import FaceCenteredCubic
from ase import units
from ase.md.velocitydistribution import
    MaxwellBoltzmannDistribution
from ase.md.verlet import VelocityVerlet

symbol = "Ar"
size = (3, 3, 3)
```

```
atoms = FaceCenteredCubic(symbol=symbol, size=size, pbc=True)
MaxwellBoltzmannDistribution(atoms, 300 * units.kB)
```

## 9.0.2 Comparison and relative error

Text.

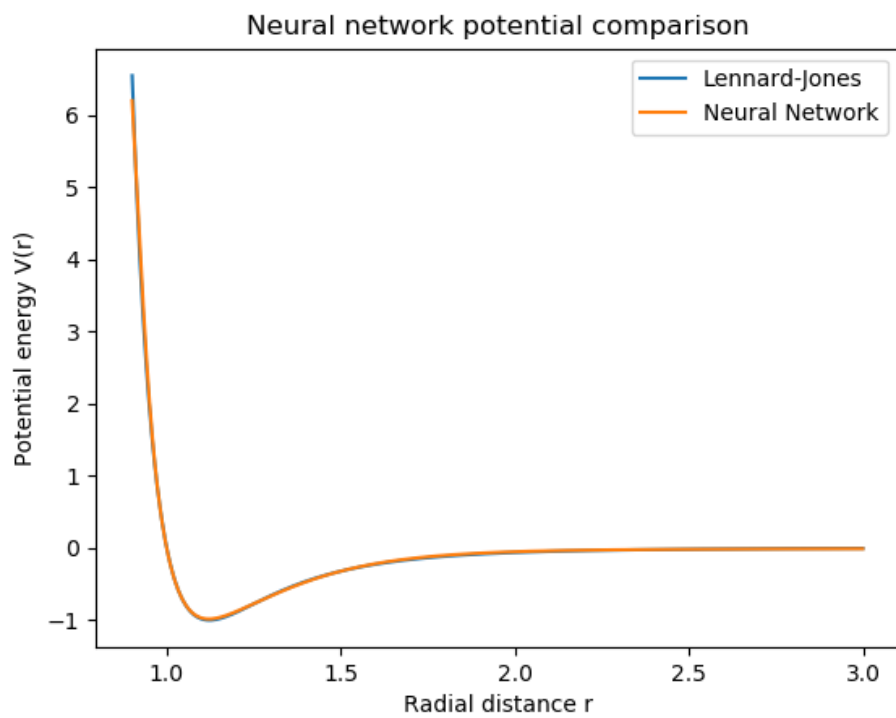


Figure 9.1: Neural network potential compared to Lennard-Jones.

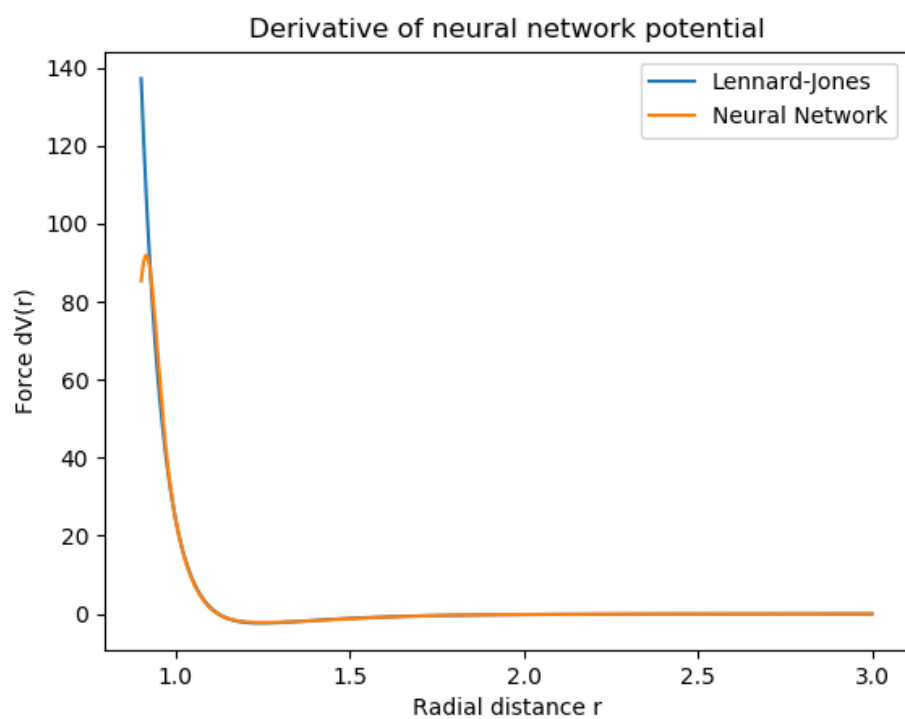


Figure 9.2: Neural network derivative compared to Lennard-Jones.

