Constructing high-dimensional neural network potentials for molecular dynamics

by

John-Anders Stende

Thesis

for the degree of

Master of Science

Faculty of Mathematics and Natural Sciences
University of Oslo

May 2017

Abstract

We implement a method for constructing analytic inter-atomic potentials by fitting artifical neural networks to data from molecular dynamics simulations. The LAMMPS distribution is extended with an algorithm for sampling atomic configurations and energies, and TensorFlow, in combination with the Behler-Parrinello method, are employed to perform the fits. Further, we add a custom potential to LAMMPS that apply trained neural networks in simulations.

We construct a neural network potential for Si, and its performance is evaluated by comparing the radial distribution function and various mechanical properties with Stillinger-Weber simulations and/or experiments (SHOULD I COMPARE WITH EXPERIMENTS?). The crystal structure of Si is reproduced with small devations, and the bulk modulus, shear modulus and Poisson ratio are measured to be similar to Stillinger-Weber values.

Acknowledgements

I would like to thank my supervisors Anders Malthe-Srenssen and Morten Hjorth-Jensen for everything they have taught me during my two years as a computational physics student. You have created a great learning environment that motivates students to *want* to work hard for learning's own sake, which is quite rare. I appreciate the numerous talks with Anders, which always left me reassured and encouraged.

Anders Hafreager has been like an additional supervisor for the whole duration of my thesis work. Thanks for always being available when I asked for guidance. Your extensive knowledge of molecular dynamics and LAMMPS have been of tremendous help.

I want to thank Morten Ledum and especially Håkon Treider for all cooperation and discussions. We have gone through a lot of ups and downs regarding symmetry functions and forces, and fixing these problems would have been difficult on my own. I would also like to thank everybody else at the computational physics group for all the enjoyable hours playing Mario Kart, along with good conversations.

A big thanks to my mom, who have been very supportive through all my years of studies, and to my sister, for all the nice dinners at Frederikke. ÅKG also deserves appreciation for being a wanted distraction during all my hours in front of the computer.

Lastly, a huge thanks to my dear Malene for all the patience and support you have granted me. I am very happy to have had you waiting for me at home after a long day at Blindern.

Contents

Contents

Chapter 1

Introduction

The field of artifical intelligence (AI) research has experienced a surge of interest in the new millenium, mainly due to the recent success of artifical neural networks (NNs). NNs are computer programs that try to mimic biological brains by giving computers the ability to learn without being explicitly programmed. They have beaten previous records in typical AI tasks like image recognition [46] and speech recognition [47], and have also shown great promise in natural language understanding [48].

In later years, ANNs have also been employed in the natural sciences, where they have outperformed other machine learning teqhniques at analyzing particle accelerator data [49] and reconstructing brain circuits [50]. Further, physicists have developed NNs that can solve the quantum many-body problem [51] and even automatically search for new quantum physics experiments [52].

1.1   Potentials in molecular dynamics

The main strength of NNs is their ability to learn features and find patterns in data sets. This has led to their application in the construction of inter-atomic potentials for molecular dynamics (MD) simulations, which is the topic of this thesis.

In MD, we are able to simulate thousands or even millions of atoms over considerable time scales (in a microscopic perspective). This is made possible by modelling the atomic interaction forces as the gradient of a potential energy surface (PES), and predicting new trajectories according to Newtonian dynamics. The PES thus determines the dynamics of the system, and is the key factor for a MD simulation to be physically plausible.

A PES can have very different forms and properties. The ideal potential[1] should be accurate, i.e. produce thermodynamic properties that agree with experiments, and be computationally efficient to enable large-scale simulations. For

---

[1]The word "potential" is frequently used in place of "potential energy surface" in this thesis.

many systems, a realistic PES must also be high-dimensional to incorporate all
relevant degrees of freedom.

There is usually a trade-off between speed and accuracy when it comes to MD
simulations. The most accurate scheme is to recalculate the PES on-the-fly with
a quantum mechanical (QM) method like Hartree-Fock or Density Functional
Theory. This *ab inito* approach to MD is however very computationally expen-
sive, and severely restricts the feasible space and time scales of the simulation
[58].

In *classical* MD, the PES is instead approximated by a *pre-defined* mathe-
matical function. The construction of such potentials has been one of the main
focus areas in MD research, and there are two main strategies. The "physical"
approach is to start out with a functional form that incorporates some of the
known physics of the system under investigation. The function has a set of pa-
rameters that are then fitted to experimental data. These *empirical* potentials
usually have a quite simple form containing only a few parameters, making them
fast to evaluate, but also biased and possibly inaccurate. Also, the parameter
fitting may be difficult and time consuming. Examples are the Lennard-Jones
[29] and the Stillinger-Weber [30] potentials.

The "mathematical" approach is to employ a QM method to produce a data
set of atomic configurations and their corresponding potential energies, called a
reference set. Then, a general functional form containing many parameters are
fitted to this data set with an interpolation algorithm of choice. The use of a
very flexible, generic functional form makes this type of potential accurate and
bias-free. If a small fitting error is achieved, only the error of the underlying QM
method is present. This enables MD simulations with effective *ab inito* accuracy
that are several orders of magnitude faster than recomputing the PES on-the-fly.
However, since no physics can be extracted from its functional form, great care
must be taken during its construction.

1.2   Neural network potentials

Several methods have been proposed to fit data sets of atomic configurations
and energies, like interpolating moving least squares [28], and modified Sheppard
interpolation [19]. Recently, also NNs have been employed for this purpose [1],
resulting in so-called *neural network potentials* (NNPs).

NNs have very flexible functional forms with an arbitrary number of inputs,
outputs and parameters. The various attempts at constructing NNPs vary sig-
nificantly in scope and complexity, especially in the number of networks that are
used. Examples of NNPs where a single NN was employed are Raff et al. [35],
who studied the dissociation of the SiO_2 molecule, and Prudente et al. [53], who
calculated the vibrational levels of the H_3^+ ion.

Using a single NN restricts the dimensionality of the system to a few atoms.

Two high-dimensional approaches have been developed in parallel by Bholoa et al. [36], and Behler and Parrinello [2]. These methods are applicable to systems containing an arbitrary number of atoms. This thesis concerns itself with the Behler-Parrinello (BP) method for constructing NNPs, described by Raff et al. [21] as the most powerful technique for executing dynamics of large systems using PESs obtained from *ab inito* calculations. The BP scheme has been applied to systems like copper [55], sodium [59], and solid-liquid interfaces [54].

As mentioned above, a NN is a purely mathematical concept, which can pose problems in a physical context. There are certain physical symmetries any PES must possess, like translational and rotational invariance of the energy of a system. Cartesian coordinates are not invariant with respect to these operations, and can therefore not be used. The BP method utilizes a set of symmetry functions that ensures the relevant symmetries are satisfied.

In principle, the method can be used to construct potentials for any chemical system, since no intial biased functional form is required. Only the symmetry function set and the reference data differ from system to system, the interpolation procedure itself is automated. Further, the NNP can be fitted to arbitrary accuracy if performed correctly. Thus, this scheme can be viewed as a multi-scale physics tool that bridges the gap between the quantum realm and the classical realm, enabling large-scale MD simulations with *ab inito* accuracy.

1.3   Goals

The main goal of this thesis is to use artificial neural networks and the Behler-Parrinello method to construct high-dimensional potential energy surfaces and apply them in molecular dynamics simulations. We will not attempt to construct a PES based on data from *ab inito* calculations, but instead perform a demonstration and validation of the complete methodology by reproducing the Stillinger-Weber potential for Si.

The main goal can be split into the following intermediate objectives:

a) **Sample data from MD simulations**
   We want to sample atomic configurations and energies from MD simulations with known, empirical potentials. The samples will be assembled to a data set, and neural networks applied to fit a functional form to this data set.

   1) **Develop a method for sampling atom-centered configurations and energies from LAMMPS**
      Specifically, we wish to dump the positions of all neighbours inside the cutoff sphere of a central atom, along with the total potential energy of the central atom. LAMMPS does not contain this functionality, a goal is therefore to develop such a method an integrate it in the LAMMPS distribution.

    2) Implement a sampling algorithm Certain statistical properties of the data set is of great importance for a NNP to perform well. Thus, we should add an adequate sampling algorithm to the LAMMPS extension above.

b) **Develop a neural network model for regression**.
We will use the machine learning software TensorFlow to implement feedforward neural networks as a set of Python scripts. The code will be organized like a library that can fit a functional form to any data set.

c) **Implement the Behler-Parrinello method**
Develop Python scripts employing the BP symmetry functions to transform the cartesian coordinates of atomic configurations to a set of symmetry values. The code should be embedded in the neural network regression model, and the combination of these scripts should be able to create NNPs for systems with an arbitrary number of chemical elements. We also wish to write code that can analyze and visualize data sets and symmetry function sets.

d) **Apply neural network potentials in molecular dynamics simulations**.
We need to extend the LAMMPS library with our own neural network potential. This includes finding a way to transfer a NN from Python to C++ and evaluate the NN as an analytic potential.

e) **Develop, apply and validate a NNP for Si**
To demonstrate the methods described in this thesis, we will fit a NN to a data set consisting of samples from Stillinger-Weber simulations of Si. The performance of the NNP will be tested by measuring the radial distribution function, the bulk modulus, the shear modulus and the Poisson ratio, and compare with Stillinger-Weber.

1.4   Our contributions

In this thesis, we have decided to employ LAMMPS to perform MD simulations and TensorFlow to train neural networks. The author has previously written a basic molecular dynamics code[2], which has been employed to do some preliminary testing on how to evaluate and differentiate a neural network in C++, that has been trained in Python. However, LAMMPS has been used in the final implementation of the neural network potential for speed purposes.

    We realized from the beginning that writing our own neural network code would not be worth the effort. We wanted to focus on the whole process of

---

[2]github.com/johnands/FYS4460-MD

constructing a NNP, not just the fitting stage. Developing code from scratch would have given deep insights in the underlying concepts of neural networks, but this has nonetheless been obtained by careful study of the processes initiated with each TensorFlow command.
In this thesis we have:

- Extended the LAMMPS library with a custom `compute` that samples atom-centered configurations and potential energies. This also incluces a modified version of a sampling algorithm.

- Developed a method to transfer a trained NN from Python to C++ by manually saving all relevant parameters to file. This file is then read in C++ and the NN reconstructed as a set of matrices and vectors.

- Extended the LAMMPS library with a custom class to enable the use of NNPs in MD simulations. This has been implemented both manually and with the TensorFlow C++ API[3]. To differentiate the network and obtain forces, we have implemented a modified version of the backpropagation algorithm.

- Written a set of post-processing and analysis scripts.

All the source code can be found at GitHub[4].

1.5   Structure of the thesis

This thesis has four parts: The first part consists of an introduction to the theories of molecular dynamics and neural networks, in addition to a description of the Behler-Parrinello method. The second part provides introductions to the LAMMPS and TensorFlow libraries with implementational details, advanced theory on the construction of NNPs, and results demonstrating the complete method. The fourth part concludes the text and contains a discussion of possible future work and applications.

---

[3]Application Programming Interface.
[4]github.com/johnands (pinned repositories).

Part I

Theory

Chapter 2

Molecular dynamics

Molecular dynamics (MD) is a method to simulate the physical movements of atoms and molecules in gases, liquids and solids. It is thus a type of N-body simulation. The atoms are modelled as point-like particles with interactions described by force fields. Their time-evolution is governed by Newton's equations of motion. MD allows us to study the microscopic movement of thousands or millions of atoms, enabling the sampling of macroscopic properties such at temperature, pressure, diffusion, heat capacity and so on.

The dynamics of an ensemble of particles is governed by their interactions. In MD, the interactions are described by a force field $\mathbf{F}$, which is defined as the negative gradient of a potential energy surface (PES) $V(\mathbf{r})$,

$$\mathbf{F} = -\nabla V(\mathbf{r}) \tag{2.1}$$

The PES describes the potential energy of a system for any configuration of atoms $\mathbf{r} = (\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N)$, and thereby the forces, thus controlling the behaviour of the system. In other words, the PES is clearly the most important factor for a MD simulation to be physically plausible.

In this chapter, we will start out with a discussion on different ways to construct an adequate PES, followed by details about time-integration and functions that are commonly used to define a PES.

2.1   Potential energy surfaces

We know that atomic interactions are governed by the laws of quantum mechanics (QM). In principle, we can solve the time-dependent Schrödinger equation to obtain the trajectories of an ensemble of atoms. This is however a very computationally expensive task, and is not feasible except for small systems and time scales. Numerical methods for this purpose scales non-linearly with system size, while MD simulations scale linearly, if implemented in a certain way.

An alternative is to solve the time-*independent* Schrödinger equation for each time step in a simulation. The resulting PES is differentiated to obtain the forces on all atoms, and the new positions and velocities are calculated by employing Newton's laws. This method is called *ab inito* molecular dynamics. Although this is the most accurate approach to MD, it also suffers from bad scalability when the system size grows.

In *classical* MD, the PES is represented by a predefined, mathematical function, i.e. we do not have to update the PES during the simulation, resulting in better performance. The cost is a loss of accuracy, but for many systems it is fully possible to construct a predefined PES that yields the correct thermodynamic properties. In this thesis, we present a method to construct potential energy surfaces for classical MD.

2.1.1   From quantum mechanics to classical potentials

It it not obvious how an ensemble of atoms, which is a quantum mechanical system, can be simulated using laws from classical mechanics. To be able to do this, we must make the assumption that the motion of atomic nuclei and electrons in a molecule can be seperated, known as the Born-Oppenheimer approximation. Under this assumption, we can solve the electronic Schrödinger equation with the nucleonic degrees of freedom as parameters. The resulting wavefunction serves as a potential in a *nucleonic* Schrödinger equation, yielding a PES for the nuclear motion. In this way, the positions and velocities of the electrons are baked into the PES, allowing us to model the atomic nuclei as point particles that follow Newtonian dynamics.

2.1.2   Constructing potential energy surfaces

As mentioned above, we represent the PES by a predefined functional form in classical MD. We assume that this function can be constructed as a sum of n-body terms V_n,

$$V(\mathbf{r}) \approx \sum_i^N V_1(\mathbf{r}_i) + \sum_{i,j}^N V_2(\mathbf{r}_i, \mathbf{r}_j) + \sum_{i,j,k}^N V_3(\mathbf{r}_i, \mathbf{r}_j, \mathbf{r}_k) + \ldots \qquad (2.2)$$

where the sums are over all single atoms i, all pairs (i,j), all triplets (i,j,k) and so on. Each term is determined by empirical knowledge of the system, by quantum mechanical calculations or a combination of both. The resulting PES is called an empirical, quantum mechanical or semi-empirical potential respectively.

When constructing empirical potentials, we typically start out with a functional form that captures some of the known, basic properties of the system under investigation. The function has a set of parameters that are fitted to experimental data, hence the name. Empirical potentials are often quite simple and only

have a few parameters, making their computatioal performance superior. However, their accuracy is limited by the choice of functional form; there is always a risk that not all the physics of a system is captured. Further, the parameters need to be fitted specifically for a given system and are typically not transferable from one system to another. Also, the fitting process can be difficult and time-consuming.

For QM potentials, electronic structure calculations are performed for a number of atomic configurations, yielding a data set of configurations and corresponding potential energies. A very flexible functional form, often containing many parameters, are then fitted to this data set. Potentials of this type can in principle be fitted to arbitrary accuracy, and may thus be treated as *ab inito* potentials. QM potentials are purely mathematical, in the sense that no physics can be extracted from their functional form. Therefore, the construction of the potential has to be done with great care, i.e. the data set must contain all relevant configurations for the intended application of the PES. However, the unbiased nature of QM potentials make them more general than their empirical counterparts, and the methodology can be applied to any system.

Semi-empirical potentials are constructed as a combination of the two schemes above. Their general functional form are based on methods from computational QM, but many approximations are made and some parameters are obtained from empirical data.

Truncation and configuration space

Determining how many terms in (2.2) that is sufficient to describe a system adequately is not trivial. Noble gases like argon (Ar) have weak interactions and may be well described by a two-body potential V_2, while molecules with strong angular dependencies on their bond, like silicon (Si), will need more terms. We note that the one-body term V_1 is only included if the system is experiencing an external force.

Another challenge is to identify the configuration space of the system. For three-body potentials and more the set of probable configurations is large, and running electronic structure calculations for each configuration is expensive. The problem of identifying the configuration space is described in section 7.1.

Fitting procedure

In this thesis we present a method to construct many-body QM potentials for MD simulations. An important step in this process is the fitting of the data set of atomic configurations and energies by a suitable functional form. Several methods have been proposed, like different types of splines, interpolating moving least squares [28] and modified Sheppard interpolation [19]. We will use *artificial neural networks* (NNs) for this purpose. The basic theory of NNs is presented

in chapter 2, while chapter 3 deales with physical symmetries that has to be considered when constructing neural network potentials (NNPs). Finally, in the implementation part, the details of constructing and applying a NNP in MD simulations are described.

2.2   Common empirical potentials

Although the method described in this thesis is for the construction of QM potentials, empirical potentials are useful as test cases and validations of the method. We can define two types of empirical potentials based on how they relate to groups of atoms. *Bonded* potentials compute the energy of each (covalent) bond for a predefined set of atoms or molecules, while *non-bonded* potentials depend on the spatial configuration of *each possible set* of particles of a specific kind, e.g. all pairs of Ar atoms or all triplets of Si atoms. The total potential V_{tot} is the sum of the contributions from these two types of potentials,

$$V_{\text{tot}} = V_{\text{bonded}} + V_{\text{non-bonded}} \tag{2.3}$$

We will only work with non-bonded potentials in this thesis. In the following we will discuss three non-bonded potentials that are common in the literature, two of which we have employed to validate our PES construction method.

2.2.1   Lennard-Jones

The Lennard-Jones (LJ) potential [29] is one of the simplest and most widely used potentials. It has a simple functional form which is inexpensive to compute, and is able to reproduce weakly interacting systems like noble gases well. It is a two-body potential that only depends on the distance r_{ij} between pairs of atoms i and j. The general form is (Figure 2.1)

$$V_{\text{LJ}}(r_{ij}) = 4\epsilon \left[\left(\frac{\sigma}{r_{ij}} \right)^{12} - \left(\frac{\sigma}{r_{ij}} \right)^{6} \right] \tag{2.4}$$

where the two parameters ϵ and σ are the depth of the potential well and the zero-point of the potential respectively. The Lennard-Jones potential captures a basic, phyiscal property of many atomic systems: Atoms repulse each other at short distances and are attracted to each other at long distances. The first term describes the Pauli repulsion due to overlapping electron orbitals, while the second term is the attraction due to the so-called van der Waals force. The functional form of the repulsive term has, in contrast to the attractive one, no theoretical justification other than that it approximates the Pauli repulsion well.
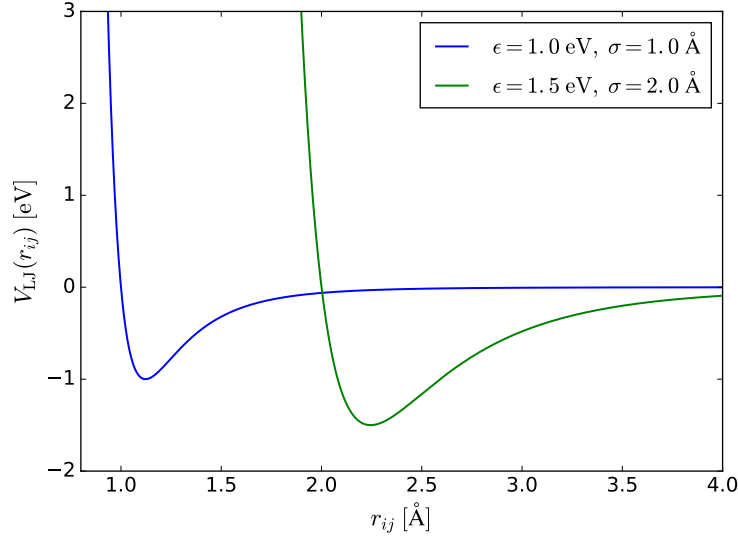
Figure 2.1: The Lennard-Jones potential (2.4) as a function of inter-atomic distance for two different parameter sets. The depth of the potential well (potential strength) is given by ϵ, while σ is the inter-atomic distance for which the potential is zero.

Calculating total potential energy

Before we move on to three-body potentials, we will briefly explain how the total potential energy of a system is calculated. For a two-body potential like LJ, each unique pair of atoms is only counted once, so that

$$V_{\text{tot}} = \sum_{i<j} V_2(r_{ij}) = \frac{1}{2} \sum_{i \neq j} V_2(r_{ij}) \qquad (2.5)$$

In the second expression we count all pairs twice, thus a factor $1/2$ is required. For the three-body case, the counting is a bit more complicated. An atomic triplet is defined by the three variables $(r_{ij}, r_{ik}, \theta_{jik})$, where r_{ij} and r_{ik} are the two inter-atomic distances to the central atom i, and θ_{jik} is the angle between the lines from i to j and from i to k. Note that there are only two *independent* variables, since

$$\theta_{jik} = \frac{\mathbf{r}_{ij} \cdot \mathbf{r}_{ik}}{r_{ij}^2 r_{ik}^2} \qquad (2.6)$$

Figure 2.2 illustrates that three atoms (i, j, k) define three different triplets depending on which atom that is the central one. All three triplets have different energies, and must therefore be tallied individually. If we assume that the potential form is symmetric with respect to exchange of the j and k indicies (which is
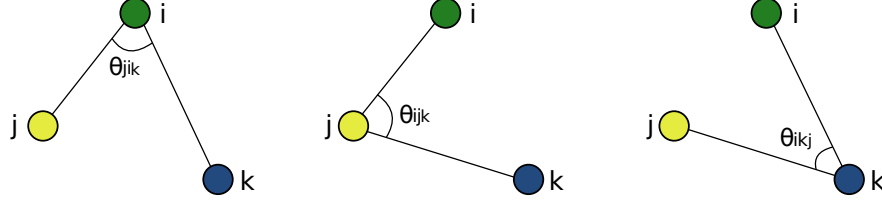
Figure 2.2: Three atoms (i, j, k) define three different triplets that generally have different energies. They must therefore be tallied individually when computing the total energy of a system.

not always the case for multielemental systems), we have

$$\sum_{i<j<k} \left[V_3^{(1)}(r_{ij}, r_{ik}, \theta_{jik}) + V_3^{(2)}(r_{ij}, r_{jk}, \theta_{ijk}) + V_3^{(3)}(r_{ik}, r_{jk}, \theta_{ikj}) \right] \tag{2.7}$$

where $V_3^{(i)}$ is one of the triplets in Figure 2.2. Alternatively, we can calculate the energy of one triplet at a time and instead sum over all j,

$$\sum_i \sum_{j \neq i} \sum_{k>j} V_3(r_{ij}, r_{ik}, \theta_{jik}) \tag{2.8}$$

This is the convention we are going to adopt in the following. The total energy is now written,

$$V_{\text{tot}} = \sum_i \sum_{j>i} V_2(r_{ij}) + \sum_i \sum_{j \neq i} \sum_{k>j} V_3(r_{ij}, r_{ik}, \theta_{jik}) \tag{2.9}$$

2.2.2   Stillinger-Weber

The Stillinger-Weber (SW) potential was proposed in 1985 [30] for simulating Si atoms. It has a two-body and a three-body term, and is thus written on the form (2.9). The two-body part consists of a Lennard-Jones factor with variable exponents, together with an exponential decay factor that ensures zeroing out of the potential at a *cutoff* distance $r_c = a\sigma$,

$$V_2(r_{ij}) = A\epsilon \left[B \left(\frac{\sigma}{r_{ij}} \right)^p - \left(\frac{\sigma}{r_{ij}} \right)^q \right] \exp \left(\frac{\sigma}{r_{ij} - a\sigma} \right) \tag{2.10}$$

The significance of the cutoff r_c is explained in section 2.4. Silicon forms 4-coordinated tetrahedral bonded structures, and the purpose of the three-body component is to enforce the tetrahedral bond angle $\theta_0 \approx 109°$,

$$V_3(r_{ij}, r_{ik}, \theta_{jik}) = \lambda\epsilon[\cos\theta_{jik} - \cos\theta_0]^2 \exp \left(\frac{\gamma\sigma}{r_{ij} - a\sigma} \right) \exp \left(\frac{\gamma\sigma}{r_{ik} - a\sigma} \right) \tag{2.11}$$

We observe that for $\theta_{jik} = \theta_0$, the three-body term is zeroed out, making this configuration energetically favourable. The expression also contains two exponential decay factors to ensure that the potential goes to zero once r_{ij} and/or r_{ik} approach the cutoff.