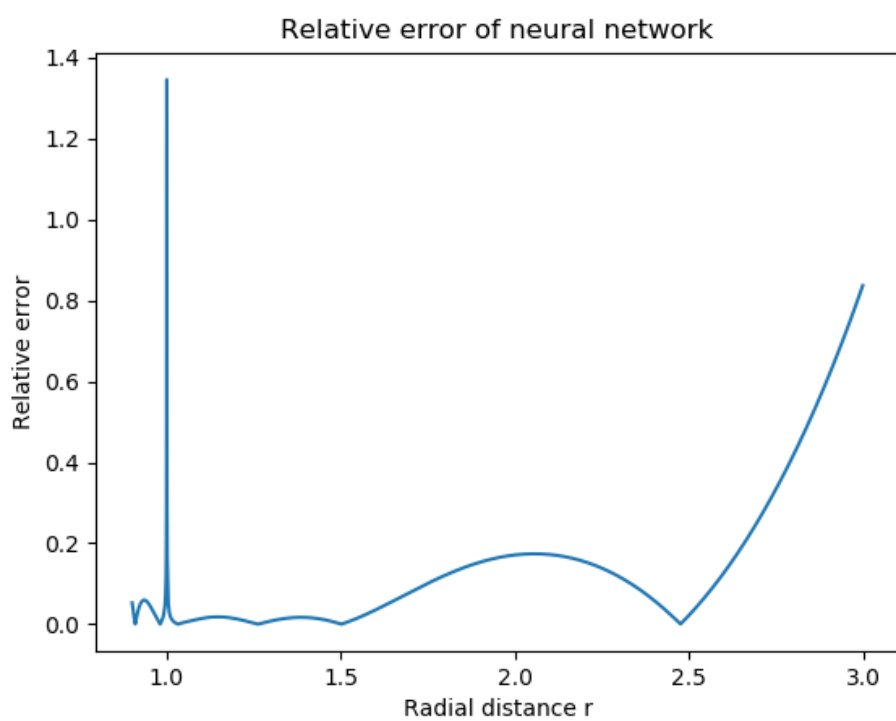


Figure 9.3: Relative error of the neural network potential compared to Lennard-Jones.

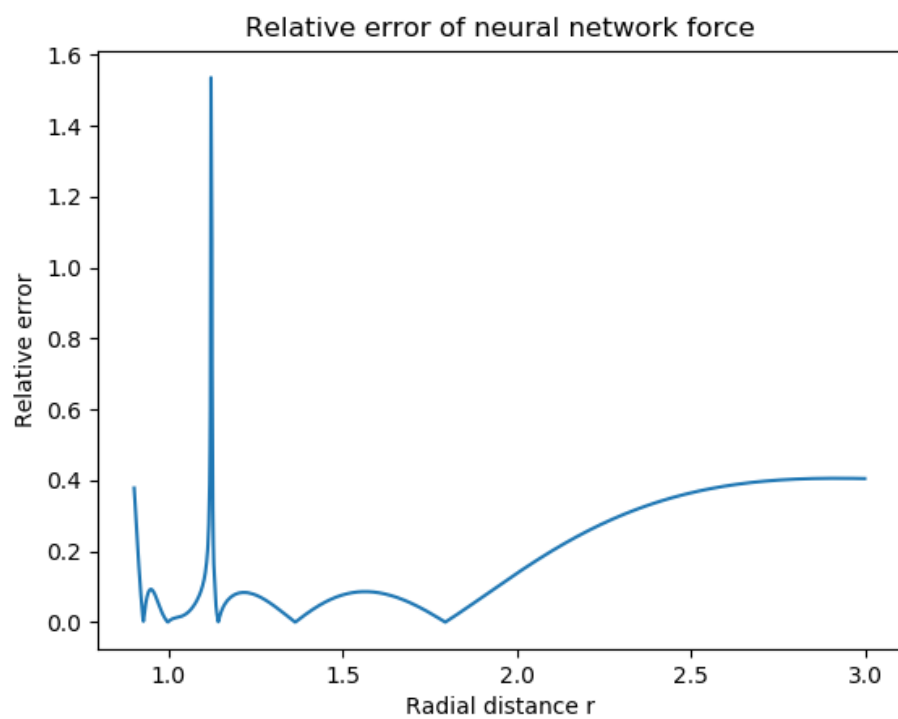


Figure 9.4: Relative error of neural network derivative compared to Lennard-Jones.



# Part III

## Results

# Chapter 10

## Empirical potentials

In order to test the validity of the Behler-Parrinello method and AMP, we train and test neural network potentials on the Lennard-Jones and Stillinger-Weber potentials. These are some of the simplest realistic potentials, which are widely applied in the molecular dynamics literature, and fitting to these potentials will provide much insight into the speed and accuracy of fitting machine-learned potentials.

### 10.0.1 Lennard-Jones

The AMP calculator is trained on a Lennard-Jones trajectory starting in a face-centered-cubic (FCC) crystal with a lattice spacing of 5.26 Å. We use  $2 \times 2 \times 2$  unit cells for a total of  $4 \cdot 2^3 = 64$  atoms, and velocities are initialized from the Maxwell-Boltzmann distribution with a temperature of 750 Kelvin. The atoms are propagated forward using the Verlet integrator in the microcanonical (NVE) ensemble for  $10^4$  steps with a timestep of  $\Delta t = 5\text{fs}$ , and the trajectory is written to file every 25 steps. The potential energy and forces is computed using the Lennard-Jones potential with the characteristic energy  $\epsilon = 1.0318 \cdot 10^{-2}\text{eV}$  and characteristic length  $\sigma = 3.405\text{\AA}$  and a radial cutoff  $r_c = 3\sigma$ .

Since the Lennard-Jones system is not a very complicated one, with interactions only reliant on distance we use the AMP defaults for the symmetry functions. The AMP defaults for Argon are 4  $G_2$  radial functions with  $\eta = 0.05, 0.23, 1.07, 5.00$  and 4  $G_4$  angular functions with  $\eta = 0.05$ ,  $\gamma = 1.0, -1.0, 1.0, -1.0$  and  $\zeta = 1.0, 1.0, 4.0, 4.0$  respectively. The smooth cutoff function is the cosine function with a cutoff  $r_c = 6.5\text{\AA}$ , which is the default

in AMP.

Plot symmetry functions?

The fitting of the neural network requires the selection of a large number of parameters in order to get an appropriate convergence. In table 10.0.1 all relevant parameters are listed. We have not performed a systematic search over hyperparameters in this section, so the choice of these is somewhat arbitrary and based on intuition and rules of thumb. The network was found to converge on an energy RMSE of approximately  $10^{-3}\text{eV}$  and a force RMSE of approximately  $7.5 \cdot 10^{-2} \text{ eV/\AA}$  after a number of trial runs, and these were therefore chosen as convergence criteria. The loss function has weightings for both the energy and the forces, and adding force terms to the loss function is thought to improve the dynamics of the trained calculator. However, this requires us to store the force on all atoms for each image, and may introduce problems with available memory.

Table 10.1: Hyperparameters used in fitting.

Hyperparameter	Value
Features	8
Hidden layers	(10, 10, 10)
Activation	Hyperbolic tangent
Timesteps	$10^4$
Max epochs	10000
Optimizer	Adam
Initial learning rate	$10^{-3}$
Energy RMSE	$10^{-3}\text{eV}$
Energy coefficient	1.0
Force RMSE	$7.5 \cdot 10^{-2}\text{eV/\AA}$
Force coefficient	0.04

In figure 10.1 you can see the convergence of the total loss, and the energy and force RMSEs. The loss has converged with minor improvements after approximately 200-300 steps, and then slowly moves toward the criteria until approximately 450 steps. The loss function is often quite noisy at the start of the training, but then converges much more smoothly towards an asymptote. The force RMSE is typically much larger than the energy RMSE, this is likely

because the force RMSE is averaged over a much larger set of components, which introduces a lot of noise into the loss.