SW simulations of solid and liquid silicon agree reasonably well with experimental data [30]. The potential has also been used on other systems than silicon, for instance in monatomic water models [31].

2.2.3   Vashishta

In 1990, Vashishta et al. [32] suggested a three-body interaction potential for SiO_2. The three-body term is of similar form to that of Stillinger-Weber,

$$
\begin{aligned}
V_3(r_{ij}, r_{ik}, \theta_{jik}) \ = B_{jik} & \frac{[\cos\theta_{jik} - \cos\theta_{0jik}]^2}{1 + C[\cos\theta_{jik} - \cos\theta_{0jik}]^2} \cdot \\
& \exp\left(\frac{\gamma_{ij}}{r_{ij} - r_{0ij}}\right) \exp\left(\frac{\gamma_{ik}}{r_{ik} - r_{0ik}}\right)
\end{aligned}
\tag{2.12}
$$

We note that indicies are included for the parameters, as they are generally different for each pair and triplet combination of elements; the parameter set must be fitted to e.g. (Si,Si,Si) and (Si,O,O) seperately. The angle θ_{0jik} has the same purpose as for SW, while r_{0ij} is the cutoff distance for each pair of elements. The two-body part is more complex,

$$
V_2(r_{ij}) = \frac{H_{ij}}{r^{\eta_{ij}}} - \frac{W_{ij}}{r^6} + \frac{Z_i Z_j}{r} \exp(-r/\lambda_{1,ij}) - \frac{D_{ij}}{r^4} \exp(-r/\lambda_{4,ij})
\tag{2.13}
$$

where we recognize the two first terms as typical LJ repulsive and attractive components. The other two describe electrostatic interactions, the third being a so-called screened Coulomb potential for charges Z_i and Z_j. We will not expand upon the physical justification of these terms here.

The Vashishta potential has also been succesfully employed on other multi-component systems like SiC [33].

2.3   Time integration

To evolve the system forward in time, we need a numerical method to integrate Newton's equations of motion,

$$
\mathbf{v}(t) = \mathbf{v}(0) + \int_0^t \mathbf{a}(t)dt
\tag{2.14}
$$

$$
\mathbf{r}(t) = \mathbf{r}(0) + \int_0^t \mathbf{v}(t)dt
\tag{2.15}
$$

Various finite different schemes can be used for this purpose, and the choice of algorithm is of vital importance to the quality of the simulation. The most important property of an integrator for use in MD, is long-term energy conservation. MD simulations are often run for millions of time steps, and a significant energy drift may cause useless results. Further, all integration schemes approximate derivatives as finite differences, introducing truncation errors. Finally, numerical stability is important to prevent unphysical behaviour on large time scales.

A good choice is the Velocity-Verlet algorithm, which is a *symplectic* integrator. Put simply, this means that it preserves the form of Hamilton's equations, thereby conserving energy[1]. Employing Velocity-Verlet allows us to sample the phase space of the microcanonical (NVE) ensemble, and collect thermodynamic properties like temperature and pressure. The steps of the algorithm are,

$$\mathbf{v}(t + \Delta t/2) = \mathbf{v}(t) + \frac{\mathbf{F}(t)}{m}\frac{\Delta t}{2} \tag{2.16}$$

$$\mathbf{r}(t + \Delta t) = \mathbf{r}(t) + \mathbf{v}(t + \Delta t/2)\Delta t \tag{2.17}$$

$$\mathbf{v}(t + \Delta t) = \mathbf{v}(t + \Delta t) + \frac{\mathbf{F}(t + \Delta t)}{m}\frac{\Delta t}{2} \tag{2.18}$$

First, the velocity at half the time step forward $\mathbf{v}(t + \Delta t/2)$ is computed. This half-velocity is used to obtain the new positions $\mathbf{r}(t + \Delta t)$. Then we calculate the new accelerations $\mathbf{F}(t + \Delta t)/m$ based on the new positions, which is used to compute the velocities at a whole time step forward $\mathbf{v}(t + \Delta t)$.

In addition to being symplectic, Velocity-Verlet is time-reversible [34], which it should be considering that we are solving time-reversible equations (2.14, 2.15). It is also computationally inexpensive; the only additional calculation compared to the primitive Euler method is the half-velocity. The global error in position and velocity is $O(\Delta t^2)$, which may not seem very impressive. However, this is sufficient for collecting statistical properties, our goal is not to reproduce the exact trajectories of the particles.

Velocity-Verlet is the integrator of choice in LAMMPS (chapter 5), and thus the only scheme used in this work.

2.4   Force calculations and cutoff radius

We need the force on each atom to compute the acceleration that is integrated in the equations of motion (2.14, 2.15). Using the Lennard-Jones potential as an example, the force on atom i from atom j is (2.1, 2.4),

$$F_{\mathrm{LJ}}(r_{ij}) = -\nabla V_{\mathrm{LJ}}(r_{ij}) = -24\epsilon \left[2\left(\frac{\sigma^{12}}{r_{ij}^{13}}\right) - \left(\frac{\sigma^6}{r_{ij}^7}\right) \right] \tag{2.19}$$

---

[1]Strictly speaking, Velocity-Verlet only conserves a pseudo-Hamiltonian approaching the true Hamiltonian in the limit of infinitesimal time steps [34].

The force is of the same shape as the potential in Figure 2.1, which means it rapidly goes to zero. If we choose units so that $\sigma = \epsilon = 1.0$, the force at a distance $r_{ij} = 3\sigma$ is

$$F_{\mathrm{LJ}}(r_{ij} = 3.0\sigma) \approx 0.01 \tag{2.20}$$

This is a general property of systems where only van der Waal-interactions are present: Atoms that are far apart do not effectively interact. This can be exploited in MD simulations by introducing a cutoff radius r_{c}, and only compute forces for atoms displaced by a distance smaller than this cutoff. In principle, we have to sum over all pairs of atoms in the system to compute the energy and forces, which scales like $O(N^2)$. With a cutoff, this is reduced to $O(N)$, and the force is now written,

$$F_{\mathrm{LJ}}(r_{ij}) = \begin{cases} -24\epsilon \left[2 \left(\frac{\sigma^{12}}{r_{ij}^{13}} \right) - \left(\frac{\sigma^6}{r_{ij}^7} \right) \right], & r_{ij} \leq r_{\mathrm{c}} \\ 0, & r_{ij} > r_{\mathrm{c}} \end{cases} \tag{2.21}$$

which greatly increases the computational performance of MD simulations. In fact, simulations with non-bonded potentials not including a cutoff is unfeasible except for small systems.

The potential itself (2.4) is also truncated at $r_{ij} = r_{\mathrm{c}}$, creating a discontinuity at the cutoff. To avoid this problem, we simply shift the LJ potential upward a little, so that the energy is excactly zero at this point. The truncated and shifted form of the potential is

$$V_{\mathrm{LJ}}^{\mathrm{trunc}}(r_{ij}) = \begin{cases} V_{\mathrm{LJ}}(r_{ij}) - V_{\mathrm{LJ}}(r_c), & r_{ij} \leq r_{\mathrm{c}} \\ 0, & r_{ij} > r_{\mathrm{c}} \end{cases} \tag{2.22}$$

Chapter 3

Machine learning

Machine learning is the science of giving computers the ability to learn without being explicitly programmed. The idea is that there exist generic algorithms which can be used to find patterns in a broad class of data sets without having to write code specifically for each problem. The algorithm will build its own logic based on the data.

Machine learning is a subfield of computer science, and is closely related to computational statistics. It evolved from the study of pattern recognition in artificial intelligence (AI) research, and has made contributions to AI tasks like computer vision [46], natural language processing [48] and speech recognition [47]. It has also, especially in later years, found applications in a wide variety of other areas, including bioinformatics, economy, physics, finance and marketing [7].

The approaches to machine learning are many, but are often split into two main categories. In *supervised learning* we know the answer to a problem, and let the computer deduce the logic behind it. On the other hand, *unsupervised learning* is a method for finding patterns and relationship in data sets withouth any prior knowledge of the system. Some authours also operate with a third category, namely *reinforcement learning*. This is a paradigm of learning inspired by behavioural psychology, where learning is achieved by trial-and-error, solely from rewards and punishment.

Another way to categorize machine learning tasks is to consider the desired output of a system. Some of the most common tasks are:

- Classification: Outputs are divided into two or more classes. The goal is to produce a model that assigns inputs into one of these classes. An example is to identify digits based on pictures of hand-written ones. Classification is typically supervised learning.

- Regression: Finding a functional relationship between an input data set and a reference data set. The goal is to construct a function that maps input data to continous output values.
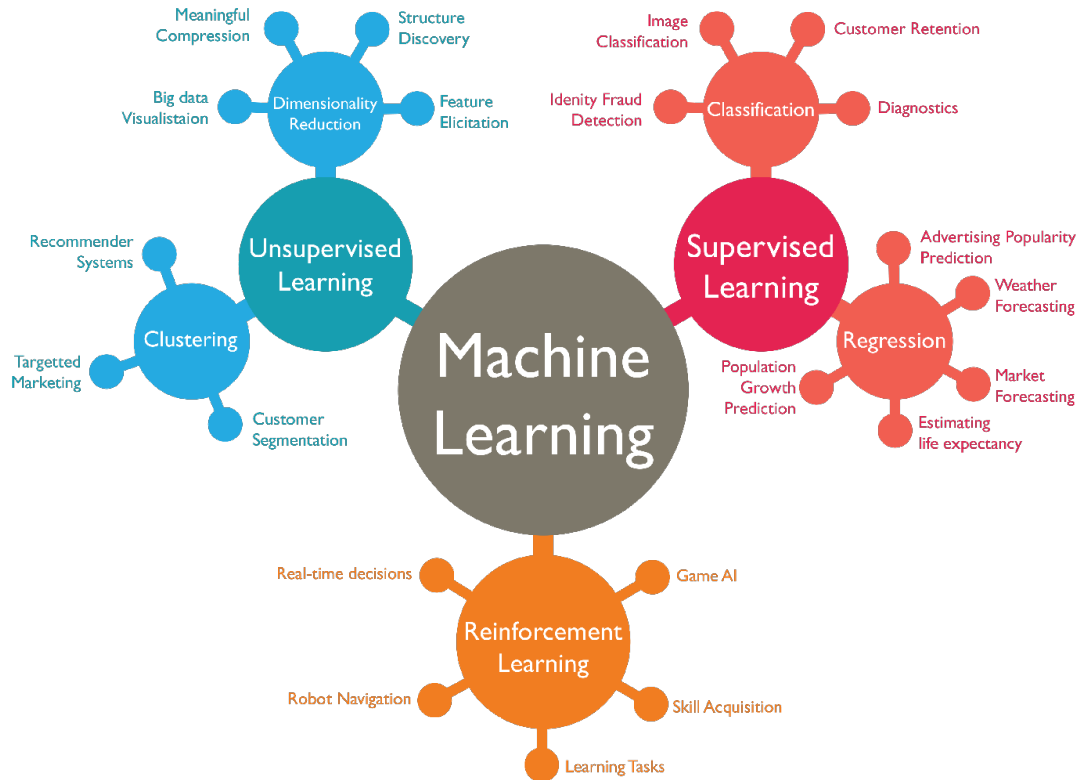
Figure 3.1: Overview of various approaches to machine learning and their applications. Source: isaziconsulting.co.za.

- Clustering: Data are divided into groups with certain common traits, without knowing the different groups beforehand. It is thus a form of unsupervised learning.

Figure 3.1 gives an (incomplete) overview of machine learning methods with some selected applications. We will not go into further detail on all these approaches to machine learning, as this work is only concerned with one specific area: artificial neural networks. The rest of this chapter is devoted to the ideas behind this machine learning approach.

3.1   Artificial neurons

The field of artificial neural networks has a long history of development, and is closely connected with the advancement of computer science and computers in general. A model of artificial neurons was first developed by McCulloch and Pitts in 1943 [23] to study signal processing in the brain and has later been refined by others. The general idea is to mimic neural networks in the human brain, which is composed of billions of neurons that communicate with each other by
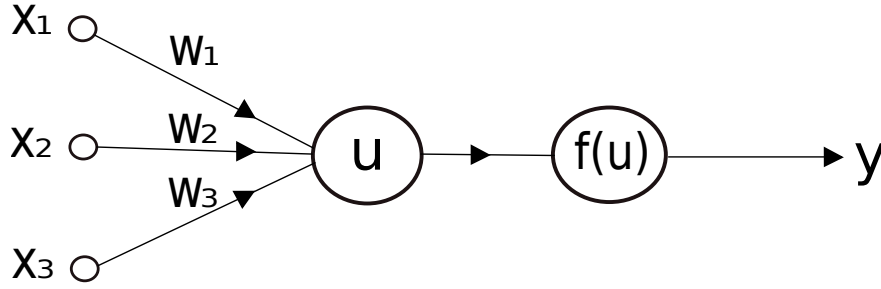
Figure 3.2: Mathematical model of an artificial neuron. The neuron receives input signals $x_i, \ldots, x_n$ from n other neurons. Each signal x_i is associated with a weight w_i, and the neuron accumulates all input signals as a weighted sum u. This sum is then used as input to its activation function f, which serves as the neuron's output signal.

sending electrical signals. Each neuron accumulates its incoming signals, which must exceed an activation threshold to yield an output. If the threshold is not overcome, the neuron remains inactive, i.e. has zero output.

This behaviour has inspired a simple mathematical model for an artificial neuron [5],

$$y = f\left(\sum_{i=1}^{n} w_i x_i\right) = f(u) \tag{3.1}$$

Here, the output y of the neuron is the value of its activation function, which have as input a weighted sum of signals $x_i, \ldots, x_n$ received by n other neurons. The model is illustrated in Figure 3.2. We will now take a closer look at if and how this mathematical model is biologically plausible.

Biological model

Artificial neurons are designed to mimic certain aspects of their biological counterparts [24]. A schematic model of a biological nervous cell is depicted in Figure 3.3. The *dendrites* in a neuron acts as the input vector, and allow the cell to receive signals from a large number of neighbouring neurons. As in the mathematical model, each dendrite is associated with the multiplication of a "weight value", which is achieved by an increase or decrease of chemical neurotransmitters that amplifies or weakens the input signals. The cell can also transmit signal inhibitors (oppositely charged ions) to accomplish "negative" weight values.

The positive and negative ions arriving from the dendrites are mixed together in the solution inside the *soma*, which corresponds to the summation in (3.1). Finally, the *axon* serves as the activation function by sampling the electrical potential of this solution. If the potential reaches a certain strength, the axon transmits a signal pulse down its own length. The axon is connected to another
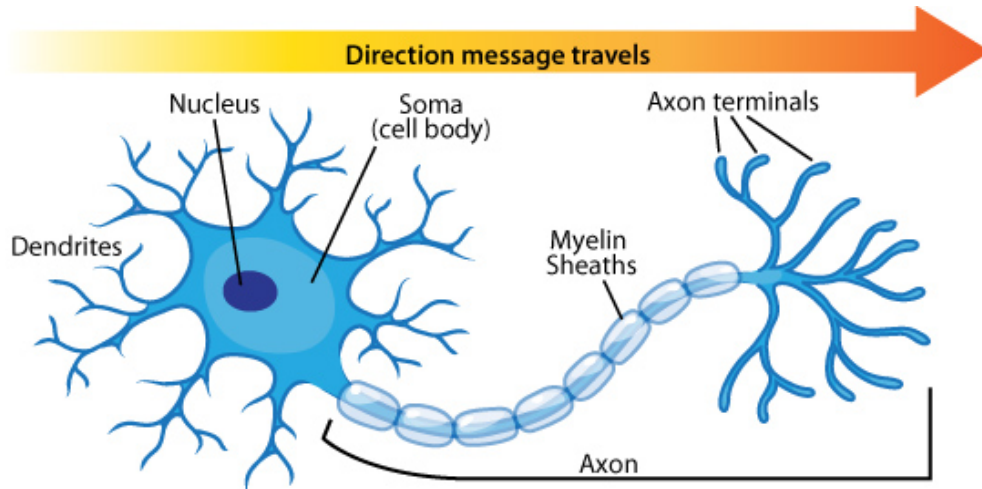
Figure 3.3: Biological model of a nervous cell. The various parts of the cell are briefly explained in the text. Source: askabiologist.asu.edu.

neuron's dendrites (or other parts of the cell), enabling the neurons to communicate.

This discussion of nervous cells helps, to a certain degree, to validate the mathematical form (3.1) of artifical neurons. However, we note that the above description is a highly simplified picture of the great complexity of biological neurons, which is not yet fully understood. Our mathematical neurons are thus only caricatures of nature. Next, we will look at how networks of these mathematical objects can be formed to emulate the neural networks in the brain.

3.2   Neural network types

An artificial neural network (ANN), is a computational model that consists of layers of connected neurons, or *nodes*. It is supposed to mimic a biological nervous system by letting each neuron interact with other neurons by sending signals in the form of mathematical functions between layers. Each node is modelled according to (3.1). A wide variety of different ANNs have been developed, but most of them consist of an input layer, an output layer and eventual layers in-between, called *hidden layers*. All layers can contain an arbitrary number of nodes, and each connection between two nodes is associated with a weight variable.

The main factor that seperates the different types are how the neurons are *connected*. This section contains a short presentation of some of the most common types of ANNs, before we move on to a more detailed description of the ANN architecture used in this thesis. In the following we will refer to artificial neural networks simply as neural networks (NNs), i.e. not referring to its biological
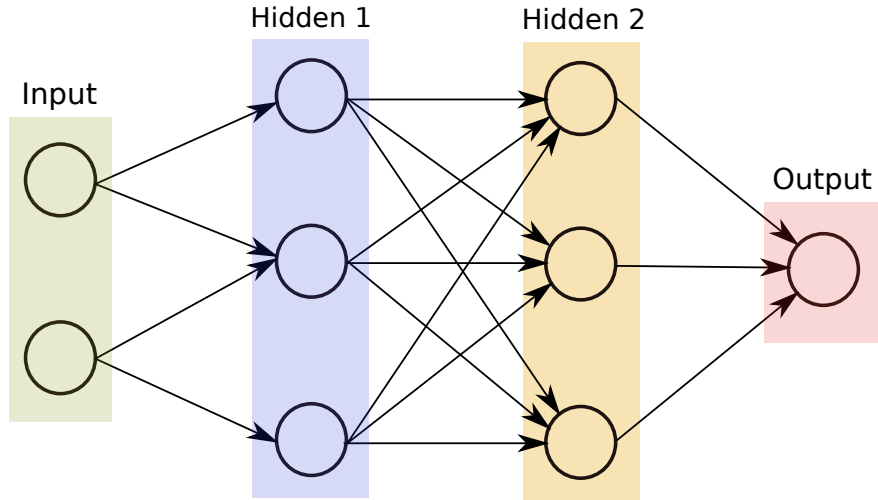
Figure 3.4: Example of a fully-connected feed-forward neural network with an input layer, two hidden layers and an output layer. Each node in a layer is connected to *all* nodes in the subsequent layer, and information only flows forward through the layers, hence the name.

counterpart.

3.2.1   Feed-forward neural networks

The feed-forward neural network (FFNN) was the first and simplest type of NN devised. In this network, the information moves in only one direction: forward through the layers. An example FFNN is shown in Figure 3.4, consisting of an input layer, two hidden layers and an output layer. Nodes are represented by circles, while the arrows display the connections between the nodes, including the direction of information flow. Additionaly, each arrow corresponds to a weight variable, not displayed here. We observe that each node in a layer is connected to *all* nodes in the subsequent layer, making this a so-called *fully-connected* FFNN. This is the type of NN that is used in the present work, and will be further investigated in section 3.3.

A different variant of FFNNs are *convolutional neural networks* (CNNs) [25], which have a connectivity pattern inspired by the animal visual cortex. Individual neurons in the visual cortex only respond to stimuli from small sub-regions of the visual field, called a receptive field. This makes the neurons well-suited to exploit the strong spatially local correlation present in natural images. The response of each neuron can be approximated mathematically as a convolution operation.

CNNs emulate the behaviour of neurons in the visual cortex by enforcing a *local* connectivity pattern between nodes of adjacent layers: Each node in a convolutional layer is connected only to a subset of the nodes in the previous layer, in contrast to the fully-connected FFNN in Figure 3.4. Often, CNNs consist of

several convolutional layers that learn local features of the input, with a fully-connected layer at the end, which gathers all the local data and produces the outputs. They have wide applications in image and video recognition [7].

3.2.2   Recurrent neural networks

So far we have only mentioned NNs where information flows in one direction: forward. *Recurrent neural networks* on the other hand, have connections between nodes that form directed *cycles*. This creates a form of internal memory which are able to capture information on what has been calculated before; the output is dependent on the previous computations. Recurrent NNs make use of sequential information by performing the same task for every element in a sequence, where each element depends on previous elements. An example of such information is sentences, making recurrent NNs especially well-suited for handwriting and speech recognition.

3.2.3   Other types of networks

There are many other kinds of NNs that have been developed. One type that is specifically designed for interpolation in multidimensional space is the radial basis function (RBF) network. RBFs are typically made up of three layers: an input layer, a hidden layer with non-linear radial symmetric activation functions and a linear output layer ("linear" here means that each node in the output layer has a linear activation function). The layers are normally fully-connected and there are no cycles, thus RBFs can be viewed as a type of fully-connected FFNN. They are however usually treated as a seperate type of NN due the unusual activation functions (common activation functions are presented in section 3.4).