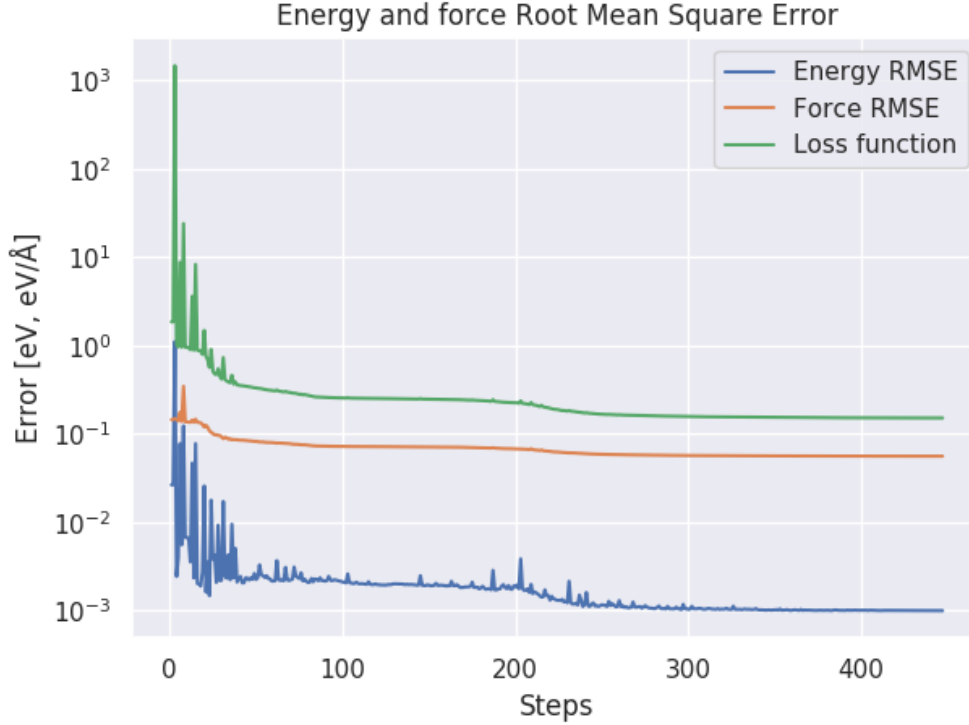


Figure 10.1: Convergence

After the AMP calculator is trained we use it to generate a test trajectory for  $10^3$  steps, which is compared to a Lennard-Jones trajectory of the same size and initial conditions. In figure 10.2 we have plotted a comparison of the potential energy of the system over time. Typical of any NVE simulation starting in an equilibrium crystal structure is that some amount of the initial kinetic energy will be converted to potential energy, and thereafter the potential energy will stabilize around some value. We see that the potential energies agree initially, but then the AMP energy diverges. This is likely a numerical artifact caused by a poor force fit allowing the atoms to move too closely, and this will increase the total energy of the system. This may be improved by further training, a larger dataset or a larger force coefficient.

In figure 10.3 we have plotted the radial distribution function, which

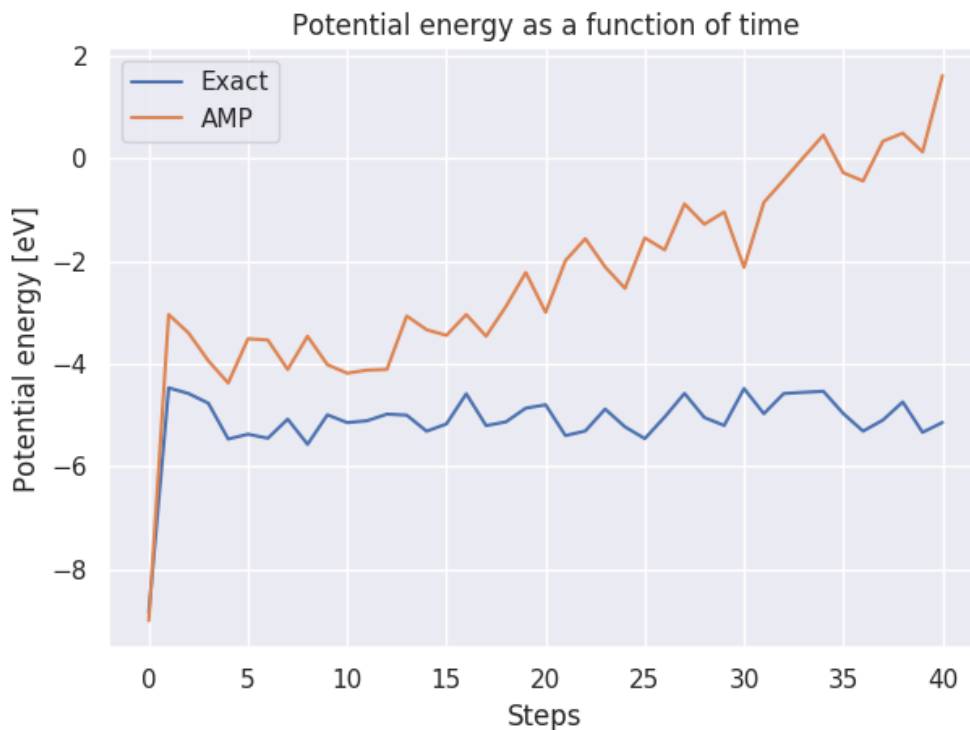


Figure 10.2: Energy difference

is a good measure of the structural composition of a system. The radial distribution function measures the density of atoms at a radial distance from an arbitrarily chosen reference atom. Any two atoms should not be too close as this implies an enormous amount of potential energy, and it typically exhibits several peaks where the atoms prefer to reside. In a homogenous infinite structure the function approaches unity as the radial distance increases. If the neural network is to reproduce the correct ensemble averages we would expect it to exhibit the same structural details. From the figure we can see the the RDFs mostly exhibit the same structure and details, which is a good sign. The RDFs do not overlap completely, which may be improved by further training and more data.

In figure 10.4 we have plotted the mean squared displacement. The mean squared displacement measures how an (average) atom moves with respect to an initial position over time, and is related to the fraction of the system

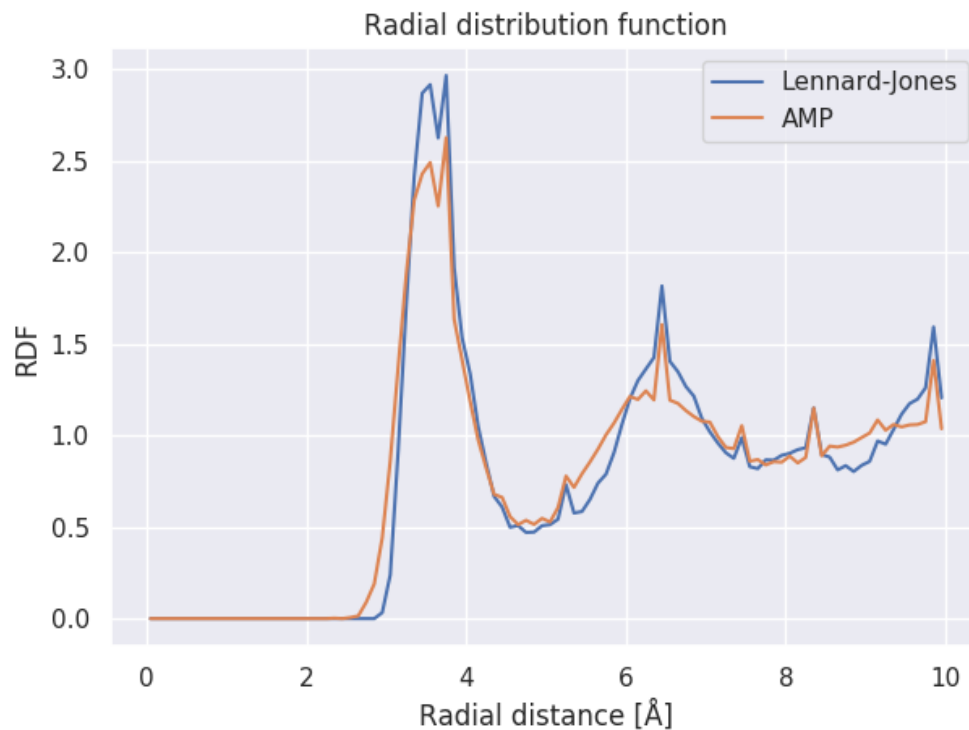


Figure 10.3: Radial distribution function

explored by a random walker. The mean squared displacement should be linear for a liquid unaffected by an external force, and can be used to compute the diffusion constant of the system. If the AMP calculator is accurate we would expect the diffusion constants to reasonably agree. We see that the AMP MSD diverges from the Lennard-Jones one and seems to be increasing sharply over time. This is likely due to numerical error, with the energy of the system increasing as the atoms are appearing too close. This should be rectified.