Other types of NNs could also be mentioned, but are outside the scope of this work. We will now move on to a detailed description of how a fully-connected FFNN works, and how it can be used to interpolate data sets.

3.3   Multilayer perceptron

In this thesis we use fully-connected feed-forward neural networks with three or more layers (an input layer, one or more hidden layers and an output layer), mainly consisting of neurons that have non-linear activation functions. Such networks are often called *multilayer perceptrons* (MLPs)[1]. The *perceptron* was first introduced by Rosenblatt in 1958 [26], and was a FFNN made up of an input layer and an ouput layer only. It served as a *binary classifier*, i.e. a function

---

[1]The terms "feed-forward neural network" and "multilayer perceptron" are used interchangeably in the literature, although the MLP is just one type of FFNN, namely a fully-connected one with mainly non-linear neurons.

that maps a real-valued vector $\mathbf{x}$ to a single binary value $f(\mathbf{x})$, employing a linear activation function. The addition of hidden layers and non-linear activation functions led to the MLP, which could produce an arbitrary number of continous output values.

3.3.1   Why multilayer perceptrons?

We have chosen to use MLPs to interpolate data sets for the construction of inter-atomic potentials. Other NN types could also have been used, but our choice is well justified. According to the *Universal approximation theorem* [4], a feed-forward neural network with just a single hidden layer containing a finite number of neurons can approximate a continous multidimensional function to arbitrary accuracy, assuming the activation function for the hidden layer is a "non-constant, bounded and monotonically-increasing continous function". Note that the requirements on the activation function only applies to the hidden layer, the output nodes are always assumed to be linear, so as to not restrict the range of output values.

The only requirement on the NN information flow is that it must be feed-forward, which is not satisfied by recurrent NNs. Anyhow, computing energies and forces in MD simulations at a given time step do not require any knowledge of the earlier states of the system, i.e. no memory or recurrency is involved. Further, it is not obvious how local connectivity should be helpful in a MD context, thus we might as well use the simplest multilayer FFNN available, the MLP.

We note that this theorem is only applicable to a NN with *one* hidden layer. Therefore, we can easily construct a NN that employs activation functions which do not satisfy the above requirements, as long as we have at least one layer with activation functions that *do*. Furthermore, although the universal approximation theorem lays the theoretical foundation for regression with neural networks, it does not say anything about how things work in practice: A NN can still be able to approximate a given function reasonably well without having the flexibility to fit *all other* functions.

3.3.2   Mathematical model

In an MLP, each node (neuron) is modelled according to (3.1). To increase the flexibility of the NN, we also add a *bias b* to this model, resulting in

$$y = f\left(\sum_{i=1}^{n} w_i x_i + b_i\right) = f(u) \tag{3.2}$$

In a FFNN of such neurons, the *inputs* x_i are the *outputs* of the neurons in the preceding layer. Furthermore, a MLP is fully-connected, which means that each
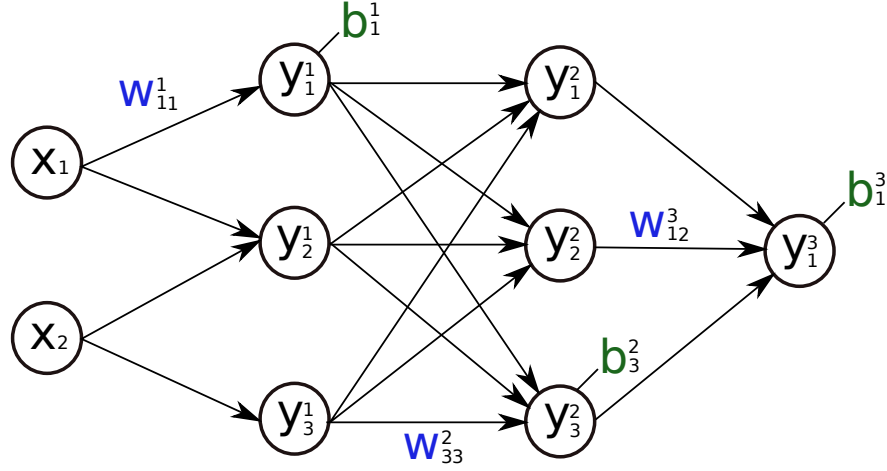
Figure 3.5: Example of a fully-connected feed-forward neural network with an input layer, two hidden layers and an output layer. Nodes are depicted as circles, while the arrows shows the connections between neurons, including their directions. Each connection has a weight w (blue), where the notation explained in the text is applied. Every node is marked with its output y and its associated bias b (green), while all input nodes are labeled with an x. Only a few weights and biases are displayed.

neuron receives a weighted sum of the outputs of *all* neurons in the previous layer.

Figure 3.5 displays the same NN as in Figure 3.4, now with all nodes labeled by their output y, except for the input nodes, which are labeled with x. The input nodes are strictly speaking not artifical neurons, as they have no activation function. Their only purpose is to forward the NN inputs to the first hidden layer. A few weights w (blue) and biases b (green) are also included, next to the connections and nodes which they belong to respectively. The following notation is introduced: y_i^l is the output of the i-th node in layer l, where $l = 0$ is the input layer, $l = 1$ is the first hidden layer and so on. The same notation applies to the biases. For the weights, we have that w_{ij}^l is the weight connecting node j in layer $l-1$ with node i in layer l[2]. All weights and biases are real-valued numbers.

The MLP in Figure 3.5 maps the inputs x_i to the output y_1^3 via two hidden layers with an (arbitrary) number of hidden nodes. The hidden neurons have no physical meaning, but have the purpose of defining the functional form of the NN. In the following we describe how to calculate the output of this MLP. The description is also valid for a NN with several outputs; the result applies to each ouput node individually.

First, for each node i in the first hidden layer, we calculate a weighted sum

---

[2]This order of indicies i and j may seem illogical, but will become useful when converting to matrix equations below.

u_i^1 of the input coordinates x_j,

$$u_i^1 = \sum_{j=1}^{2} w_{ij}^1 x_j + b_i^1 \tag{3.3}$$

This value is the argument to the activation function f_1 of each neuron i, producing the output y_i^1 of all neurons in layer 1,

$$y_i^1 = f_1(u_i^1) = f_1 \left(\sum_{j=1}^{2} w_{ij}^1 x_j + b_i^1 \right) \tag{3.4}$$

where we assume that all nodes in the same layer have identical activation functions, hence the notation f_l[3]. For an arbitrary node i in layer l this generalizes to

$$y_i^l = f_l(u_i^l) = f_l \left(\sum_{j=1}^{N_{l-1}} w_{ij}^l y_j^{l-1} + b_i^l \right) \tag{3.5}$$

where N_l is the number of nodes in layer l. When the output of all the nodes in the first hidden layer are computed, the values of the subsequent layer can be calculated and so forth until the output is obtained. The output of neuron i in layer 2 is thus,

$$y_i^2 = f_2 \left(\sum_{j=1}^{3} w_{ij}^2 y_j^1 + b_i^2 \right) \tag{3.6}$$

$$= f_2 \left[\sum_{j=1}^{3} w_{ij}^2 f_1 \left(\sum_{k=1}^{2} w_{jk}^1 x_k + b_j^1 \right) + b_i^2 \right] \tag{3.7}$$

where we have substituted y_m^1 with (3.4). Finally, the NN output yields,

$$y_1^3 = f_3 \left(\sum_{j=1}^{3} w_{1m}^3 y_j^2 + b_1^3 \right) \tag{3.8}$$

$$= f_3 \left[\sum_{j=1}^{3} w_{1j}^3 f_2 \left(\sum_{k=1}^{3} w_{jk}^2 f_1 \left(\sum_{m=1}^{2} w_{km}^1 x_m + b_k^1 \right) + b_j^2 \right) + b_1^3 \right] \tag{3.9}$$

We can generalize this expression to a MLP with l hidden layers. The complete functional form is,

$$y_1^{l+1} = f_{l+1} \left[\sum_{j=1}^{N_l} w_{1j}^3 f_l \left(\sum_{k=1}^{N_{l-1}} w_{jk}^2 f_{l-1} \left(\dots f_1 \left(\sum_{n=1}^{N_0} w_{mn}^1 x_n + b_m^1 \right) \dots \right) + b_k^2 \right) + b_1^3 \right] \tag{3.10}$$

---

[3]Note that for the activation functions, the layer indicies are subscripts rather than superscripts, as having a single superscript can easily be confused with a square.
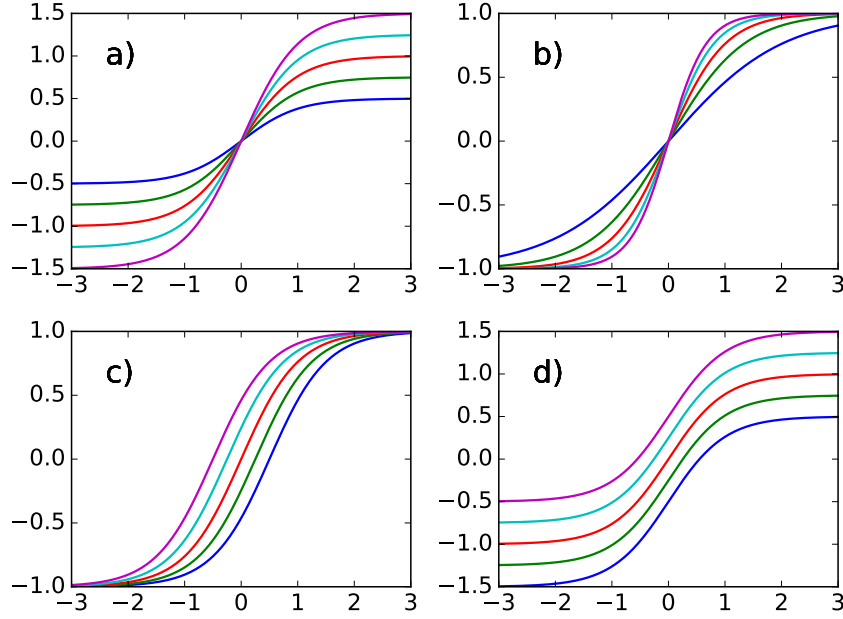
Figure 3.6: The mathematical expression for a MLP consists of nested terms of the form $h(x) = c_1 f(c_2 x + c_3) + c_4$ (3.10), where f is the activation function and c_i are NN parameters. The flexibility of the MLP is shown by adjusting c_1, c_2, c_3 and c_4 such that $h(x)$ a) is scaled, b) has a change of slope, c) is shifted left and right, d) is shifted up and down, respectively. Here f is the hyperbolic tangent (3.16). Reproduced from Behler [1].

which illustrates a basic property of MLPs: The only independent variables are the input values x_n. This confirms that a MLP, despite its quite convoluted mathematical form, is nothing more than an analytic function, specifically a mapping of real-valued vectors $\mathbf{x} \in \mathbb{R}^n \to \mathbf{y} \in \mathbb{R}^m$. In our example (Figure 3.5), $n = 2$ and $m = 1$. Consequentially, the number of input and output values of the function we want to fit must be equal to the number of inputs and outputs of our MLP.

Furthermore, the flexibility and universality of a MLP can be illustrated by realizing that the expression (3.10) is essentially a nested sum of scaled activation functions of the form

$$h(x) = c_1 f(c_2 x + c_3) + c_4 \qquad (3.11)$$

where the parameters c_i are weights and biases. By adjusting these parameters, the activation functions can be shifted up and down or left and right, change slope or be rescaled (Figure 3.6), which is the key to the flexibility of a NN.

Activation function of output neuron

We stated in subsection 3.3.1 that the output neuron always have a linear activation function for regression to avoid restricting the output values. This is a requirement for a MLP to operate as a mapping $\mathbf{x} \in \mathbb{R}^n \to \mathbf{y} \in \mathbb{R}^m$. In this thesis, we will exclusively use a specific linear activation function for the output neuron o, namely the identity function $f(x) = x$,

$$f_o = f(u_o) = u_o \tag{3.12}$$

Matrix-vector notation

We can introduce a more convenient notation for the activations in a NN. If we look at the structure of the NN in Figure 3.5, we realize that all the weights connecting two adjacent layers can be represented as a matrix W_l, where the (i, j)-th element is is the weight w_{ij}^l connecting connecting node j in layer $l - 1$ with node i in layer l, as above.

Additionaly, we can represent the biases and activations as layer-wise column vectors $\mathbf{b}_l$ and $\mathbf{y}_l$, so that the i-th element of each vector is the bias b_i^l and activation y_i^l of node i in layer l respectively.

We have that W_l is a $N_{l-1} \times N_l$ matrix, while $\mathbf{b}_l$ and $\mathbf{y}_l$ are $N_l \times 1$ column vectors. With this notation, the sum in (3.7) becomes a matrix-vector multiplication, and we can write the equation for the activations of hidden layer 2 in Figure 3.5 as[4]

$$\mathbf{y}_2 = f_2(W_2\mathbf{y}_1 + \mathbf{b}_2) = f_2 \left(\begin{bmatrix} w_{11}^2 & w_{12}^2 & w_{13}^2 \\ w_{21}^2 & w_{22}^2 & w_{23}^2 \\ w_{31}^2 & w_{32}^2 & w_{33}^2 \end{bmatrix} \cdot \begin{bmatrix} y_1^1 \\ y_2^1 \\ y_3^1 \end{bmatrix} + \begin{bmatrix} b_1^2 \\ b_2^2 \\ b_3^2 \end{bmatrix} \right) \tag{3.13}$$

and we see that the activation of node i in layer 2 is

$$y_i^2 = f_2 \left(w_{i1}^2 y_1^1 + w_{i2}^2 y_2^1 + w_{i3}^2 y_3^1 + b_i^2 \right) = f_2 \left(\sum_{j=1}^{3} w_{ij}^2 y_j^1 + b_i^2 \right) \tag{3.14}$$

which is in accordance with (3.7). Note that This is not just a convenient and compact notation, but also a useful and intuitive way to think about MLPs: The output is calculated by a series of matrix-vector multiplications and vector additions that are used as input to the activation functions. For each operation $W_l\mathbf{y}_{l-1}$ we move forward one layer. Furthermore, it is the easiest way to implement NNs with TensorFlow, which is discussed in subsection 6.2.2.

---

[4]For the same reasons as the activation functions above, the layer indicies are changed from superscripts to subscripts with this matrix-vector notation.
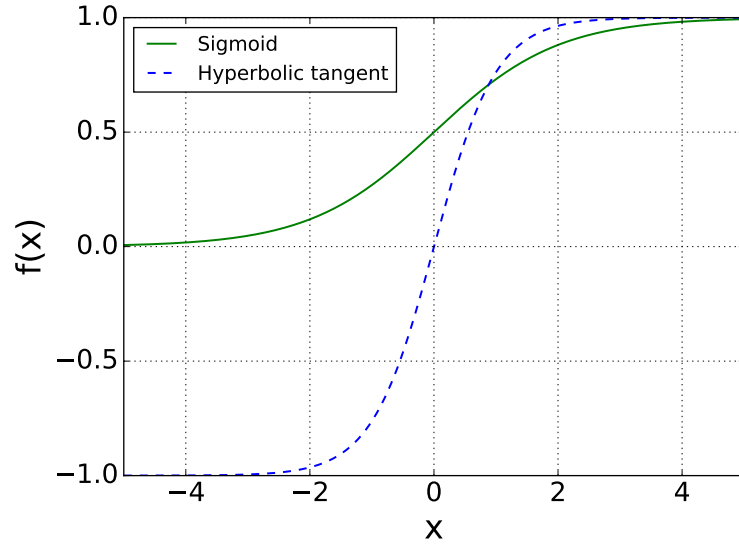
Figure 3.7: Two of the most common activation functions for neural networks, the sigmoid (3.15) and the hyperbolic tangent (3.16). Both meet the requirements of the universal approximation theorem.

3.4   Activation functions

A property that characterizes a NN, other than its connectivity, is the choice of activation function(s). As described in subsection 3.3.1, the following restrictions are imposed on an activation function for a FFNN to fulfill the universal approximation theorem [4]:

1. Non-constant

2. Bounded

3. Monotonically-increasing

4. Continous

We realize that the second requirement excludes all linear functions. Furthermore, in a MLP with only linear activation functions, each layer simply performs a linear transformation of its inputs. Consequentially, regardless of the number of layers, the output of the NN will be nothing but a linear function of the inputs. Thus we need to introduce some kind of non-linearity to the NN to be able to fit non-linear functions.

   A group of non-linear functions that meet the above requirements are the *sigmoid* functions. The two sigmoid functions that are most commonly used as activation functions [5] in neural networks are the logistic function

$$f(x) = \frac{1}{1 + e^{-x}} \tag{3.15}$$
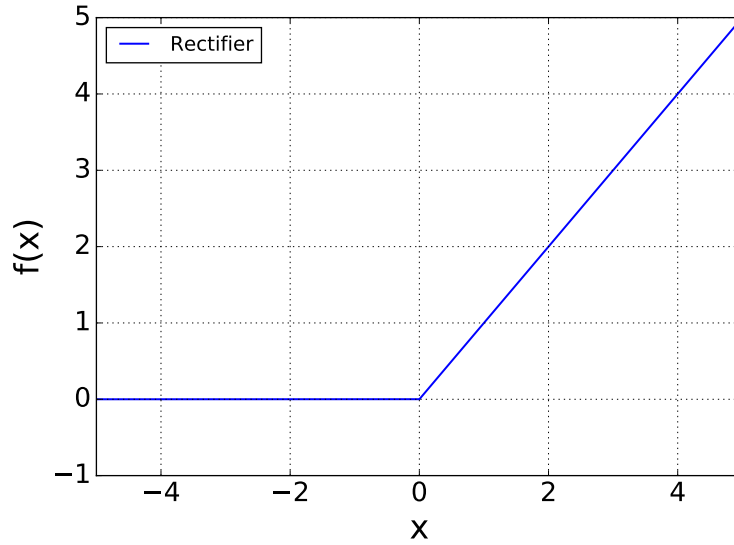
Figure 3.8: The rectifier (3.17) has become the most popular activation function for deep convolutional NNs.

and the hyperbolic tangent

$$f(x) = \tanh(x) \tag{3.16}$$

The logistic functon is often referred to as "the sigmoid", which is the naming convention we will use in the following. Both functions are depicted in Figure 3.7. We observe that they are bounded above by 1, while the lower bound is -1 and 0 for the hyperbolic tangent and sigmoid respectively. This property keeps the activations from diverging.

The sigmoid are more bilogically plausible because the output of inactive neurons are zero. Such activation function are called *one-sided*. However, it has been shown [6] that the hyperbolic tangent performs better than the sigmoid for training MLPs. This will be further discussed in subsection 7.3.2. Nevertheless, one should assess the problem at hand when deciding what activation function to use; the performance can vary from problem to problem.

In later years, the rectifier function (Figure 3.8)

$$f(x) = \max(0, x) \tag{3.17}$$

has become the most popular [7] for *deep neural networks*, i.e. NNs with several hidden layers[5]. It has been argued to be even more biologically plausible than the sigmoid and also perform better than the hyperbolic tangent, especially for deep, convolutional NNs [8]. Even though the rectifier is a piecewise *linear* function, it

---

[5]There is no consensus in the literature on how many hidden layers a NN must possess in order to qualify as "deep", but there seems to be a tendency towards setting the minimum number of layers to three.

effectively behaves as a non-linearity. However, it does not have an upper bound, and therefore does not satisfy the universal approximator requirements. Further, the function values can potentially blow up. These problems are solved in practice by employing various teqhniques, illustrating that the universal approximation theorem should not be used as an absolute guideline for the construction of activation functions.

3.5   Training

In the beginning of this chapter we defined machine learning as the "science of giving computers the ability to learn without being explicitly programmed." In the neural network case, this learning is achieved by iteratively feeding the network with data. With the help of certain learning algorithms, the network will then automatically adjust its parameters, i.e. the weights and biases, to find patterns in this data[6]. The procedure depends on what learning paradigm we work with (unsupervised or supervised learning), and the type of NN we use.

In this thesis we employ NNs for regression: We want to interpolate data sets og atomic coordinates and energies. The goal is to adjust the weights of a MLP so that it accurately represents a function that maps atomic coordinates to the corresponding potential energies. Interpolation with NNs is essentialy an optimization problem: The weights are typically initialized as random numbers, and then modified iteratively to minimize the error to a set of expected output values, in our case energies. This procedure is called *training*.

Thus, the input values to the function that we want to fit are atomic configurations, while the output values are the total energies of these configurations. Together they form what we refer to as a *data set* or a *reference set*. The output values are also known as the *target values*. The input data X and output data Y are represented as matrices, where the i-th row of X and Y, denoted X_{i*} and Y_{i*} respectively, together form a *training example*. The combined data set consists of N such training examples.

Note that the number of columns of X and Y are equal to the number of inputs and outputs of our MLP respectively. This is a restatement of the observation made in subsection 3.3.2: The number of inputs and outputs of our MLP must be equal to the number of input and output values of the function we are trying to fit. In the following we assume that the MLP only has one output, thus each row of Y is a single number, denoted $Y_{i*} = Y_i$.

---

[6]in the following, "weights" will often be equivalent to all the NN parameters, including the biases.

3.5.1 Cost functions

To minimize the error, we need a way to *define* it. As in mathematical optimization, the error is represented by an objective function, also called *loss* function or *cost* function. Training a NN therefore amounts to the minimization of this function, which can be written in the following general way,

$$\Gamma = \Gamma\big(\{W_l\}, \{\mathbf{b}_l\}, X, Y\big) \tag{3.18}$$

where $\{W_l\}$ and $\{\mathbf{b}_l\}$ are all the weights and biases of the NN respectively. The value of this function is a measure of how well the NN is able to map $X \to Y$. By adjusting $\{W_l\}$ and $\{\mathbf{b}_l\}$, we try to minimize the value of this function.

The standard cost function used in regression with NNs is the mean-square error,

$$\Gamma = \frac{1}{2N} \sum_{i=1}^{N} (Y_i - y_i)^2 \tag{3.19}$$

where $y_i = y_i(W, B, X_{i*})$ is the value predicted by the NN for training example X_{i*}. For a MLP with more than one output, Y_i and y_i are vectors. The constant $1/2$ is included to cancel out the exponent when this function is differentiated at a later stage (subsection 3.6.3).

A variety of other cost functions are used for other machine learning tasks. In classificiation the output is often binary, hence the mean-square error function is inadequate. We will not go into further detail on these cost functions here, but subsection 7.3.5 provides a discussion on an extension of (3.19) that is relevant for the construction of NNPs.

3.6 Optimization

There are a large number of algorithms that can be used to determine the set of weights minimizing the cost function (3.19). Different kinds of gradient descent-based methods are widely used, while higher order methods like the conjugate gradient algorithm or Newton's method are rarely seen in the literature because they are too computationally expensive for large NNs and data sets. Therefore, only first-order methods will be discussed here.

The idea behind gradient descent methods is to minimize a function by iteratively taking steps in the direction of steepest descent towards a minima in parameter space. This direction is defined as the negative gradient of the function with respect to all its parameters. If we define $\boldsymbol{\theta} \in \mathbb{R}^d$ as a vector containing all the weights and biases of a MLP, we get the following iterative scheme,

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \gamma \nabla_{\boldsymbol{\theta}_k} \Gamma(\boldsymbol{\theta}) \tag{3.20}$$

where γ is a step size, called the *learning rate* in the context of machine learning. The process is started by initializing the parameters as random numbers. The value of γ is of great importance for the algorithm to converge, and is allowed to change at every iteration. Note that convergence to a minimum is not guaranteed withouth certain assumptions on the function Γ and the particular choices of γ. Also, the obatained minimum is generally local, not global. Different ways to update γ is discussed below.

3.6.1   Gradient descent variants

There are three versions of gradient descent [27], which differ in the number of training examples we present to the NN before updating the parameters. According to neural network terminology, the process of adjusting the parameters based on all the training examples, either in batches or all at once, is called an *epoch*.

In *batch* gradient descent, we compute the gradient of the cost function for the *whole* data set before updating, also called *offline learning*. This approach can be very slow and is intractable for datasets that do not fit in memory. Furthermore, there is a risk of performing redundant computations if many similar examples are present in the data set.

In contrast, *stochastic* gradient descent (SDG) performs a parameter update for *each* training example. SDG avoids redundant gradient calculations and is therefore faster than batch gradient descent. However, the accuracy of each update is lower compared to offline learning, which can lead to quite large error oscillations. The concept of having more than one parameter update per epoch is called *online learning*.

Finally, we have *mini-batch* gradient descent, which is a mix of the two other approaches. The reference set is divided into n equally sized mini-batches and a parameter update is performed for each mini-batch. This is usually the algorithm of choice, as it has the optimal trade-off between speed and accuracy. Normal mini-batch sizes range between 50 and 256, but should in some degree be tailored to each problem.

We realize that SDG can be seen as a variant of mini-batch gradient descent with a mini-batch size of 1. In the literature, these two methods are often collectively referred to as SDG.

3.6.2   Optimization algorithms

In the following we will outline different optimization algorithms that are widely used in neural network research. They are all variations on the update rule (3.20). The main focus of the methods is to find a proper learning rate. A learning rate that is too small may lead to very slow convergence, while a learning rate that

is too large can cause the loss function to fluctuate around the minimum or even diverge.

Momentum

Near local minima, the surface area of the cost function (in parameter space) often curve much more steeply in one direction than in another, forming ravines. SDG will have slow converge in such regions, as it will oscillate across the slopes while having slow progress along the bottom of the ravine. The Momentum method [15] avoids this problem by accelerating SDG in the downwards direction, while damping the oscillations. The update rule is

$$\begin{aligned}
\mathbf{v}_k &= \eta \mathbf{v}_{k-1} + \gamma \nabla_k \Gamma \\
\boldsymbol{\theta}_{k+1} &= \boldsymbol{\theta}_k - \mathbf{v}_k
\end{aligned} \tag{3.21}$$

where η is the momentum term, which is usually set to 0.9 or a similar value. The new update vector $\mathbf{v}_k$ is formed by adding a fraction η of the previous update vector $\mathbf{v}_{k-1}$ to the gradient. In this way, the magnitude of the update decreases for parameters (dimensions) whose gradients change direction, while it increases for parameters whose gradients points in the same direction as in the previous step. This helps to push SDG downhill towards the minima.

Adagrad

In SDG and Momentum, we have one learning rate γ that is applied to all the parameters. Ideally, the learning rate should be adapted to each individual parameter to perform larger or smaller updates depending on their importance. Adagrad [16] does just that. For brevity, we set

$$g_{k,i} = \nabla_{\theta_i} \Gamma(\boldsymbol{\theta}) \tag{3.22}$$

to be the gradient of the loss function w.r.t. parameter θ_i at step k. Adagrad adjusts the general learning rate γ at each step k for every parameter θ_i based on the past gradients for that parameter,

$$\theta_{k+1,i} = \theta_{k,i} - \frac{\gamma}{\sqrt{G_{k,ii} + \epsilon}} \cdot g_{k,i} \tag{3.23}$$

where $G_k \in \mathbb{R}^{d \times d}$ is a diagonal matrix where each diagonal element $G_{k,ii}$ is the sum of squares of the gradient with respect to θ_i up to step k. The smoothing constant $\epsilon \sim 10^{-8}$ is present to avoid division by zero. The vectorized version of this equation yields,

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \frac{\gamma}{\sqrt{\text{diag}(G_k) + \epsilon}} \odot \mathbf{g}_k \tag{3.24}$$

where $\odot$ stands for element-wise multiplication. In other words, we do not have to manually tune the learning rate, Adagrad does that for us. A weakness of this method is that the learning rates inevitably shrinks for each step, resulting in arbitrary small values, at which point the learning stops.

For the rest of subsection 3.6.2, we assume for brevity that all vector operations are element-wise, i.e. $\mathbf{g}\,\mathbf{h} = \mathbf{g} \odot \mathbf{h}$ and $\mathbf{g}^2 = \mathbf{g} \odot \mathbf{g}$.

Adadelta

Adadelta [17] is an extension of Adagrad that reduces the rate at which the learning rate decreases. The sum of all past squared gradients is replaced by a *exponentially decaying average* of all the previous squared gradients. We introduce the vector $\mathbf{E}[\mathbf{g}^2]_k$ containing the decaying averages of the gradient with respect to all the parameters at step k. This vector is defined recursively as a weighted average of the previous averages $\mathbf{E}[\mathbf{g}^2]_{k-1}$ and the current gradient $\mathbf{g}_k$,

$$\mathbf{E}[\mathbf{g}^2]_k = \eta \mathbf{E}[\mathbf{g}^2]_{k-1} + (1 - \eta)\mathbf{g}_k^2 \tag{3.25}$$

where η is a decay constant similar to that in the Momentum method, and is usually set to the same value (0.9). We also define a new parameter update vector $\Delta\boldsymbol{\theta}_k = -\mathbf{v}_k$ so that

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k + \Delta\boldsymbol{\theta}_k \tag{3.26}$$

Replacing the vector $\mathrm{diag}(\mathrm{G}_k)$ in (3.24) with the decaying average vector (3.25) yields the following parameter update vector,

$$\Delta\boldsymbol{\theta}_k = -\frac{\gamma}{\sqrt{\mathbf{E}[\mathbf{g}^2]_k + \epsilon}}\mathbf{g}_k \tag{3.27}$$

Additionally, the learning rate γ is replaced by a decaying average of previous squared parameter updates $\mathbf{E}[\Delta\boldsymbol{\theta}^2]$ up to step $k-1$, yielding the Adadelta update rule,

$$\mathbf{v}_k = \frac{\sqrt{\mathbf{E}[\Delta\boldsymbol{\theta}^2]_{k-1} + \epsilon}}{\sqrt{\mathbf{E}[\mathbf{g}^2]_k + \epsilon}}\mathbf{g}_k \tag{3.28}$$

The final replacement is done to obtain correct units for the update vector and to eliminate the learning rate from the equation. Thus, we do not even need to set a default learning rate with Adadelta.

Adam

The final algorithm we are going to discuss is Adapte Moment Estimation (Adam) [18]. This method computes adaptive learning rates for each parameter by storing

exponentially decaying averages of both the gradients *and* the squared gradients (3.25),

$$\begin{aligned} \mathbf{E}[\mathbf{g}]_k &= \beta_1 \mathbf{E}[\mathbf{g}]_{k-1} + (1 - \beta_1)\mathbf{g}_k \\ \mathbf{E}[\mathbf{g}^2]_k &= \beta_2 \mathbf{E}[\mathbf{g}^2]_{k-1} + (1 - \beta_2)\mathbf{g}_k^2 \end{aligned} \qquad (3.29)$$

We set $\mathbf{a}_k = \mathbf{E}[\mathbf{g}]_k$ and $\mathbf{b}_k = \mathbf{E}[\mathbf{g}^2]_k$ so that

$$\begin{aligned} \mathbf{a}_k &= \beta_1 \mathbf{a}_{k-1} + (1 - \beta_1)\mathbf{g}_k \\ \mathbf{b}_k &= \beta_2 \mathbf{b}_{k-1} + (1 - \beta_2)\mathbf{g}_k^2 \end{aligned} \qquad (3.30)$$

where $\mathbf{a}_k$ and $\mathbf{b}_k$ are estimates of the first moment (the mean) and the second moment (the uncentered variance) of the gradients respectively, hence the name of the method. The elements of these two vectors are initialized as zeros, which make them biased towards zero, especially during the first time steps. This bias is counteracted by computing bias-corrected version of the vectors (derived in [18]),

$$\begin{aligned} \hat{\mathbf{a}}_k &= \frac{\mathbf{a}_k}{1 - \beta_1^k} \\ \hat{\mathbf{b}}_k &= \frac{\mathbf{b}_k}{1 - \beta_2^k} \end{aligned} \qquad (3.31)$$

The parameter update vector for Adam is obtained by replacing the gradient $\mathbf{g}_k$ with $\mathbf{a}_k$ in (3.27), yielding the update rule,

$$\boldsymbol{\theta}_{k+1} = \boldsymbol{\theta}_k - \frac{\gamma}{\sqrt{\hat{\mathbf{b}}_k} + \epsilon}\hat{\mathbf{a}}_k \qquad (3.32)$$

The default values of the hyperparameters are $\beta_1 = 0.9$, $\beta_2 = 0.999$ and $\epsilon = 10^{-8}$.

Which optimizer to use?

There is no general answer to the question stated above. The adaptive learning rate methods like Adagrad, Adadelta and Adam are generally more robust because they do not rely on manual fine-tuning of the learning rate. Also, an adaptive learning rate generally handles *sparse* data better, i.e. data where some learning features are poorly represented.

Adam has been proven to perform favourably over the other optimization methods for certain applications like classification with MLPs and convolutional NNs [18]. However, the performance of various optimizers should be tested on the problem at hand to determine which method that works best.

3.6.3   Backpropagation

In the preceding sections we have discussed different optimizers that implements various update rules for the parameters of a loss function. In our case, these

parameters are the weights and biases of a neural network and the cost function Γ is the mean-square-error (3.19)

$$\Gamma = \frac{1}{2}(Y_i - y_i)^2 \tag{3.33}$$

here written as a function of only *one* training example i. What we have not discussed yet, is how we calculate the gradient of the cost function, i.e. how we obtain the partial derivatives g_{ij} and h_i with respect to all the weights and biases respectively,

$$g_{ij} = \frac{\partial \Gamma}{\partial w_{ij}} \tag{3.34}$$

$$h_i = \frac{\partial \Gamma}{\partial b_i} \tag{3.35}$$

where we have dropped layer indicies for clarity. A common method to obtain these derivatives is *backpropagation* [9]. In backpropagation, a training example is propagated forward through the NN to produce an output. This output is compared to the desired output (target values), and the error is then propagated *backwards* through the layers to obtain the amount of which each parameter should be adjusted, hence the name. The method is essentially an implementation of the chain rule, and will allow us to calculate the partial derivatives of the cost with respect to all the parameters, thereby obtaining the gradient of the network.

1. Forward propagation

The first stage of the algorithm is forward propagation, producing an output as described in subsection 3.3.2[7]. In the following, we will use the same notation as in subsection 3.3.2. We also define

$$A_i = \{j : w_{ij}\} \tag{3.36}$$

as the set $\{j\}$ of nodes anterior to node i and connected to node i with weights w_{ij}, in addition to

$$P_j = \{i : w_{ij}\} \tag{3.37}$$

as the set $\{i\}$ of nodes posterior to node j and connected to node j with weights w_{ij}.

---

[7]We assume that the MLP only has one output. As before, the results are valid for each individual output in a many-output MLP.

2. Backward propagation

The second step is to backpropagate the error of the NN output. The weight derivative (3.34) can be expanded into two factors by use of the chain rule,

$$g_{ij} = \frac{\partial \Gamma}{\partial u_i} \frac{\partial u_i}{\partial w_{ij}} \tag{3.38}$$

Now we move in the opposite direction of the feed-forward stage: First we differentiate the cost w.r.t the input of neuron i, then we differentiate the input w.r.t. weight w_{ij} connecting neurons j (in the preceding layer) and i. The first term on the r.h.s. is defined as the the error δ_i of node i,

$$\delta_i = \frac{\partial \Gamma}{\partial u_i} \tag{3.39}$$

This definition can be justified by the following argument. To change the value of a cost function (3.18), we need to change the outputs of the neurons in the network. Changing the input u_j to neuron j by a small amount Δu_j results in the output

$$y_j = f_j(u_j + \Delta u_j) \tag{3.40}$$

This change will propagate through subsequent layers in the network, finally causing the overall cost to change by an amount $\frac{\partial \Gamma}{\partial u_j} \Delta u_j$. If $\partial \Gamma / \partial u_j$ is close to zero, we are not able improve the cost much by perturbing the weighted input u_j; the neuron is already quite near the optimal value. This is a heuristic argument for $\partial \Gamma / \partial u_j$ to be a measure of the error of the neuron.
The second term on the r.h.s. of (3.38) is

$$\frac{\partial u_i}{\partial w_{ij}} = \frac{\partial}{\partial w_{ij}} \sum_{m \in A_i} w_{im} y_m = y_j \tag{3.41}$$

Combining the the two terms (3.39) and (3.41) yields

$$g_{ij} = \delta_i y_j \tag{3.42}$$

To compute this quantity, we thus need to know the outputs and the errors of all nodes in the network. The outputs are generated during the feed-forward stage (3.5),

$$y_i = f_i(u_i) = f_i \left(\sum_{j \in A_i} w_{ij} y_j + b_i \right) \tag{3.43}$$

and need to be stored for all nodes. The errors are obtained by backpropagating the error of the output neuron o,

$$\delta_o = \frac{\partial \Gamma}{\partial u_o} = \frac{\partial \Gamma}{\partial y_o} \frac{\partial y_o}{\partial u_o} = (Y - y_o) \frac{\partial y_o}{\partial u_o} \tag{3.44}$$
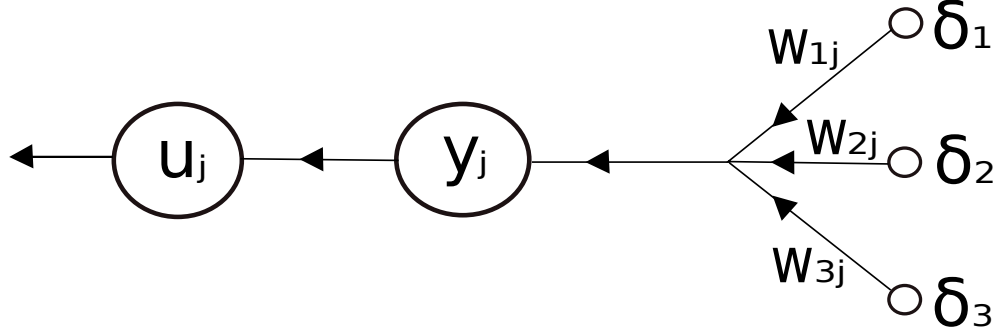
Figure 3.9: Illustration of the backpropagation algorithm. The direction of information flow is opposite of Figure 3.2. A hidden neuron j receives a weighted sum of errors of all nodes in the posterior layer. Then, we differentiate in the backwards direction: the activation function (output) y_j of node j is differentiated w.r.t to its net input u_j.

We remember from subsection 3.3.2 that the output neuron has the identity activation function $y_o = f_o(u_o) = u_o$, which reduces the error to

$$\delta_o = Y - y_o \tag{3.45}$$

This error is then propagated backwards through the network, layer by layer. The error of each neuron in layer l thus depends on the errors of all nerons in layer $l+1$. Consequentially, the error of an arbitrary hidden neuron can be written as a recursive equation (Figure 3.9),

$$\delta_j = \frac{\partial \Gamma}{\partial u_j} = \sum_{i \in P_j} \frac{\partial \Gamma}{\partial u_i} \frac{\partial u_i}{\partial y_j} \frac{\partial y_j}{\partial u_j} \tag{3.46}$$

We observe that also the inputs u_i to all nodes need to be stored during forward propagation for this scheme to work. The first term on the r.h.s. of (3.46) is the error δ_i of node i, which is assumed to be known. As long as there are no cycles in the network, there is an ordering of nodes from the output back to the input that respects this condition. It is therefore valid only for feed-forward NNs.

The second term is the derivative of the net input of all posterior nodes $\{i\}$ w.r.t. the output of node j,

$$\frac{\partial u_i}{\partial y_j} = \frac{\partial}{\partial y_j} \left(\sum_{m \in A_i} w_{im} y_m + b_i \right) = w_{ij} \tag{3.47}$$

while the third is the derivative of node j's activation function w.r.t. its net input:

$$\frac{\partial y_j}{\partial u_j} = \frac{\partial f_j(u_j)}{\partial u_j} \equiv y_j' \tag{3.48}$$

Putting all the pieces together we obtain

$$\delta_j = y'_j \sum_{i \in P_j} w_{ij} \delta_i \tag{3.49}$$

This expression is illustrated in Figure 3.9. Each hidden neuron j recieves a weighted sum of the errors of all nodes in the posterior layer. Then we differentiate in the backwards direction (compare with Figure 3.2): the output y_j of node j is differentiated w.r.t. its input u_j.

The errors are propagated backwards through the whole NN until we reach the weights connecting the input layer and the first hidden layer. By propagating the error of only one ouput neuron, we thus obtain the errors of all the neurons at once. This is the main strength of the backpropagation algorithm, and the reason for its popularity in neural network research.

For the biases, we have (3.35),

$$h_j = \frac{\partial \Gamma}{\partial b_j} = \sum_{i \in P_j} \frac{\partial \Gamma}{\partial u_i} \frac{\partial u_i}{\partial y_j} \frac{\partial y_j}{\partial u_j} \frac{\partial u_j}{\partial b_j} \tag{3.50}$$

The only new term here is the last one:

$$\frac{\partial u_j}{\partial b_j} = \frac{\partial}{\partial b_j} \left(\sum_{m \in A_j} w_{jm} y_m + b_j \right) = 1 \tag{3.51}$$

Consequentially, the bias gradient is simply the error of each neuron:

$$h_j = \delta_j \tag{3.52}$$

Matrix notation

In subsection 3.3.2 we rewrote the forward propagation framework as matrix-vector equations. This is easily done also for the backpropagation case. From subsection 3.3.2 we have the set of vectors $\mathbf{b}_l$ and $\mathbf{y}_l$ for $l = 1, \ldots, L+1$, where L is the number of hidden layers. We now extend this set with the vectors $\boldsymbol{\delta}_l$ and $\mathbf{u}_l$, i.e. the errors and inputs (or *preactivations*) of layer l respectively. These are all column vectors of size $N_l \times 1$.

Comparing the sums in (3.43) and (3.49), we realize that since

$$\sum_j w_{ij} y_j \;\Rightarrow\; \mathrm{W}\mathbf{y} \tag{3.53}$$

we have

$$\sum_i w_{ij} \delta_i \;\Rightarrow\; \mathrm{W}^T \boldsymbol{\delta} \tag{3.54}$$

i.e. the weight matrices employed in backpropagation are the transpose of the matrices used in forward activation. Thus, for nodes $\{j\}$ in layer l and nodes $\{i\}$ in layer $l+1$, the vectorized equation for the error of each hidden neuron is

$$\delta_j = y_j' \sum_{i \in P_j} w_{ij} \delta_i \;\Rightarrow\; \boldsymbol{\delta}_l = \mathbf{y}_l' \odot (W_{l+1} \boldsymbol{\delta}_{l+1}) \tag{3.55}$$

where $\odot$ signifies element-wise multiplication. The expression for the weight gradients (3.42) is converted into an outer product of the errors of layer $l+1$ and the outputs of layer l,

$$g_{ij} = \delta_i y_j \Rightarrow \mathrm{G}_l = \boldsymbol{\delta}_{l+1} \mathbf{y}_l^T \tag{3.56}$$

where G_l is a matrix containing all gradients for the weights connecting layer l and $l+1$. The corresponding matrix for the biases is a $N_l \times 1$ column vector $\mathbf{H}_l$,

$$\mathbf{H}_l = \boldsymbol{\delta}_l \tag{3.57}$$

The set $\{\mathrm{G}_l, \mathbf{H}_l\}$ for $l = 1, \ldots, L+1$ thus make up the total gradient of the cost function (3.19).

Training algorithm

We are now ready to express the complete training algorithm for a MLP with backpropagation using matrix-vector notation. We introduce $\mathrm{G}_{l,i}$ and $\mathbf{H}_{l,i}$ as the weight gradient matrix and the bias gradient vector for training example i.

1. Input a set of n training examples.

2. For each training example $\mathbf{x} = X_{i*}$:

   (a) Initialize the input layer:
   $$\mathbf{y}_0 = \mathbf{x} \tag{3.58}$$

   (b) Propagate the activity forward. For $l = 1, \ldots, L+1$:
   $$\mathbf{u}_l = \mathrm{W}_l \mathbf{y}_{l-1} + \mathbf{b}_l \tag{3.59}$$
   $$\mathbf{y}_l = f_l(\mathbf{u}_l) \tag{3.60}$$

   Store all vectors $\mathbf{u}_l$ and $\mathbf{y}_l$.

   (c) Calculate and store the error in the output layer:
   $$\boldsymbol{\delta}_{L+1} = \mathbf{Y} - \mathbf{y}_{L+1} \tag{3.61}$$

   (d) Backpropagate the error. For $l = L, L-1, \ldots, 1$:
   $$\boldsymbol{\delta}_l = \mathbf{y}_l' \odot (\mathrm{W}_{l+1}^T \boldsymbol{\delta}_{l+1}) \cdot \tag{3.62}$$

   Store all errors $\boldsymbol{\delta}_l$.

(e) Compute and store the weight and bias gradients.
   For $l = L + 1, L, \ldots, 1$:

$$G_{l,i} = \boldsymbol{\delta}_l \mathbf{y}_{l-1}^T, \quad \mathbf{H}_{l,i} = \boldsymbol{\delta}_l \tag{3.63}$$

3. Calculate the total gradients of all training examples i.
   For $l = L + 1, L, \ldots, 1$:

$$G_l = \sum_i^n G_{l,i} \quad \text{and} \quad \mathbf{H}_l = \sum_i^n \mathbf{H}_{l,i} \tag{3.64}$$

4. Update weights and biases with an update rule of choice.

The value of n depends on the learning paradigm that is used, described in subsection 3.6.1.

Chapter 4

Neural network potentials

In subsection 2.1.2 we discussed various ways to construct a PES for use in molecular dynamics, and stated that this thesis deales with neural network potentials (NNP). Then, in chapter 3 we outlined the basic theory of MLPs, which is our choice of neural network type. In this chapter, we will discuss the symmetry problems that arise when constructing NNPs, and introduce the Behler-Parrinello method for solving these symmetry issues.

4.1   Neural network potentials

NNPs are constructed by fitting an analytic function to a set of reference data obtained by quantum mechanical calulations. The main goal of the NN approach to constructing a PES, is to create a potential that has the accuracy of *ab inito* methods and the speed of empirical potentials (or at least of the same order).