### 10.0.2 Stillinger-Weber

Text.

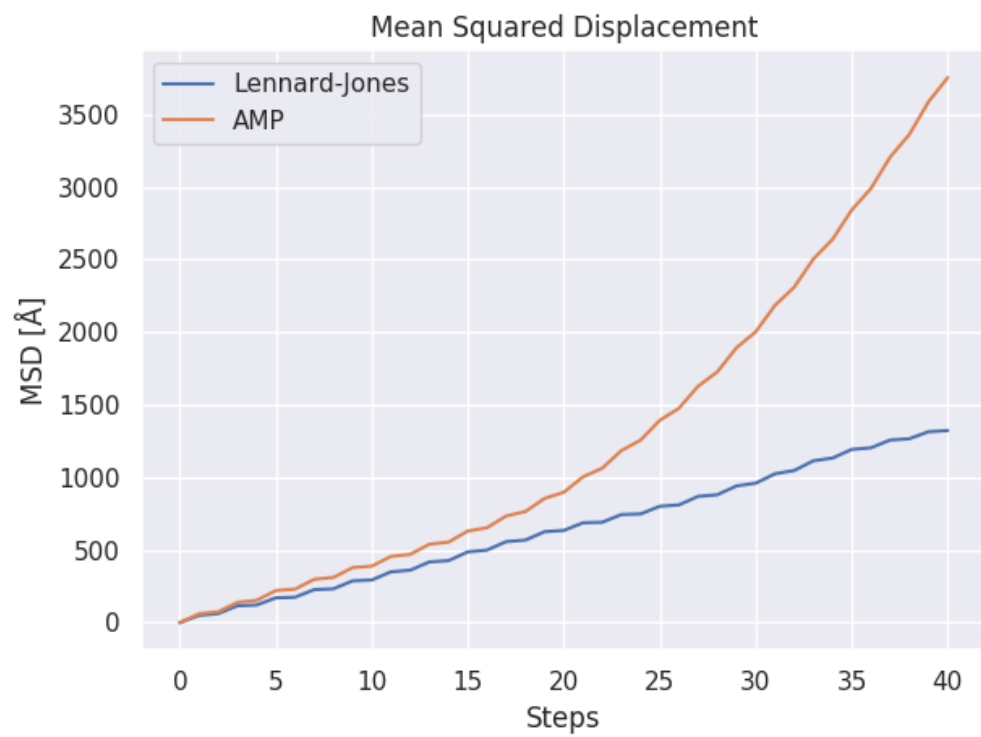


Figure 10.4: Mean Squared Displacement

# Chapter 11

## Ab-Initio Molecular Dynamics

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the



alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

# Chapter 12

## Conclusion

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

And after the second paragraph follows the third paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

After this fourth paragraph, we start a new paragraph sequence. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain

all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

This is the second paragraph. Hello, here is some text without a meaning. This text should show what a printed text will look like at this place. If you read this text, you will get no information. Really? Is there no information? Is there a difference between this text and some nonsense like “Huardest gefburn”? Kjift – not at all! A blind text like this gives you information about the selected font, how the letters are written and an impression of the look. This text should contain all letters of the alphabet and it should be written in of the original language. There is no need for special content, but the length of words should match the language.

# Appendices

# Appendix A

## Title 1

# Appendix B

## Title 2