Several different methods have been developed to construct NNPs. A comprehensive review of many of these has been written by Behler [1]. They mainly differ in the number of NNs that are employed to represent the total energy of the system, the dimensionality of each NN and the choice of input coordinates. By dimensionality we mean the number of input nodes, as the number of output nodes is assumed to be 1 if not specified otherwise.

4.1.1   Potentials using single neural network

In the literature, the dominant part of NN potentials employ a single feed-forward NN to calculate the total energy of the system. This is the most straightforward approach and is the easiest to implement. For instance, Agrawal et al. [35] have used a single, three-dimensional NN to study the dissociation of the SiO_2 molecule. However, these types of NNPs suffer from a number of limitations. Most importantly, they have a fixed dimensionality. This is a general property of MLPs where the atomic coordinates are fed directly to the network without

any transformations: Once fitted, the NNP can only be applied to systems with the same number of atoms as the number of inputs. Further, a large number of input nodes requires many hidden neurons, and there comes a point where the number of parameters of the NN makes training unfeasible. Thus, the number of degrees of freedom of the system must be small.

4.1.2 Potentials using multiple neural networks

For NNPs to be recognized as competitive potentials in molecular dynamics, they should be applicable to large systems of various sizes, containing thousands of atoms. This can not be achieved by employing only one NN to represent the total energy of the system. A solution is to replace the single NN by a set of *atomic* NNs, where each NN provides the contribution E_i of *one* atom to the total energy E of the system,

$$E = \sum_{i=1}^{N} E_i \tag{4.1}$$

Two methods using this approach have been developed in parallel by Bholoa et al. [36], and Behler and Parrinello [37]. Both methods compute E_i based on the chemical environment of atom i, but they differ in how the environment is represented and also how the atomic NNs are structured.

Bholoa et al. compute each atomic energy E_i by a NN of variable size, using a set of vectors as input to the NN. Each vector specifies the geometry of a four atom chain $i - j - k - l$ describing the chemical environment. The input layer and first hidden layer are replicated once for each vector, thus the problem of fixed dimensionality is solved.

Behler and Parrinello have developed a different approach to constructing a NNP. We have chosen to use the Behler-Parrinello (BP) method in this thesis, which will be the topic of the rest of the chapter.

4.2 The Behler-Parrinello method

In the Behler-Parrinello method, the architecture of each NN is fixed, and the atomic energies E_i depend on the local chemical environments up to a cutoff radius R_c. This is analogous to the introduction of a cutoff radius in MD simulations (section 2.4). Thus, the NNP can be applied to MD simlulations in excactly the same way as ordinary, empirical potentials: It is simply an analytic function that yields the potential energy of a central atom in the presence of neighbouring atoms contained in a cutoff sphere of radius R_c. In other words, the BP method employs an *atom-centered* approach. In the following we will refer to all atoms inside the cutoff sphere of an atom i as *neighbours*.

The cutoff is introduced by a set of so-called symmetry functions. These functions describe the local chemical environment of an atom i by transforming the cartesian coordinates of *all* its neighbours to a predefined number of function values, forming a symmetry vector. The symmetry vector is used as input to the atomic NN of atom i, that in turn calculates its potential energy E_i. The length of this vector must therefore be equal to the number of NN inputs.

Each atom has an individual atomic NN along with a fixed set of symmetry functions, but all atoms of the same element have identical atomic NNs and symmetry functions sets. Thus, for a diatomic system, only two unique NNs need to be trained.

As an example, let us look at a system containing two different elements: Si and O. All Si atoms have identical atomic NNs and symmetry function sets, the former represented by the function F_{Si}. A given Si atom i has at a given time n neighbours, including both Si and O atoms. We define $\mathbf{r}_{ij} = (r_{i1}, r_{i2}, \ldots, r_{in})$ as the vector containing the distances from atom i to all neighbours $j = 1, \ldots, n$, where $r_{ij} = |\mathbf{r}_j - \mathbf{r}_i|$. Further, $\mathbf{G}_{\text{Si}}(\mathbf{r}_{ij}) = \{G_{\text{Si},s}(\mathbf{r}_{ij})\}$ is the vector of Si symmetry functions $s = 1, \ldots, M_{\text{Si}}$[1]. The process of calculating the potential energy E_i described above can then be written,

$$\mathbf{r}_{ij} \rightarrow \mathbf{G}_{\text{Si}}(\mathbf{r}_{ij}) \rightarrow F_{\text{Si}}[\mathbf{G}_{\text{Si}}(\mathbf{r}_{ij})] \rightarrow E_i \qquad (4.2)$$

To obtain the corresponding expression for an O atom, we simply substitute Si $\rightarrow$ O. The expression is valid for any element e, and for a system with an arbitrary number of elements,

$$\mathbf{r}_{ij} \rightarrow \mathbf{G}_{\text{e}}(\mathbf{r}_{ij}) \rightarrow F_{\text{e}}[\mathbf{G}_{\text{e}}(\mathbf{r}_{ij})] \rightarrow E_i \qquad (4.3)$$

While the symmetry function vector of an element e is of a fixed size equal to the number of inputs to F_{e}, the length of $\mathbf{r}_{ij}$ can be arbitrary. This is enabled by the functional form of the symmetry functions, which is the topic of the next section.

We note that although the NNs and the symmetry vectors are identical for all atoms of the same element, the resulting energies E_i are not equal since they depend on the unique atomic environments $\mathbf{r}_{ij}$ of each atom.

4.2.1   Symmetry functions

We have mentioned two requirements that the symmetry function set have to meet. First, they must contain a cutoff r_c that defines a chemical environment. Secondly, since we employ atomic NNs with a fixed number of inputs, they have

---

[1]This notation signifies that each symmetry function, i.e. each element of the symmetry vector, depends on the coordinates of all atoms in the cutoff sphere. However, it does not say anything about the form of the dependence, e.g. if it involves triplets or just pairs of atoms etc.

to be able to produce a constant number of function values independent of the number of neighbours, which can change during MD simulations.

In addition to this, the NN has symmetry problems. It does not take into account that exchanging two or more atoms can lead to an energetically equivalent configuration. For example, exchanging the positions of both hydrogen atoms in a water molecule will not alter the energy. However, noting that all weights have numerically different values, changing the order of the input coordinates to the NN will result in a different energy value. This problem can be solved by employing symmetry functions that are invariant with respect to the order of the atoms.

Further, the energy of a system is invariant under a translation or rotation. Cartesian coordinates can therefore not be used directly as input to the NN, as their numerical values change under such operations, resulting in a different NN output.

We start with defining the chemical environments by introducing a cutoff function [2],

$$f_c(r_{ij}) = \begin{cases} 0.5\left[\cos\left(\frac{\pi r_{ij}}{r_c}\right) + 1\right], & r_{ij} \leq r_c \\ 0, & r_{ij} > r_c \end{cases} \tag{4.4}$$

which is the monotonically decreasing part of a cosine function on $r_{ij} \in [0, r_c]$ (Figure 4.1). This function have the desirable property that it decreases with increasing distance r_{ij} between the central atom i and its neighbour j. At the cutoff radius r_c, is has zero value and slope, which is important to avoid discontinuities when computing energies and forces. Atoms beyond the cutoff radius are not a part of the central atom's chemical environment and therefore do not contribute to its energy.

Several types of many-body symmetry functions can be constructed based on the cutoff function (4.4). They can be divided into two classes: *Radial* symmetry functions, describing the radial distribution of neighbours up to the cutoff radius, and *angular* symmetry functions, specifying their angular arrangement. The symmetry functions discussed in this section are all suggested by Behler [2].

Radial symmetry functions

The most basic radial symmetry function is simply the sum of the cutoff functions for all the neighbours j of atom i,

$$G_i^1 = \sum_{j=1}^{N} f_c(r_{ij}) \tag{4.5}$$

We immediately realize that this simple function meets the symmetry requirement stated above: The value of the sum is independent of the order of neighbours j. This is the case for each of the following symmetry functions, as they are
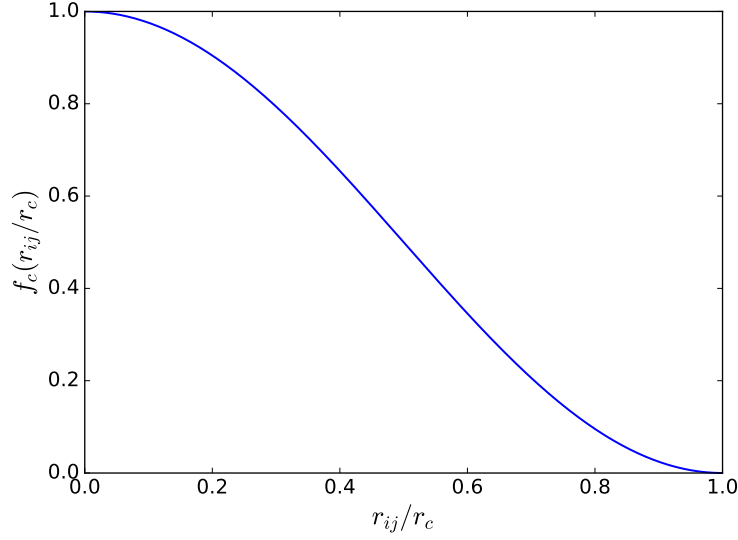
Figure 4.1: Plot of the cutoff function (4.4) applied in this thesis. This function is used to define a chemical environment around a central atom: only the atoms within the cutoff radius r_c contribute to its energy. These are called neighbouring atoms. The closer a neighbouring atom is, the larger the energy contribution, as is the case for most physical systems.

all defined as sums over neighbours. We also note that all symmetry functions, including the cutoff function, are unitless.

Physically, (4.5) can be thought of as a "coordination number" on $r_{ij} \in [0, r_c]$. A set of such functions with different cutoff radii r_c can be used to describe the radial arrangement of the neighbouring atoms by yielding coordination numbers on distance intervals of varying length. Thus, a symmetry function is defined by its parameter values, and a set of symmetry functions with different parameters form a symmetry vector **G**.

A better alternative is to use a sum of products of Gaussians and the cutoff function,

$$G_i^2 = \sum_{j=1}^{N} \exp[-\eta(r_{ij} - r_s)^2] \, f_c(r_{ij}) \tag{4.6}$$

We now have two parameters that can be adjusted to probe different radii. The width parameter η determines the radial extension of the symmetry functions, while the shifting parameter r_s displaces the Gaussians to improve the sensitivity at specific radii. A third option is

$$G_i^3 = \sum_{j=1}^{N} \cos(\kappa r_{ij}) \, f_c(r_{ij}) \tag{4.7}$$
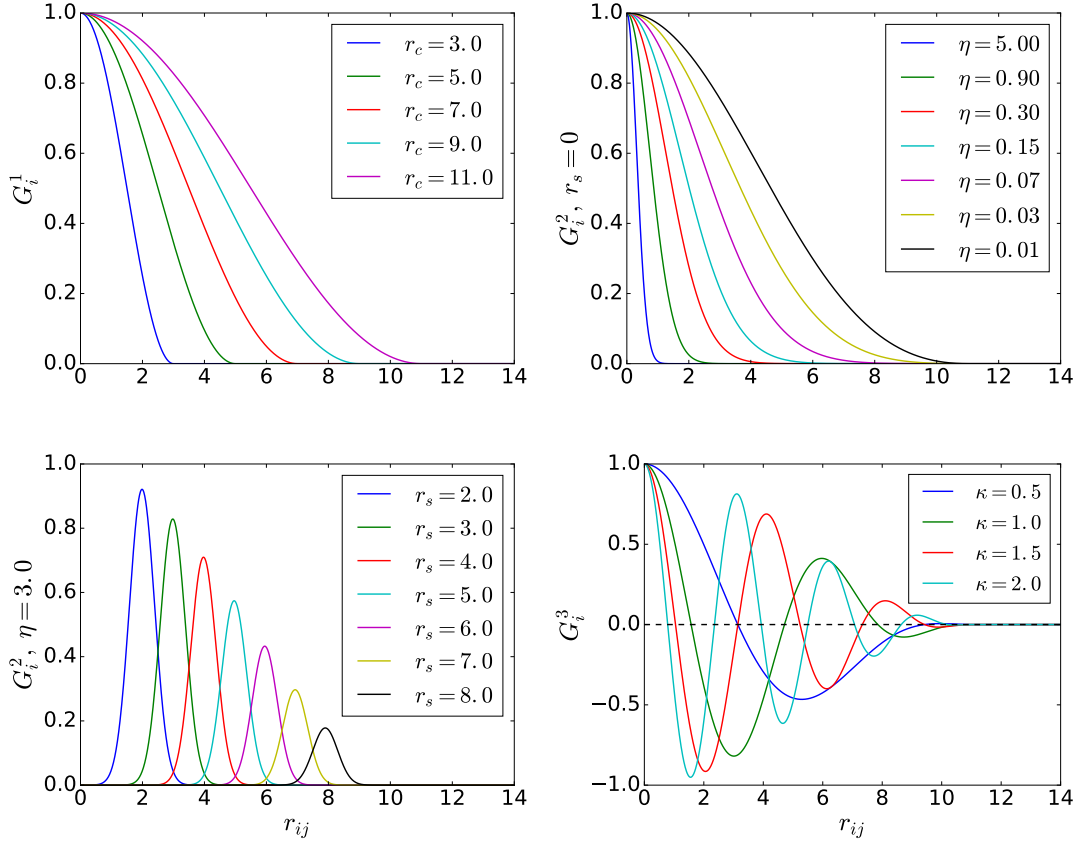
Figure 4.2: Radial symmetry functions G^1, G^2 and G^3 for an atom with one neighbour only. A set of such functions represents the radial distrubution of neighbours around a central atom placed at the origin. For G^2 and G^3 a cutoff $r_c = 11.0\,\text{Å}$ has been used. r_{ij}, r_c and r_s are in units of Å, while η and κ have units Å^{-2} and Å^{-1} respectively.

which are damped cosine functions with a period length adjusted by parameter κ. We will however not use this function because of the existence of negative function values, which can lead to atoms canceling each other's contribution to the sum.

In Figure 4.2 we see the radial symmetry functions for several different parameters. It is clear that a set of such functions can exhibit a large flexibility by tuning the parameters, enabling us to adequately represent the radial distribution of neighbours around a central atom.

Angular symmetry functions

To obtain a suitable structural fingerprint of the atomic environments, we also need the angular distribution of neighbouring atoms. This can be achieved by

using functions depending on θ_{jik}, which is the angle formed by the central atom i and the two interatomic distances r_{ij} and r_{ik}. The potential is periodic with respect to this angle, so we can use the cosine of θ_{jik} instead. We thus define an angular symmetry function as a sum over all cosines with respect to any possible triplet (i, j, k), multiplied by Gaussians of the three interatomic distances and the respective cutoff functions,

$$G_i^4 = 2^{1-\zeta} \sum_{j\neq i} \sum_{k>j} \Big[(1 + \lambda \cos\theta_{jik})^\zeta \, \exp(-\eta(r_{ij}^2 + r_{ik}^2 + r_{jk}^2))$$

$$f_c(r_{ij}) f_c(r_{ik}) f_c(r_{jk}) \Big] \tag{4.8}$$

This function is zeroed out if any of the interatomic distances is larger than r_c. The parameter η takes here into account that the angular contribution depends on the atomic seperations. The angular arrangement can be investigated by using different values for ζ while the normalization factor $2^{1-\zeta}$ ensures that the range of values is independent of the choice of ζ. The parameter $\lambda \in -1, 1$ can be used to invert the shape of the cosine function. For $\lambda = +1$ the maxima of the cosine terms are at $\theta_{jik} = 0°$, while for $\lambda = -1$ they are located at $\theta_{jik} = 180°$. The presence of the cutoff function r_{jk} ensures that only triplets where all three interatomic distances are within the cutoff radius are included. Another function can be defined that has no constraint on r_{jk},

$$G_i^5 = 2^{1-\zeta} \sum_{j\neq i} \sum_{k>j} \Big[(1 + \lambda \cos\theta_{jik})^\zeta \, \exp(-\eta(r_{ij}^2 + r_{ik}^2)) \, f_c(r_{ij}) f_c(r_{ik}) \Big] \tag{4.9}$$

which will generally lead to larger function values than (4.8) because the lack of constraint on r_{jk} results in a larger number of non-zero terms in the summation. The angular parts of G_i^4 and G_i^5 are identical,

$$G_i^\theta = 2^{1-\zeta}(1 + \lambda \cos\theta_{jik})^\zeta \tag{4.10}$$

and is shown in Figure 4.3 for different values of ζ.

Determining the symmetry parameters

The parameter values r_c, η, r_s, ζ and λ are not automatically optimized, like the weights. They need to be pre-determined, and remain fixed during the training. In the context of machine learning, such variables are called *hyperparameters*, while the variables that are optimized during training are the *model parameters*, i.e. the weights and biases. Other examples of hyperparameters are the learning rate (section 3.6) and the number of NN nodes and layers. Various techniques for deciding on satisfactory hyperparameters are discussed in section 7.3.

They symmetry function set is an integral part of the NN, and needs to be evaluated together with the NN itself when applied in simulations. The symmetry parameters must be customized for the system under cosideration; different strategies regarding this is described in section 7.2.

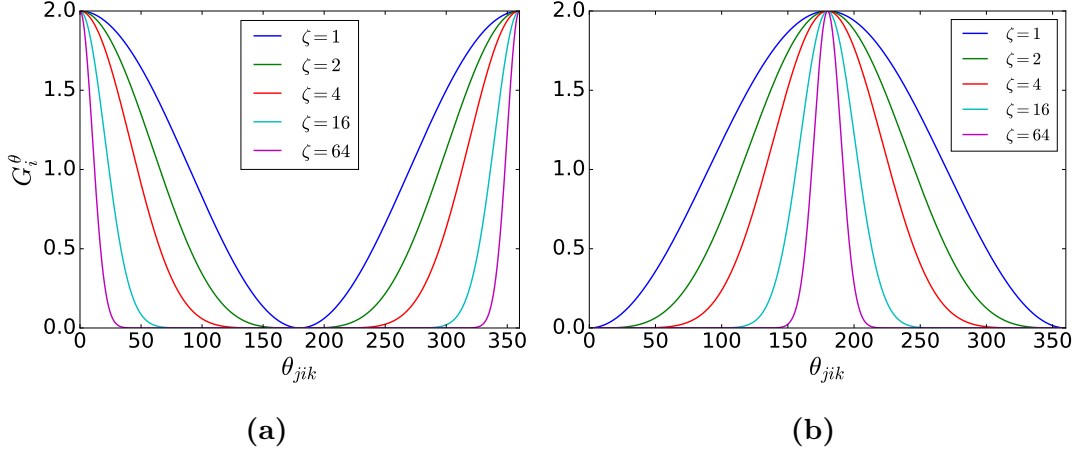**Figure 4.3:** Angular part of symmetry functions G_i^4 and G_i^5 for an atom with one neighbour only. A set of such functions represents the angular distrubution of neighbours around a central atom placed at the origin. $\lambda = +1$ for Figure 4.3a, $\lambda = -1$ for Figure 4.3b.

4.2.2   Symmetry functions and forces

To integrate Newton's laws in MD simulations we need the forces on all the atoms. The force field associated with a PES was introduced in section 2.1 and is restated here for reference:

$$\mathbf{F} = -\nabla E \tag{4.11}$$

In the present case, the total energy E of the system is defined as a sum over all the atomic energies (4.1). Further, we have applied a cutoff so that each atom only gets energy contributions from its atomic environment up to a certain distance r_c. The force component $F_{k,\alpha}$, $\alpha = (x, y, z)$ acting on atom k with respect to coordinate $r_{k,\alpha}$, can therefore be written [2]

$$F_{k,\alpha} = -\frac{\partial E}{\partial r_{k,\alpha}} = -\sum_{i=1}^{N} \frac{\partial E_i}{\partial r_{k,\alpha}} = -\sum_{i=1}^{N_k+1} \frac{\partial E_i}{\partial r_{k,\alpha}} \tag{4.12}$$

N is as before the total number of atoms, while N_k is the number of neighbours of atom k. Note that E_k is included in the above sum - atom k itself contributes to the total energy of its atomic environment. We therefore have $N_k + 1$ contributions.

NNs have as we have seen well-defined functional forms, and analytic derivatives are therefore readily available. However, when symmetry transformations is applied, there is only an indirect relation between the energy and the cartesian coordinates of the atoms. To calculate the force (4.12) we therefore need to apply

the chain rule,

$$F_{k,\alpha} = -\sum_{i=1}^{N_k+1} \frac{\partial E_i}{\partial r_{k,\alpha}} = -\sum_{i=1}^{N_k+1} \sum_{s=1}^{M_i} \frac{\partial E_i}{\partial G_{i,s}} \frac{\partial G_{i,s}}{\partial r_{k,\alpha}} \tag{4.13}$$

where M_i is the number of symmetry functions of atom i. In other words, we need to sum over all the symmetry functions of all the neighbours i of atom k and k itself. This is the reverse process of (4.3),

$$E_i \to F_{\mathrm{e}}[\mathbf{G}_{\mathrm{e}}(\mathbf{r}_{ij})] \to \mathbf{G}_{\mathrm{e}}(\mathbf{r}_{ij}) \to \mathbf{r}_{ij} \tag{4.14}$$

where the differentiation is perfomed in the direction of the arrows. We realize that the first term $\partial E_i/\partial G_{i,s}$ is the derivative of the output of the atomic NN of atom i with respect to the inputs, i.e. the gradient of the NN. This is an analytic gradient which is given by the architecture of the NN, including the values of the weights and biases. We recall from subsection 3.6.3 that the backpropagation algorithm calculates the derivatives of all neurons in a NN to obtain the amount of which each weight and bias should be adjusted during training. We can thus use a slightly modified version of this algorithm to compute the gradient of a NN,

1. Instead of backpropagating the derivative of the cost function, we backpropagate the derivative of the output neuron itself. For output neurons with the identity activation function (3.12), this derivative is 1.

2. The derivative is propagated all the way back to the input nodes. During training, the propagation stops at the weights connecting the input layer and the first hidden layer.

This procedure clarifies how a fully-connected feed-forward NN is built: To get the derivative of the output neuron w.r.t. the input neurons, we have to also compute the derivative of all the neurons in-between because all nodes in each layer are connected to all nodes in the following layer.

The second term $\partial G_{i,s}/\partial R_{k,\alpha}$ in (4.13) is given by the definition of the employed symmetry functions. The derivatives of the Behler symmetry functions in subsection 4.2.1 are listed in Appendix A.

Change of coordinates

In a MD simulation we only work with relative or inter-atomic coordinates and potentials. The training of neural networks is also based on an atom-centered approach; the cartesian coordinates that is given as input to the symmetry functions are always atom-centered. This means that the atom in question is placed at the origin and the positions of all neighbouring atoms are measured relative to
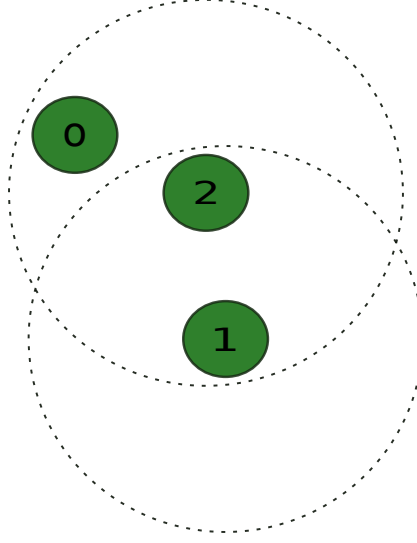
Figure 4.4: Example system of three atoms of the same type. Only atom 2 is inside the cutoff sphere of atom 1, thus we only need to take the energy of atom 1 and 2 into account to find the force on atom 1. Since the atoms are of the same type, they have identical atomic NNs and symmetry function sets.

the origin. Effectively, we are making a change of coordinates to the neighbours j of the central atom i,

$$\mathbf{r}_j \to \mathbf{r}_j - \mathbf{r}_i \tag{4.15}$$

This is in accordance with the Behler symmetry functions, which only operate with relative coordinates. It is however not trivial how this affects the above equations for the forces, which are not written in an atom-centered way. The process of obtaining the forces can be illustrated with an example system consisting of three atoms of the same element with positions $\mathbf{r}_i$ where $i = 0, 1, 2$, shown in Figure 4.4. We want to find the force on atom 1 in the x-direction. Atom 0 is outside the cutoff sphere of atom 1, so according to (4.12) the force is,

$$F_{1,x} = \frac{\partial E_1}{\partial x_1} + \frac{\partial E_2}{\partial x_1} \tag{4.16}$$

For simplicity, we assume that each atom only have two different symmetry functions describing their atomic environment; one $G_{i,0}^2$ function describing the radial arrangement of neighbours and one $G_{i,1}^5$ dealing with the angular distribution. Each atomic NN thus has two inputs and one output. Since the three atoms are all of the same type, they will have identical NNs F and symmetry vectors $\mathbf{G} = (G_{i,0}^2, G_{i,1}^5)$. We will therefore denote the the two symmetry functions simply as G_2 and G_5. In the following we apply the notation $r_{ij} = |\mathbf{r}_j - \mathbf{r}_i|$. The

expressions for E_1 and E_2 are,

$$E_1 = F\big[G_2(r_{12})\big] \tag{4.17}$$

G_5 is zero because atom 1 only has one neighbour. Further,

$$E_2 = F\big[G_2(r_{20}, r_{21}) + G_5(r_{20}, r_{21}, \theta_{021})\big] \tag{4.18}$$

We note that atom 2 receives an energy contribution from both atom 0 and 1, so that its G_2 function is a sum containing two terms. To compute the derivatives of these energies with respect to the x-coordinate of atom 1, we must take into account the change of coordinates (4.15). We have that

$$\frac{\partial G_2(x_{ij})}{\partial x_j} = \frac{\partial G_2(x_{ij})}{\partial x_{ij}}\frac{\partial x_{ij}}{\partial x_j} = \frac{\partial G_2(x_{ij})}{\partial x_{ij}}\frac{\partial (x_j - x_i)}{\partial x_j} = \frac{\partial G_2(x_{ij})}{\partial x_{ij}} \tag{4.19}$$

and

$$\frac{\partial G_2(x_{ij})}{\partial x_i} = \frac{\partial G_2(x_{ij})}{\partial x_{ij}}\frac{\partial (x_j - x_i)}{\partial x_i} = -\frac{\partial G_2(x_{ij})}{\partial x_{ij}} \tag{4.20}$$

which is also valid for the other radial symmetry functions G_1 and G_3. This symmetry, which is a manifestation of Newton's third law, enable us to calculate the derivatives in (4.16). For radial symmetry functions dependent on several atom pairs, (4.19) and (4.20) apply to each term in the sum seperately. Further, the symmetry is also present for each atom in a triplet,

$$\frac{\partial G_5(x_{ij}, x_{ik}, \theta_{jik})}{\partial x_i} = -\frac{\partial G_5(x_{ij}, x_{ik}, \theta_{jik})}{\partial x_{ij}} - \frac{\partial G_5(x_{ij}, x_{ik}, \theta_{jik})}{\partial x_{ik}} \tag{4.21}$$

and

$$\frac{\partial G_5(x_{ij}, x_{ik}, \theta_{jik})}{\partial x_j} = \frac{\partial G_5(x_{ij}, x_{ik}, \theta_{jik})}{\partial x_{ij}} \tag{4.22a}$$

$$\frac{\partial G_5(x_{ij}, x_{ik}, \theta_{jik})}{\partial x_k} = \frac{\partial G_5(x_{ij}, x_{ik}, \theta_{jik})}{\partial x_{ik}} \tag{4.22b}$$

i.e. the derivative w.r.t. i is the negative sum of the derivatives w.r.t. j and k. This is a symmetry we have exploited in our code, see section 5.4.

To obtain the derivatives of the energies (4.17) and (4.18), we apply the chain rule

$$\frac{\partial E_1}{\partial x_1} = \frac{\partial F}{\partial G_2(r_{12})}\frac{\partial G_2(r_{12})}{\partial x_1} = -\frac{\partial F}{\partial G_2(r_{12})}\frac{\partial G_2(r_{12})}{\partial x_{12}} \tag{4.23}$$

where we have applied the above change of coordinates (4.20). Using (4.19), the derivative of E_2 can be calculated correspondingly,

$$\frac{\partial E_2}{\partial x_1} = \frac{\partial F}{\partial G_2(r_{20}, r_{21})}\frac{\partial G_2(r_{20}, r_{21})}{\partial x_1} + \frac{\partial F}{\partial G_5(r_{20}, r_{21}, \theta_{021})}\frac{G_5(r_{20}, r_{21}, \theta_{021})}{\partial x_1} \tag{4.24a}$$

$$= \frac{\partial F}{\partial G_2(r_{21})}\frac{\partial G_2(r_{21})}{\partial x_{21}} + \frac{\partial F}{\partial G_5(r_{20}, r_{21}, \theta_{021})}\frac{G_5(r_{20}, r_{21}, \theta_{021})}{\partial x_{21}} \tag{4.24b}$$

From the definition of G_2 (4.6) we see that the sum contains two terms in the case of atom 2, one for each pair interaction between atom 2 and its two neighbours. The interaction term between atom 0 and 2 cancels when differentiated with respect to the coordinates of atom 1, and is thus not included in the force on atom 1.

The sum in G_5 (4.9) however, consists of only one term corresponding to the only unique triplet with atom 2 as the central atom. This term is directly dependent on the coordinates of atom 0, which means that an atom which is outside the cutoff sphere of atom 1 still contributes to the force. This is an artefact of many-body potentials like the NNPs employed in this thesis: The force acting on an atom i depends on the positions of the atoms as far as $2r_c$ away because *neighbours of neighbours* are included in the energies that are differentiated to obtain the force. This large effective range of atomic interactions is however not of great importance, taking into account that the interaction strength falls rapidly with distance for most systems where Coulomb forces are not included.

4.3   Summary

In this chapter we have introduced the Behler-Parrinello method for constructing NNPs to be used in MD simulations. This scheme employs certain symmetry functions that have the following desirable properties,

1. They are atom-centered and have an inherent cutoff r_c that defines a chemical environment of tractable size.

2. They yield the same number of outputs independent of the number of inputs.

3. The order of the inputs is irrelevant.

4. They are invariant with respect to translation and rotation.

5. They have analytic derivatives.

1. and 2. solves the technical issues of the limited and fixed dimensionality of MLPs respectively. 3. and 4. deals with physical symmetries, namely that the energy of an atomic system does not change if two identical atoms are exhanged, or if the system undergoes a translation or a rotation. 5. is a wanted property because the symmetry functions are differentiated to obtain forces.

In chapter 7 we present the complete workflow for constructing a NNP, including a description of how to assemble a symmetry function set, found in section 7.2.

Part II

Implementation and results

Chapter 5

LAMMPS

LAMMPS (Large-scale Atomic/Molecular Massively Parallel Simulator) is a classical molecular dynamics simulation package developed at Sandia National Laboratories [12]. It is written in highly portable and extendable objected-oriented C++. The package contains functionality encompassing a wide variety of potentials, atom types, thermostats and ensembles, and is able to simulate a large number of different systems, including atoms, molecules, coarse-grained particles, polymers, graunular materials and metals. LAMMPS can be run on a single processor or in parallell.

There exist several other molecular dynamics packages, like GROMACS, OpenMD, Aber and NAMD that could also have served the purpose of this thesis. We have chosen to work with LAMMPS because it is well documented[1], easy to expand upon[2], and because the Computational physics group at UiO has a lot of experience with using it for molecular dynamics simulations.

In this chapter we will outline the basic usage of LAMMPS, including a walk-through of an example input script. We also describe its class hierarchy and how we can add our own functionality to the package.

5.1 Installing LAMMPS

We have chosen to install LAMMPS by cloning the Github repository and compile the source code by running `make` in the source directory. To compile a serial version of the software, we simply run the command

```
make serial
```

while the corresponding command for the parallel MPI version is

---

[1]lammps.sandia.gov/doc/Manual.html
[2]lammps.sandia.gov/doc/Developer.pdf

```
make mpi
```

New updates can be downloaded by setting an upstream git remote to point to the LAMMPS GitHub repository and doing a `git fetch upstream`. LAMMPS also have several additional packages that can be installed. This can be done by running

```
make yes-<package name>
```

e.g. `manybody` to install many-body potentials like Stillinger-Weber and Vashishta.

5.2   LAMMPS input script

LAMMPS is run by providing an input script as argument to the excecutable. This input script is read line by line and has its own syntax. A good way to show the basics an input file is to look at a simple example. We have chosen to have a look at a script to measure the radial distribution function of a Si crystal with the Stillinger Weber (SW) potential (subsection 2.2.2),

```
# initialization
units           metal
dimension       3
boundary        p p p
atom_style      atomic

# create geometry and atoms
lattice         diamond 5.431
variable        L equal 10
region          simBox block 0 ${L} 0 ${L} 0 ${L}
create_box      1 simBox
create_atoms    1 box

# set mass and initial temperature
mass            1 28.06
variable        temp equal 300
velocity        all create ${temp} 87287

# compute radial distribution function
compute         radialDist all rdf 100

# potential
pair_style      sw
pair_coeff      * * Si.sw Si
neighbor        0.5 bin
neigh_modify    every 20 delay 0 check no
```

```
# integration
timestep          0.01
run_style         verlet
fix               integration all nve

# output
thermo            50
thermo_style      custom step temp density press ke pe etotal
thermo_modify     norm yes

# thermalize before sampling rdf
run               10000
fix radial all ave/time 100 10 1000 c_radialDist[*] file &
    rdf.txt mode vector
run               10000
```

We will in the following briefly explain what processes that are evoked in LAMMPS when the above commands are read. We start with the initialization of three basic properties of the simulation: Units, boundary conditions and particle types.

```
units       metal
```

This command defines the units that are used in the simulation and the output. LAMMPS have eight different unit sets. The `metal` set measures distance in Ångstrøm, energy in eV and temperature in Kelvin. The choice of units depends on the system that is investigated and the scale we are looking at.

```
boundary    p p p
```

We wish to measure the radial distribution function in a *bulk* Si crystal, thus we want to have periodic boundary conditions in all three dimensions, annotated by a `p`. LAMMPS can also handle stiff `f` and adaptive `s` non-periodic boundaries. Adaptive means that the position of the face is set so as to encompass the atoms in that dimension.

```
atom_style atomic
```

Different systems need different information to be stored for each atom. For style `atomic`, only the default attributes are associated with each atom, namely coordinates, velocities, atom IDs and types. This is sufficient for non-bonded potentials like Stillinger-Weber.
Next, we create the geometry of the simulation domain.

```
lattice     diamond 5.431
```

This defines the initial configuration of the atoms. Here, a diamond cubic lattice with a lattice constant of 5.431 is used, which is the crystal structure of Si. Other types of lattices based on cubic or square unit cells are also available.

```
variable      L equal 10
region        simBox block 0 ${L} 0 ${L} 0 ${L}
```

The geometry of the system is defined with the `region` command. The `block` style is simply a 3-dimensional straight-faced box with a size of $L = 10$ unit cells in each dimension. We have labelled the region `simBox`. The number of unit cells `L` is defined as a LAMMPS `variable`. Variables can be referenced elsewhere in the script by writing `${variable_name}` to become part of a new input command like above. LAMMPS enables many styles of variables to be defined.

```
create_box      1 simBox
create_atoms    1 box
```

The command `create_box` creates a simulation box based on the specified region, in our case the `block` region defined above. The argument specifies the number of atom types that will be used in the simulation. Next, `create_atoms` with the argument `box` fills the simulation domain with atoms of type 1 on the lattice. LAMMPS also lets you create a random collection of atoms or single atoms at specified coordinates.

```
mass            1 28.06
variable        temp equal 300
velocity        all create ${temp} 87287
```

We need to assign mass to the atoms. For `metal` units, mass is measured in grams/mole. The atoms are also given an initial velocity that corresponds to the given initial temperature.

```
pair_style      sw
pair_coeff      * * Si.sw Si
```

The choice of potential is made with the `pair_style` command. We want to simulate interactions using the Stillinger-Weber potential `sw`. The SW parameters are set with `pair_coeff`, where `Si.sw` is a text file from which all parameters are read. Further, asterisks signifies that the potential file contains parameters for all elements, in this case for Si only.

```
neighbor        0.5 bin
neigh_modify    every 20 check yes
```

These commands sets parameters that affect the building of neighbor lists. The first argument to `neighbor` is the skin size, while the second selects what algorithm is used to build the lists. The `bin` style creates the lists by binning, which in most cases (including ours) is the fastest method. Further, we can control how often the lists are built with `neigh_modify`. The above arguments specifies that new neighbour lists are built every 20 steps, but every step LAMMPS checks if any atom has moved more than half the skin distance. If that is the case, new lists are built.

```
timestep          0.01
run_style         verlet
fix               integration  all  nve
```

LAMMPS integrates Newtons' equations of motion with the velocity-Verlet algorithm by defualt. This is the integrator of choice for most MD applications due to its simplicity and symplectic nature (section 2.3). The rRESPA integrator [11] scheme is also available. LAMMPS does however not integrate and update the positions and velocities of the particles if not explicitly told so. This is done with a `fix`, which is any operation that is applied to the system during timestepping. The above `fix` tells LAMMPS to integrate all atoms in the system so that they follow trajectories consistent with the microcanonical ensemble.

```
thermo            50
thermo_style      custom  step  temp  press  ke  pe  etotal
thermo_modify     norm  yes
```

We can control what thermodynamic properties to calculate and output with `thermo_style`, while `thermo` decides how often they should be computed. We want to output the time step, temperature, pressure, kinetic energy, potential energy and total energy. Also, we want to normalize the extensive quantities (the energies) by the number of atoms.

```
compute           radialDist  all  rdf  100
fix  radial  all  ave/time  100  10  1000  c_radialDist[*]  file  &
     rdf.txt  mode  vector
```

The radial distribution function are measured with a LAMMPS `compute`, which defines a computation that is perfomed on a group of atoms, in this case `all` atoms. For the `compute` to actually be executed, it needs to be evoked by other LAMMPS commands, in this case the fix `ave/time`. The three numerical arguments specify how the distribution is averaged over time. Note that the `compute` is referenced by `c_` followed by its label.

```
run        5000
```

Lastly, we run the simulation for 5000 timesteps. LAMMPS allowes several run commands to be issued after one another. This comes in handy if we want to thermalize the system before measuring thermodynamic quantities,

```
run     5000
fix radial all ave/time 100 10 1000 c_radialDist[*] file &
    rdf.txt mode vector
run     10000
```

When parsing the second run command, LAMMPS will continue the dynamics while computing eventual new fixes, computes, dumps etc. defined since the last run command.

This section has illuminated some of the basic funcionality of LAMMPS through several comments on a simple input script. Now we move on to how class hierarchy of LAMMPS is structured.

5.3   LAMMPS structure

LAMMPS is written in C++ in an object-oriented fashion. The class hierarchy is shown schematically in Figure 5.1. The top-level classes (blue) are visible anywhere in LAMMPS. Some selected sub-classes (green and yellow) are displayed. We recognize many of the input script commands among these sub-classes; a rule of thumb is that every input script command has a corresponding class and a corresponding file name in the source directory. For instance, we observe that the Stillinger-Weber potential "PairSW" is a sub-class of Pair. This potential was evoked by the command `pair_style sw` above and the source files are named `pair_sw.cpp` and `pair_sw.h`.

We will not go into detail about the functionality and communication of these classes. The one that is the most relevant for this work is the `Pair` class. This is the parent class of non-bonded potentials, which in LAMMPS includes many-body potentials like Stillinger-Weber and Vashishta. The NN potential in this thesis falls under this category, and thus inherits from `Pair`. In the following section we will describe how LAMMPS can be extended with such a new `pair_style`.

5.4   Extending LAMMPS

To extend LAMMPS with a new component, we simply add a new `*.cpp` and `*.h` file to the source directory and re-compile. The new class will (in theory) work with all other LAMMPS classes without breaking any functionality. A good strategy is to start with an existing potential source file that is somewhat similar to the one we want to make, instead of programming from scratch. We aim to
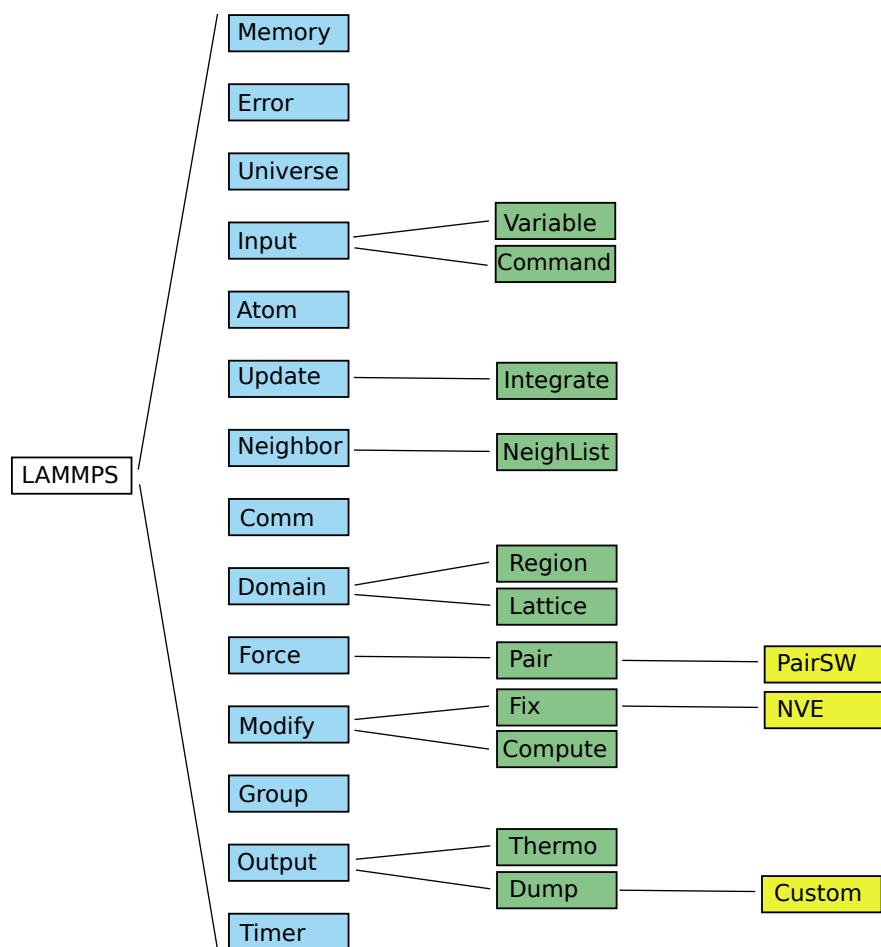
Figure 5.1: Class hierarchy of LAMMPS. The inheritance is from left to right. The top-level classes (blue) are visible everywhere in the code. Only a selected set of sub-classes (green and yellow) is displayed, where we recognize many of the input script commands.

construct a many-body non-bonded neural network potential (NNP), and have chosen to base our implementation on the Stillinger-Weber source files.

5.4.1   Potential initialization

The new potential class is added to LAMMPS with the following function call in the header file,

```
PairStyle(nn/manybody,PairNNManyBody)
```

The first argument is the name the potential will have in the input script, while the second is the name of the class. To use our new NN potential in a simulation we simply write,

```
pair_style       nn/manybody
```

in the input script. This command evokes two functions in the source files that need to be implemented: `settings` and `init_style`. The former controls the number of arguments to `pair_style`, which in our case is zero,

```cpp
void PairNNManyBody::settings(int narg, char **arg) {
  if (narg != 0) error->all(FLERR,"Illegal pair_style command");
}
```

while the latter decides what kind of neighbour lists LAMMPS should provide to the potential,

```cpp
void PairNNAngular2::init_style() {
  // ...
  if (force->newton_pair == 0)
    error->all(FLERR,"Pair style NN requires newton pair on");

  // need a full neighbor list
  int irequest = neighbor->request(this);
  neighbor->requests[irequest]->half = 0;
  neighbor->requests[irequest]->full = 1;
}
```

Our NN potential needs full neighbour lists to be able to compute three-body terms. Further, we ensure that `newton_pair` is set to 1, which tells LAMMPS to communicate forces attributed to ghost atoms to the original atoms.

The parent class `Pair` has three *pure virtual member functions*, i.e. functions that are required to be implemented by a derived class: `compute`, `coeff` and the aforementioned `settings`. The `compute` method is where the forces and energies are calculated. This function is called once each time step by the Verlet algorithm and the resulting forces are used to integrate Newton's equations of motion.

The coefficients of the employed potential are set in `coeff`, which in the case of a NN potential is the weights, biases, symmetry function parameters and other hyperparameters. These are read from file, and based on information on the architecture of the NN, the network is reconstructed as a set of matrices and vectors, according to the description in subsection 3.3.2. This is achieved with the input script command,

```
pair_coeff      * * <path−to−saved−NN> <cutoff>
```

where the wildcard asterisks signify that the coefficients span all atom types and are read from file. If there are more than one element, a seperate NN file is read for each one. An example is SiO_2,

```
pair_coeff      * * <path−1> <path−2>  Si  O <cutoff>
```

5.4.2   Computing energies and forces

Now that the NNP is set up, we are ready to employ it in simulations. The process of computing energies and forces with the NNP is described as follows,

1. Initialize various variables like the vector of atom coordinates and the neighbour lists of each atom.

2. Loop through all atoms. For each atom i:

   (a) Construct the input vector to the NN by evaluating the symmetry functions for all atomic pairs and triplets below the cutoff radius. The coordinates, types, angles etc. of all pairs and triplets must be stored for later force calculations.

   (b) Evaluate the NN corresponding to the element of atom i to obtain its total potential energy.

   (c) Calculate the gradient of the energy w.r.t. the symmetry function values by backpropagation.

   (d) Differentiate the symmetry functions w.r.t. to the coordinates of all neighbours to obtain the total force on atom i.

This process determines the structure of the `nn/manybody compute` function. We start out by accessing various variables needed to compute the forces,

```
double **x = atom->x;       // atom coordinates
double **f = atom->f;       // forces
int *type = atom->type;     // type of each atom
int nlocal = atom->nlocal;  // no. of atoms belonging to current processor
```

```
int inum = list->inum;      // length of list containing all neighbour lists
int *ilist = list->ilist;   // local indicies of all local atoms
int *numneigh = list->numneigh; // lengths of each neighbour list
int **firstneigh = list->firstneigh; // all neighbour lists
```

Next, we loop through all atoms in `ilist`. For each atom, we evaluate the two-body and three-body symmetry functions with the symmetry parameters read from file. We also store certain information on all neighbours.

```
for (int ii = 0; ii < inum; ii++) {
  int i = ilist[ii];
  int itype = map[type[i]];

  double xtmp = x[i][0];
  double ytmp = x[i][1];
  double ztmp = x[i][2];

  // neighbour list of atom i
  int *jlist = firstneigh[i];
  int jnum = numneigh[i];

  // collect all pairs
  for (int jj = 0; jj < jnum; jj++) {
    int j = jlist[jj];
    j &= NEIGHMASK;
    tagint jtag = tag[j];
    int jtype = map[type[j]];

    double delxj = x[j][0] - xtmp;
    double delyj = x[j][1] - ytmp;
    double delzj = x[j][2] - ztmp;
    double rsq1 = delxj*delxj + delyj*delyj + delzj*delzj;

    if (rsq1 >= cutoffSquared) continue;
    double rij = sqrt(rsq1);

    // store relative coordinates (delxj, delyj, delzj), distance rij,
    // index j and type jtype. Initialize input vector.

    // apply two-body symmetry
    for (int s=0; s < m_numberOfSymmFunc; s++)
      if ( m_parameters[s].size() == 3 )
          inputVector(0,s) += G2(rij, m_parameters[s][0],
                               m_parameters[s][1], m_parameters[s][2]);

    // collect all triplets
    for (int kk = jj+1; kk < jnum; kk++) {
      // apply three-body symmetry and
      // store configurations of all triplets...
```

We are now ready to evalute the NN with the input vector of symmetry values.

The NN is implemented as a function, shown here in the monatomic case for brevity,

```cpp
double PairNNManyBody::network(arma::mat inputVector) {

  // no activation for input layer
  m_preActivations[0] = inputVector;
  m_activations[0] = m_preActivations[0];

  // hidden layers
  for (int i=0; i < nLayers; i++) {
    m_preActivations[i+1] = m_activations[i]*m_weights[i] + m_biases[i];
    m_activations[i+1] = sigmoid(m_preActivations[i+1]);
  }

  // identity activation for output layer
  m_preActivations[nLayers+1] = m_activations[nLayers]*m_weights[nLayers] +
                                m_biases[nLayers];
  m_activations[nLayers+1] = m_preActivations[nLayers+1];

  // return activation of single output neuron
  return m_activations[nLayers+1](0,0);
}
```

where `sigmoid(A)` is a function that returns the sigmoid (3.15) of each element of a matrix `A`. Each variable with the `m_` prefix is an `std::vector` containing a set of Armadillo matrices. For instance, `m_weights[0]` is a matrix of all weights connecting the input layer and the first hidden layer. This results in highly readable code that is very similar to the mathematical description (3.60) [3],

$$\mathbf{u}_l = W_l \mathbf{y}_{l-1} + \mathbf{b}_l$$
$$\mathbf{y}_l = f_l(\mathbf{u}_l) \tag{5.1}$$

The NN is differentiated by employing backpropagation according to the procedure oulined in subsection 4.2.2,

```cpp
arma::mat PairNNManyBody::backPropagation() {

  // derivative of output neuron is 1
  arma::mat output(1,1); output.fill(1);
  m_derivatives[nLayers+1] = output;

  // we can thus compute the error vectors for the other layers
  for (int i=nLayers; i > 0; i--) {
      m_derivatives[i] = ( m_derivatives[i+1]*m_weightsTransposed[i] ) %
```

---

[3]The weight-activation multiplication is in reverse order in the code because we use row vectors instead of column vectors. Also, the indexing of the weight and bias vectors are different than in (5.1), as the activation and preactivation vectors also include the input layer. This also applies to the implementation of backpropagation.

```
            sigmoidDerivative(m_preActivations[i]);
    }

    // no activation function for input layer
    m_derivatives[0] = m_derivatives[1] * m_weightsTransposed[0];

    return m_derivatives[0];
```

where % signifies element-wise multiplication ⊙, and `m_derivatives[i]` is the error vector $\boldsymbol{\delta}$ of layer i. We compare to (3.62)

$$\boldsymbol{\delta}_l = \mathbf{y}'_l \odot (\mathrm{W}^T_{l+1} \boldsymbol{\delta}_{l+1}). \tag{5.2}$$

Consequentially, the output and gradient of the NN is obtained in a compact way,

```
evdwl = network(inputVector);
arma::mat dEdG = backPropagation();
```

We remember from subsection 4.2.2 that the force is the product of the gradient of the NN and the symmetry functions (4.13),

$$F_{i,\alpha} = -\sum_{j=1}^{N_i+1} \frac{\partial E_j}{\partial R_{i,\alpha}} = -\sum_{j=1}^{N_j+1} \sum_{s=1}^{M_j} \frac{\partial E_j}{\partial G_{j,s}} \frac{\partial G_{j,s}}{\partial R_{i,\alpha}} \tag{5.3}$$

Here the NN potential differs from ordinary many-body potentials, which compute forces from each neighbour on-the-fly: The potential energy of atom i can only be calculated w.r.t all neighbours *simultaneously*. We therefore have to loop over the neighbours once more to calculate the forces. By storing the configurations of all pairs and triplets during energy computation above, we avoid recalculations. We show an excerpt of the code that implements (5.3). First, for G_i^2 (4.6),

```
for (int s=0; s < m_numberOfSymmFunc; s++) {
  if ( m_parameters[s].size() == 3 ) {
    arma::mat dG2(1,neighbours);

    // calculate cerivative of G2 for all pairs
    dG2dR(Rij, m_parameters[s][0],
          m_parameters[s][1], m_parameters[s][2], dG2);

    // chain rule: all pair forces
    arma::mat fpairs = -dEdG(0,s) * dG2 / Rij;

    // loop through all pairs
    for (int l=0; l < neighbours; l++) {
      double fpair = fpairs(0,l);

      // force on atom i
```

```
      fx2 -= fpair*drij(0,l);
      fy2 -= fpair*drij(1,l);
      fz2 -= fpair*drij(2,l);

      // force on atom j
      f[tagsj[l]][0] += fpair*drij(0,l);
      f[tagsj[l]][1] += fpair*drij(1,l);
      f[tagsj[l]][2] += fpair*drij(2,l);
    }
  }
```

We observe that the outer loop is over symmetry function values, while the inner loop is over all atom pairs (i, j). The vectors Rij and drij contain the distances to and relative coordinates of all neighbours of atom i. Note that we assign each pair force $\mathbf{F}_{ij}$ both to atom i and neighbour j, according to the symmetry considerations (4.19, 4.20) in subsection 4.2.2.

The derivatives of G_i^5 are also computed inside the outer loop over symmetry function values,

```
  // G5
  else {
    for (int l=0; l < neighbours-1; l++) {
      int numberOfTriplets = arma::size(Riks[l])(1);

      // calculate forces for all triplets (i,j,k) for this pair (i,j)
      double fj3[3], fk3[3], dGj[3], dGk[3];
      for (int m=0; m < numberOfTriplets; m++) {

        // calculate derivatives dGj and dGk of G5 w.r.t.
        // coordinates of atom j and k respectively...

        // chain rule: force atoms i-j
        fj3[0] = -dEdG(0,s) * dGj[0];
        fj3[1] = -dEdG(0,s) * dGj[1];
        fj3[2] = -dEdG(0,s) * dGj[2];

        // chain rule: force atoms i-k
        fk3[0] = -dEdG(0,s) * dGk[0];
        fk3[1] = -dEdG(0,s) * dGk[1];
        fk3[2] = -dEdG(0,s) * dGk[2];

        // force on atom i
        fix -= fj3[0] + fk3[0];
        fiy -= fj3[1] + fk3[1];
        fiz -= fj3[2] + fk3[2];

        // force on atom j
        f[tagsj[l]][0] += fj3[0];
        f[tagsj[l]][1] += fj3[1];
```

```
        f[tagsj[l]][2] += fj3[2];

        // force on atom k
        f[tagsk[l][m]][0] += fk3[0];
        f[tagsk[l][m]][1] += fk3[1];
        f[tagsk[l][m]][2] += fk3[2];

        // compute virial...
```

We exploit the symmetries (4.21, 4.22a, 4.22b) in subsection 4.2.2 also here, by assigning the negative of the sum of the computed forces `fj3` and `fk3` to atom i, while giving the negative of these forces to atom j and k seperately.

In fact, this is a requirement for our implementation to work: (5.3) states that the force on atom i is the derivative of the energies of all neighbours and i itself. This also applies to all neighbours j of atom i: Each neighbour experiences a force contribution $\partial E_i / \partial R_{j,\alpha}$ from atom i. To avoid having to store all atomic energies E_i, we therefore take care of this contribution when E_i is available. The alternative is to compute and store the energies and the environments of all atoms before calculating the forces, but this is tedious and requires too much storage for large systems.

5.5   Summary

In this chapter we have illustrated some of the basic usage of LAMMPS by explaining the different commands of an example input script. Further, we have briefly discussed the class structure of LAMMPS, and how to extend this structure with a custom potential.

The example code from the NNP class is an implementation of different theoretical aspects introduced in chapter 2, 3, and 4. We showed how a NN can be evaluated as a series of matrix-vector operations (subsection 3.3.2), how to differentiate a NN with the backpropagation algorithm (subsection 3.6.3), and how to obtain the forces when employing the Behler-Parrinello method (subsection 4.2.2).

We will now continue with an introduction on how to train NNs with the machine learning library TensorFlow.

Chapter 6

TensorFlow

TensorFlow [13, 14] is an open source software library developed by Google for numerical computation using data flow graphs (DFG). A DFG is a graphical representation of how data is advanced through a system, including what kind of information is input to and output from the system and where data is stored. The data flow is represented as a directed graph with a number of *nodes* that is connected by *edges.* In TensorFlow, nodes represent mathematical operations, while the edges represent the multidimensional data arrays (tensors) communicated between them.

A simple example of a DFG is displayed in Figure 6.1, which is a visualization of the flow of data for the (real) solutions

$$x_1 = \frac{-b + \sqrt{b^2 - 4ac}}{2a} \tag{6.1a}$$

$$x_2 = \frac{-b - \sqrt{b^2 - 4ac}}{2a} \tag{6.1b}$$

of a quadratic equation

$$ax^2 + bx + c = 0 \tag{6.2}$$

We observe that each node (green circles) contain the mathematical operations needed to compute (6.1a, 6.1b), while the arrows (edges) show what data each node recieve and where the output is sent. The inputs and outputs of the graph is marked with rectangular boxes.

TensorFlow was originally developed by the Google Brain Team[1] with the purpose of conducting machine learning and deep neural networks research, but is applicable in other domains as well, e.g. for solving partial differential equations. The package incorporates a wide variety of machine learning algorithms, including convolutional neural networks and recurrent neural networks. A computation expressed using TensorFlow can be executed on a wide variety of devices and systems, from mobile phones to large-scale distributed systems and GPUs.

---

[1]research.google.com

Figure 6.1: Data flow graph for the solutions (6.1a, 6.1b) of a quadratic equation (6.2). The nodes (green circles) are mathematical operations connected by edges (arrows) that show the data flow direction. Inputs and outputs are marked as rectangular boxes. Reproduced from: web.cecs.pdx.edu

In this chapter we will go through a simple example on how to build a computational graph with TensorFlow, and how to train a feed-forward neural network. We begin with a brief discussion on how to install the software.

6.1   Installing TensorFlow

The easiest way to install TensorFlow on Ubuntu[2] is to use pip[3], a package management system for software written in Python, installed with the command

```
$ sudo apt-get install python-pip python-dev
```

TensorFlow can be installed with or without GPU support. The GPU version requires a system with CUDA-enabled NVIDIA GPUs [4]. We have used the CPU version for the development of our code, which is installed with the command

```
$ pip install tensorflow
```

A simple

---

[2]www.tensorflow.org/install/
[3]pip.pypa.io/en/
[4]https://developer.nvidia.com/cuda-gpus

```
$ python -c "import tensorflow"
```

will test if the installation was succesful.

6.2 TensorFlow basic usage

TensorFlow relies on highly-optimized C++ for its computations, but offers APIs [5] in both Python, C++ and a few other languages. The Python API is at present the most complete, the easiest to use and the best documented. We have therefore chosen to use Python to train neural networks. For a discussion on ways to import a trained NN to C++ see section 5.4 and **??**.

6.2.1 Hello world

As already mentioned, a TensorFlow computation is described by a graph which consists of nodes that communicate with multi-dimensional arrays along inter-connected edges. Each node has zero or more outputs, and represents the instantation of an *operation*. Each operation occurs within a so-called *session*. A Hello world! program can be written as

```python
import tensorflow as tf

hello = tf.constant('Hello world!')
sess = tf.Session()
print sess.run(hello)
# --> Hello world!
```

This simple example outlines the general workflow for building a TensorFlow graph,

1. Declare nodes (operations)

2. Initiate a session

3. Run operations within session

The only operation in the above example is the built-in TF constant that holds the Hello world! string, and the corresponding DFG only consists of one node. In more involved examples, visualizing the DFG can be useful for finding out how the script works. This is done with the included TF visualization tool TensorBoard. We will now look at how to construct a neural network with TF and visualize the corresponding DFG.

---

[5]https://www.tensorflow.org/api_docs/

6.2.2   Creating a neural network

In subsection 3.3.2 we demonstrated that a NN can be constructed as a set of matrix-vector multiplications and additions which are input to certain activation functions. The equations for evaluting a NN were (3.60),

$$\mathbf{u}_l = W_l \mathbf{y}_{l-1} + \mathbf{b}_l \tag{6.3a}$$

$$\mathbf{y}_l = f_l(\mathbf{u}_l) \tag{6.3b}$$

We now want to construct a NN with 3 inputs neurons, one hidden layer with 10 neurons and 1 output neuron. We use the sigmoid (3.15) activation function for the hidden layer, while the output layer have the identity activation function (3.12). This can be achieved with the following Python script,

```python
import tensorflow as tf

# NN parameters
inputs  = 3
nodes   = 10
outputs = 1

# declare input tensor
inputLayer = tf.placeholder( tf.float32, [None,inputs] )

# weights input layer -> hidden layer and biases hidden layer
W1 = tf.Variable( tf.random_normal([inputs,nodes]) )
b1 = tf.Variable( tf.constant(0.1, shape=[nodes]) )

# activation hidden layer
preAct1 = tf.add( tf.matmul(inputLayer, W1), b1 )
act1 = tf.nn.sigmoid(preAct1)

# weights hidden layer -> output layer and bias output layer
W2 = tf.Variable( tf.random_normal([nodes,outputs]) )
b2 = tf.Variable( tf.constant(0.1, shape=[outputs]) )

# activation output layer
preActOutput = tf.add( tf.matmul(act1, W2), b2 )
activationOutput = preActOutput
```

There are several aspects of the TensorFlow Python API to comment on here. We observe that the input layer is declared as a TF `placeholder`. This is a special type of TF tensor that is not initialized with any values, but is rather fed with data on its execution. Specifying the first component of its shape as `None` signifies that we can supply an arbitrary number of data points simultaneously. These properties enables the same NN to be evaluated for different data vectors of varying size on-the-fly during training, explained in more detail below.

The weights and biases are declared as TF `Variables`. These are the trainable parameters of the computational graph, and are normally constructed as random

numbers. In the present case, the weights are normally distributed numbers while the biases are all set to the same value 0.1. We note that the shapes of the weight matrices and bias vectors are in accordance with the description in subsection 3.3.2.

The activations (outputs) of each layer are calculated according to (6.3a) and (6.3b). Together with the input `placeholder`, these are of type `tf.Tensor`, i.e. nodes in the computational graph.

6.2.3   Visualizing the graph

We can use TensorBoard to visualize the DFG of the above script. The DFG for the present example is quite simple, but for more involved models it quickly becomes difficult to keep track of all the different components. A convenient tool to order and group the DFG is *name scoping*. Both `Tensors` and `Variables` can be name scoped and TensorBoard employs this information to define a hierarchy of the nodes in the graph. An example of name scoping is

```python
with tf.name_scope('layer1'):
    W1 = tf.Variable( tf.random_normal([inputs,nodes]), name='weights')
    b1 = tf.Variable( tf.constant(0.1, shape=[nodes]), name='biases')
```

which results in the following names,

- `layer1/weights`

- `layer1/biases`

Additionally, TensorBoard will group the two variables together under `layer1`. To save the graph to file we have to start a TF session. The following script deploys name scoping, initiates a session and writes the DFG to file,

```python
import tensorflow as tf

# NN parameters
inputs = 3
nodes = 10
outputs = 1

with tf.name_scope('inputLayer'):
    inputData = tf.placeholder(tf.float32, [None,inputs], name='inputData')

with tf.name_scope('layer1'):

    W1 = tf.Variable( tf.random_normal([inputs,nodes]), name='weights' )
    b1 = tf.Variable( tf.constant(0.1, shape=[nodes]), name='biases' )

    preAct1 = tf.add( tf.matmul(inputData, W1), b1, name='preActivation' )
    act1 = tf.nn.sigmoid(preAct1, name='activation')
```

```python
with tf.name_scope('outputLayer'):

    W2 = tf.Variable( tf.random_normal([nodes,outputs]), name='weights' )
    b2 = tf.Variable( tf.constant(0.1, shape=[outputs]), name='biases' )

    preActOutput = tf.add( tf.matmul(act1, W2), b2, name='preActivation' )
    actOutput = tf.identity(preActOutput, name='activation')

# initialization node
initOperation = tf.global_variables_initializer()

# start session
with tf.Session() as sess:

    # initialize all variables
    sess.run(initOperation)

    # write summaries
    tf.summary.FileWriter('Summaries', sess.graph)
```

There are two new TF function calls to comment on in this script. First, we have
the line

```python
initOperation = tf.global_variables_initializer()
```

Constants are initalized when we call `tf.constant`, and their value can never
change. By contrast, variables need to be initialized using a seperate operation
within a session. The function `tf.global_variables_initializer` adds such
an operation to the graph, which initializes all global variables when run with
`sess.run(initOperation)`.

Secondly, the graph is written to file for visualization with the function
`tf.summary.FileWriter`. The first argument is the location of the files, while
the second is the graph that is launched in the present session.

The resulting graph is displayed in Figure 6.2. By default, only the top of the
hierarchy defined by name scoping is shown. In our case, this is the three NN
layers. In this figure, the output layer is expanded to reveal all the nodes in this
scope. The nodes are marked by ellipses, while the rectangles are namespaces
defining another level in the hierarchy, each containing more nodes. The edges
show how the nodes are connected[6].

---

[6]Each TF variable and constant are actually made up of several nodes, including *assign* and
read operations. See TF API for more information.

Figure 6.2: Example of a computational graph in TensorBoard of a NN consisting of an input layer, one hidden layer and an output layer. Nodes are marked by ellipses, the rectangles are namespaces containing several nodes.

Next, we will demonstrate how to train a NN with TensorFlow.

6.2.4 Training a NN with TensorFlow

To train the NN described in the previous section, we need two additional components: a cost function and an optimizer. The mean-square-error cost function (3.19) is available as a built-in function in TF, `tf.nn.l2_loss`. To compute the error, we need to supply known target values that is fed into a new `placeholder`,

```
outputData = tf.placeholder(tf.float32, [None,outputs], name='outputData')
```

The cost function is now defined as

```
cost = tf.nn.l2_loss( tf.subtract(actOutput, outputData), name='cost')
```

This adds an operation to the graph that calculates the mean-square-error between the NN prediction `actOutput` on the data set and the known target values `outputData`. In this work, the input data is configurations of atoms, while the target values are the potential energies of these configurations.

A number of gradient descent optimizers are implemented in TensorFlow, including vanilla gradient descent (3.20), Momentum [15], Adagrad [16], Adadelta [17] and Adam [18]. To add the Adam optimizer operation to the graph, we simply write

```python
optimizer = tf.train.AdamOptimizer()
trainStep = optimizer.minimize(trainCost)
```

To sum up, we have added the following to the script in the previous section (now with name scoping),

```python
outputData = tf.placeholder(tf.float32, [None,outputs], name='outputData')

# cost function
with tf.name_scope('cost'):
    error = tf.subtract(actOutput, outputData, name='deviation')
    cost = tf.nn.l2_loss(error, name='L2norm')

# optimizer
with tf.name_scope('optimizer'):
    optimizer = tf.train.AdamOptimizer(learning_rate=0.001, name='Adam')
    trainStep = optimizer.minimize(trainCost, name='trainStep')
```

The resultant graph can be seen in Figure 6.3. The optimizer namespace is expanded to display its contents. We observe that the optimizer is connected directly to all other parts of the graph, enabling it to compute the derivative of the cost function with respect to all the weights and biases. This task is performed by the `gradients` operation, that in turn sends the total gradient to `trainStep`, which implements the Adam algorithm for updating the NN parameters.

We are now ready to launch this graph in a training session. We assume that input data and target values are available as `NumPy` arrays `xTrain` and `yTrain` respectively. The code example implements the different variants of gradient descent described in subsection 3.6.1.

```python
for epoch in xrange(numberOfEpochs+1):

    # for shuffling the data set
    indicies = np.random.choice(trainSize, trainSize, replace=False)

    # offline learning / batch gradient descent
    if batchSize == trainSize:

        # pick whole set in random order
        xBatch = xTrain[indicies]
```

```
        yBatch = yTrain[indicies]

        # train
        sess.run(trainStep,
                 feed_dict={inputLayer: xBatch, outputData: yBatch})

    # online learning
    else:
        # loop through whole set, train each iteration
        for b in xrange(numberOfBatches):
            batch = indicies[b*batchSize:(b+1)*batchSize]
            xBatch = xTrain[batch]
            yBatch = yTrain[batch]

            # train
            sess.run(trainStep,
                     feed_dict={inputLayer: xBatch, outputData: yBatch})

    trainError = sess.run(cost,
                          feed_dict={inputLayer: xBatch, outputData: yBatch})
```

We observe that the training step is performed with the single line,

```
sess.run(trainStep, feed_dict={inputLayer: xBatch, outputData: yBatch})
```

which runs the `trainStep` operation defined above. This operation initiates a series of calculations that is performed automatically by TensorFlow. First, the NN is evaluated on the training batch `xBatch` and the output is compared to the corresponding batch of target values `yBatch`. The training batch is fed to the NN through the placeholder `inputLayer`, while `outputData` receives the target batch. Next, the cost function is calculated, and the error is backpropagated through the DFG to obtain its gradient. Finally, the Adam update rule is applied and the weights and biases adjusted.

Figure 6.3: Example of a computational graph in TensorBoard for training a NN consisting of an input layer, one hidden layer and an output layer. Nodes are marked by ellipses, the rectangles are namespaces containing several nodes. A cost function describing the error of the NN fit and an optimizer that minimizes this error is included in the graph. The optimizer namespace is expanded to display its contents: `gradients` computes the gradient of all nodes in the graph, `trainStep` implements the algorithm for updating the NN parameters, while `beta1_power` and `beta2_power` are the parameters of the Adam optimizer algorithm.

Chapter 7

Constructing a neural network potential

The general workflow for constructing and applying a high-dimensional neural network potential (NNP) is as follows:

1. Generate training data relevant for the intended application of the NNP.

2. Train a NN to fit the data, employing a suitable set of hyperparameters.

3. Use the analytical expression for the eventual NNP as a potential in MD simulations.

In this chapter we will go into details regarding these three steps. We describe different ways of assembling an adequate reference set for the NN, which is the most crucial factor for the performance of the final potential. Further, we discuss how suitable hyperparameters can be determined, including the symmetry function parameters. Lastly, we compare the performance of two different methods for evaluating a trained NN in simulations.

7.1 Selecting the reference set

The number of possible configurations of atomic or molecular systems quickly grows very large. Consider for instance water H_2O, which is a three-body molecule. The configuration of the three atoms can be described by three internal coordinates, each with a certain range of values. If each range is divided into N equal segments, the total number of distint configurations is N^3. If we were to construct an *ab inito* potential that characterizes this system, we would have to compute the energies and forces of all these configurations. Such a brute-force approach may be feasible for a three-atom molecule, but for more complicated systems it quickly becomes practically impossible due to the computational cost.

In this work we do not use *ab inito* methods to compute energies and forces, but this problem is still highly relevant due to the high-dimensional nature of our approach. The average number of neighbours for the Stillinger-Weber Si simulations in this thesis is about $n = 6$ (chapter 9), which is a small number compared to most other systems. This number is of great importance to the efficiency of the fitting process because it defines the dimensionality of the problem,

1. When increasing n, each symmetry function is composed of more terms. The number of terms N_{terms} scales like $N_{terms} \propto n$ for the radial symmetry functions, but for the angular symmetry functions we have that $N_{terms} \propto n^2$.

2. A larger n leads in general to a larger configuration space to fit, thus additional symmetry functions and a larger data set is required.

3. The number of NN parameters must also be increased to obtain the same error as for a lower-dimensional problem

These observations demonstrate that performance is a critical factor whether we fit *ab inito* data or not. However, we know that only a small subset of the possible configurations of a system is physically realizable. Most of the configurations will not be visited by a system obeying Newton's laws, and are therefore not important to the dynamics. We thus need some kind of importance sampling method that adequately samples this subset. We realize that the obvious way to achieve this is to sample configurations from molecular dynamics simulations, which yields precisely the subset we are looking for.

7.1.1   Iterative molecular dynamics sampling

In this thesis we employ an iterative scheme to construct a NNP with data sampled from MD simulations. The method have been used by e.g. Raff et al [21] and Behler [1], and is as follows

1. Sample an initial data set from MD simulations.

2. Calculate the energies and optionally the forces of this initial data set with a chosen method.

3. Train a NN on the data from 1. with reference values from 2. to construct a preliminary NN potential.

4. Carry out simulations using the NNP to find new relevant structures.

5. Compute the energies (and forces) of the new structures with the method from 2. and add them to the data set

6. Train a NN on the extended data set to improve the NNP

7. Repeat 4-6. until no new relevant configurations are found.

The idea is to improve the NNP iteratively and in a self-consistent way by applying it in MD simulations to find "holes" in the data set, i.e. configurations on which the NNP predicts incorrect energies and forces. The process is different depending on whether we are constructing a NNP from *ab inito* electronic structure data or reproducing known, empirical MD potentials.

For the latter, we construct the initial data set by sampling configurations from MD simulations that is relevant for the intended applications of the NNP. Furthermore, we are extracting both coordinates and energies from the simulations, i.e. we get both the input data in step 1 and the target values in step 2 simultaneously. To compute the energies in step 4, we simply run a pseudo-simulation on the new structures. This is easily done in LAMMPS by using the `rerun` command, which tells LAMMPS to compute the energies and forces of atomic configurations supplied from file without doing any integration.

When constructing an *ab inito* NNP however, we are immediately faced with a problem when trying to execute this process: to run MD simulations in step 1 we need the PES we are trying to develop. This is often termed as the "self-starting problem", and several procedures have been suggested for handling it.

One option is obviously to generate random configurations. This does however take us back to the original problem discussed in the beginning of this chapter: The large configuration space volume of high-dimensional systems enforces the use of importance sampling methods. It is highly unlikely that random sampling will yield the important regions of configuration space, thus vast amounts of data is needed. Nonetheless, random sampling can be applicable to small systems of 1-3 atoms where the complete configuration space can be represented without too much effort, see chapter 8.

Another option is to use chemical intuition to select certain configurations of the system under investigation. For instance, if a specific chemical reaction is studied, one can use previous knowledge about this reaction to select configurations along the reaction path [19]. The disadvantage of this approach is that such information is not always available.

A more effective method has been presented by Raff et al. [20], where a semi-empirical potential has been deployed to initiate the sampling. If a reasonable potential is chosen and an adequate sampling method is devised (discussed below), several thousand physically relevant structures can be obtained that will constitue the initial data set. It has also been shown [20] that the final PES obtained by the iterative sampling of MD trajectories is not sensitive to the choice of the empirical surface employed in the initial step.

In this thesis we do not construct *ab inito* NNPs, thus we will not go into further details of the above methods.

7.1.2   Sampling algorithms

We have seen that MD simulations supply the data that is needed to construct a NN potential, but we also have to consider how this sampling should be performed to identify new relevant structures. NNs have a very flexible functional form which makes them able to accurately fit data sets, but they can also yield potentially large energy and force errors when evaluated for configurations that are very different from the ones included in the data set. It is therefore of vital importance to include all structures that are relevant for the intended application of the NNP.

Sampling methods should in general be optimized to produce a reasonably uniform density of data points in those regions of configuration space that are relevant to the application of the fit [21]. If there are regions with very high point density compared to others, the NN will be trained to accurately fit these, at the cost of regions that are poorly represented. An exception to this rule is regions where the potential gradient, or forces, are large. These regions, typically configurations where two atoms are close to each other and experience strong repulsive forces, are the most difficult for the NN to fit, and should therefore have a higher density.

Initial sampling

The initial data set is supposed to function as a starting point for the construction of the NNP. In this work we use empirical MD potentials to do the initial sampling. Before the process is initiated, we must decide on the specific application for the NNP that we want to develop. This entails defining a range of temperatures, pressures etc. that the final NNP will be valid for, and sample configurations based on these.

The time interval between selecting configurations during the MD simulations must also be decided. One alternative is to simply sample points at constant time intervals. This approach will however result in a non-uniform density of data points considering that atoms spend most of their time in regions where the forces are small, at least in equilibrium MD.

To sample configuration space with something close to a uniform density, the time interval τ between sampling needs to be a function of the atomic accelerations (or forces). One such variable interval sampling algorithm, suggested by Pukrittayakamee et al. [22], is given as

$$
\tau = \begin{cases} \text{floor}\left[\alpha/a_{\max}\right]\Delta t, & \text{if floor}\left[\alpha/a_{\max}\right] > 0 \\ \text{floor}\left[\alpha/a_{\max}\right]\Delta t + \Delta t, & \text{if floor}\left[\alpha/a_{\max}\right] = 0 \end{cases} \tag{7.1}
$$

where a_i is the absolute value of the acceleration of atom i, Δt is the integration step size, $a_{\max}$ is the maximum acceleration of a chosen molecule, and $\text{floor}(x)$

yields the integer part of x. The constant α is system dependent and must be determined empirically. It is the value of α that determines the sampling frequency. We see from the second condition in (7.1) that configurations with an acceleration $a_i > \alpha$ yields $\tau = \Delta t$, i.e. the shortest possible sampling interval. On the other hand, if $a_i < \alpha$, the interval increases by a factor floor$(\alpha/a_{\max})$. This approach ensures that we sample all configurations where the gradient of the PES is large, while skipping many of the low-gradient regions of configuration space.

We have chosen to use the forces $|\mathbf{F}_i| \propto a_i$ instead of the accelerations, simply because they are readily available in LAMMPS. When sampling systems with periodic boundary conditions, we have to consider the way forces are updated during a simulation. In LAMMPS, any forces attributed to ghost atoms are added back to the originals after *all* forces are computed. This is done in `verlet.cpp`,

```
if (pair_compute_flag) force->pair->compute(eflag,vflag);
// ...
if (force->newton) comm->reverse_comm();
```

The first line computes the forces on all atoms with the chosen `pair_style` potential, while the second line communicates the forces on ghost atoms back to the original atoms according to Newton's third law. This means that to employ the algorithm (7.1), we need to sample configurations *after* this reverse communication is executed, if not the forces will be incorrect. This can be achieved by extending LAMMPS with a new `compute`[1], as all `computes` and `fixes` are invoked after the force calculations.

We have made an additional modification to the algorithm (7.1). Pukrittayakamee et al. [22] sample clusters of atoms for subsequent DFT calculations of the energies of these clusters, and computes τ based on the maximum acceleration of an atom in a cluster. Our strategy is to select a few atoms, usually 1-10, and compute τ seperately for each of these atoms. When the sampling criterium is met, the coordinates of all neighbours relative to the central atom is written to file, as well as the potential energy the central atom experiences in the presence of its neighbours. These coordinates and energies make up the input data and target values, respectively, for a NN to interpolate. Our modified sampling algorithm is thus as follows (written in a less convoluted way than above),

$$\tau = \begin{cases} \text{floor}\left[\alpha/F_i\right], & F_i \leq \alpha \\ 1, & F_i > \alpha \\ \tau_{\max}, & \text{floor}\left[\alpha/F_i\right] > \tau_{\max} \end{cases} \tag{7.2}$$

where τ is measured in units of Δt and $\tau_{\max}$ is the maximum value of τ. F_i is the

---

[1]See section 5.2 for an explanation of various LAMMPS input script commands, and section 5.4 for a general description on how to add functionality to LAMMPS.

total force on atom i including ghost contributions. We have implemented this sampling algorithm as follows,

```cpp
// ...
// loop through atoms that is to be sampled
for (auto ii : chosenAtoms) {
    i = ilist[ii];
    itype = type[i]-1;

    double F = sqrt(f[i][0]*f[i][0] + f[i][1]*f[i][1] + f[i][2]*f[i][2]);

    // update tau every maxDelay step
    if ( !(myStep % maxDelay) ) {

      // no delay if force is larger than alpha
      if (F > alpha[itype]) tau[atomCount] = 1;

      // calculate delay if force less than alpha
      else {
        int factor = floor( alpha[itype] / F );
        if (factor > maxDelay) tau[atomCount] = maxDelay;
        else                   tau[atomCount] = factor;
      }
    }

    // decide whether to sample or not based on current value of tau for atom i
    if ( (myStep % tau[atomCount] > 0) && useAlgo) continue;

    double xi = x[i][0];
    double yi = x[i][1];
    double zi = x[i][2];

    // loop through neighbours of atom i
    // write relative coordinates and atomic energy to file
    jlist = firstneigh[i];
    jnum = numneigh[i];
    for (jj = 0; jj < jnum; jj++) {
      j = jlist[jj];
      // ...
```

The vector `chosenAtoms` contains the indicies of the atoms that are to be sampled. These indicies are decided when the parameters of the employed potential is initiliazed with the input script command `pair_coeff`. The variable `maxDelay` decides how often τ is updated, and serves as a maximum value of τ. Its value does not change in the course of the simulation. Figure 7.1 displays the distribution of forces for one Si atom in a MD simulation of quartz SiO_2, with and withouth the use of the sampling algorithm (7.2). We see from Figure 7.1a that the sampling procedure cuts the number of low-gradient configurations significantly, while keeping all the environments corresponding to large forces. The
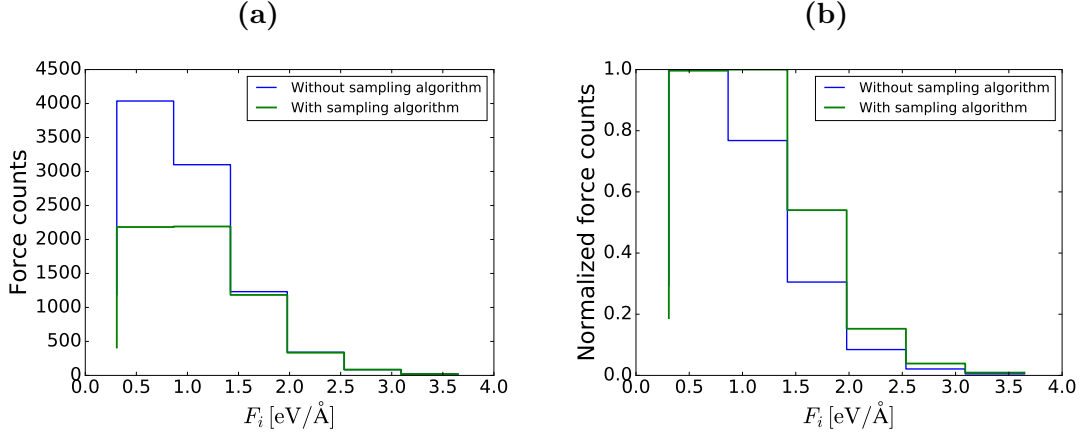
Figure 7.1: Distribution of forces for one Si atom in a quartz SiO_2 crystal with and withouth the use of the sampling algorithm (7.2). The crystal consists of 576 atoms, with an initial temperature $T = 1000\,\mathrm{K}$, run for $N = 10000$ time steps. The atoms are sampled from the microcanonical ensemble with periodic boundary conditions. Sampling algorithm parameters: $\alpha = 3.0$ and $\tau_{\mathrm{max}} = 10$.

resulting distribution is however not uniform, simply because the ratio of force counts in the range $F_i = 0\,\text{Å} - 2\,\text{Å}$ to the range $F_i = 2\,\text{Å} - 4\,\text{Å}$ is too high originally.

One way to smooth the distribution further is to manually delete configurations after the simulation is finished, but then we risk removing relevant structures. The gradient of the PES is not the only way to measure the quality of the sampled set, the time aspect is also important. By selecting samples on-the-fly, we are guaranteed that the they are distributed somewhat uniformly in the time domain. Deleting structures blindly subsequent to the simulation is therefore not a good idea.

Another option is to sample different types of systems where the forces are more evenly distributed. Examples are non-periodic systems or systems containing only a few atoms. Also, it can be a good idea to do short simulations where the atoms start out in non-equilibrium positions to obtain configurations that are not visited by systems with a regular lattice initial state.

Iterative sampling

The initial data set is usually not sufficient to construct a stable neural network PES. It is difficult to identify all the relevant structures that the NN have to learn to perform well, and since there is always a small fitting error, the NNP will generate different trajectories than the PES that has been used to train it. The data set can be extended by sampling new relevant structures from simulations using the preliminary NNP trained on the initial data set. In this

way, new configurations will "naturally present themselves", and the NNP will iteratively improve in a self-consistent way.

The obvious way to identify new structures not already in the initial data set is to evalate both the NNP and the PES we are trying to fit for each structure in an NNP MD simulation and compare the energies (and eventually the forces). If the energy deviation is above a certain treshold, the structure is added to the data set. The simulations should be carried out under the conditions of the intended application of the NNP to ensure that the relevant parts of configuration space are sampled. After a certain amount of new data is added, the NN is trained on this extended data set and thus improved. Then, new simulations are carried out with the improved NNP and so on, according to the iterative process described in the beginning of this chapter.

This approach is easily feasible when the target PES is an empirical potential, but too computationally expensive in the case where electronic structure methods are employed. Raff et al. [20] have developed a sampling algorithm called *modified novelty sampling* that does not require any quantum mechanical calculations of the energies of new structures. They describe the database as a set of vectors $\{\mathbf{q}\}$, where vector $\mathbf{q}_i$ contains the internal coordinates of the i_{th} cluster of atoms. New configurations $\mathbf{q}_n$ are generated using the NNP trained on this data base and the 2-norm difference $\mathbf{d}$ between vectors $\mathbf{q}_n$ and $\mathbf{q}_i$ are computed,

$$\mathbf{d} = |\mathbf{q}_i - \mathbf{q}_n| = [(\mathbf{q}_i - \mathbf{q}_n) \cdot (\mathbf{q}_i - \mathbf{q}_n)]^{1/2} \qquad (7.3)$$

for all clusters i. The difference vector is characterized by its minimum value $d_{\min} = \min(\mathbf{d})$ and the mean seperation distance $d_{\text{mean}} = \langle \mathbf{d} \rangle$. The basic idea is to add configuration point $\mathbf{q}_n$ to the database with high probability if $d_{\min}$ or d_{mean} is large, but low probability if both $d_{\min}$ and d_{mean} are small. The energy output of the NNP are also included in the scheme, along with a more involved selection algorithm. As this thesis only deals with empirical target potentials, the details will not be further discussed here.

In this work we use the Behler symmetry functions to transform atomic cartesian coordinates to vectors of symmetry values. These functions act as descriptors of chemical environments, and should therefore be excellent candidates to identify relevant structures during MD simulations without computing additional target energies. Behler has suggested two methods [1] where these symmetry functions are employed to improve the NNP. The first and simplest method identifies structures that fall outside of the range of input values spanned by the data set. This is achieved by comparing the symmetry function vector of each atom during the MD simulation with the minimum and maximum values of each symmetry function in the database for the respective element. When such configurations are encountered, a warning can be issued and the structure sampled.

The other scheme developed by Behler is called the *multiple-NN method* and is used to identify structures that are within the range of existing symmetry

values, but located in parts of configuration space that are poorly represented in the reference set. The idea is to use the high flexibility and reasonably fast training and evaluation of NNs to validate the potential. The procedure is as follows,

1. Fit several NNs with different architectures to the same data set. All the NNs should be trained to the same accuracy.

2. Generate a large number of trajectories and corresponding energies using one of these NNPs in MD simulations.

3. Recalculate the energies of the obtained trajectories using the other NNPs.

4. Structures for which the energy disagreement between the different NNPs is significant should be added to the data set.

This method takes advantage of the poor extrapolation abilites of NNs: The energy predictions of the different NNPs will be more or less random on structures that are far away from the training points. Thus, missing structures can be identified by comparing the energy predictions of the different NNPs on the genereated trajectories. The procedure also provides a criterion for determining the convergence of the PES without requiring the repeated computation of dynamic properties of the system under study. When, after several iterations of the multiple-NN method, no problematic structures are found, the NNP should be reliable and ready for use.

7.1.3   Summary

In this section we have presented several procedures to assemble the data set for the NN, which depends on how the target energies are computed. In this work, we know the PESs we are trying to fit, which are common empirical MD potentials. Therefore, the NNP can be validated simply by comparing its predictions to the target PES. On the other hand, if an *ab inito* NNP is to be constructed, the underlying target PES is not known, and computing energies and forces of eventual new structures are expensive. Two methods circumventing this problem has been described.

Still, there are several factors to consider when constructing the reference set, irrelevant of how the target PES looks like. We have implemented and employed a sampling algorithm (7.2) in this thesis to obtain a more uniform distribution of forces in the data set. The thermodynamical properties of the MD simulations that we sample configurations from have also been carefully considered, described in more detail in chapter 9.

7.2   Constructing the symmetry function sets

The choice of symmetry functions is important to achieve a good fit to the reference data. They need to be able to sufficiently distinct inequivalent structures to avoid fitting contradictory data. If this distinction is not possible, different reference structures with different energies and forces give rise to the same symmetry function values, which results in poor fits.

Although a NN potential is general and unbiased because it has no pre-defined functional form, the construction of the symmetry function sets is to some extent empirical. The functional form of the NN is adjusted "automatically" when trained, while the symmetry function set has to be constructed manually. This observation does however not change the fact that the high-dimensional neural network approach to constructing potentials used in this work is general, and can in principle be applied to any system. The data set and symmetry function sets have to be customized to each system, but the overarching procedure is the same.

In this section we discuss various teqhniques for building a suitable symmetry function set. A good strategy is to start out with an initial set that satisfies certain criteria, and then customize this set to fit the reference data at hand. We follow and extend upon the discussions in [2] and [10]. Note that each symmetry function can be treated as a variable that is assigned a unique value for each training example in the data set. Thus we occasionaly refer to a single symmetry function as an *input variable*. The symmetry function definitions are found in subsection 4.2.1.

7.2.1   Initial set

There are some general rules that guide the construction of the initial symmetry function set. First, *the symmetry function set should cover the configuration space in an unbiased way.* This can be realized by simply choosing an equidistant set of initial radial symmetry functions (the angular functions are discussed below). Further, the distribution of inter-atomic distances in the data set may be investigated to aquire a lower and upper bound on the range of radial parameter values. For G_i^2, a rule of thumb is that, for each element, the width (standard deviation) of the Gaussian with the largest η value should correspond to the shortest inter-atomic distance present in the data set and vice versa.

Also, a set of symmetry functions with equidistant r_s parameters should be present in the initial set to improve the sensitivity at various radii. These should all have the same, large η value to prevent too much overlap. An example of an initial radial symmetry function set is displayed in Figure 7.2.
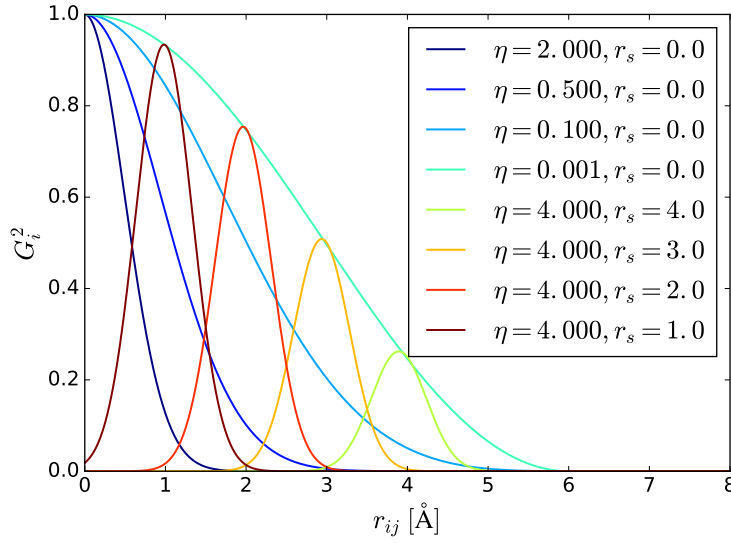
Figure 7.2: Example of an initial radial symmetry function set. The parameters η and r_s are varied to probe the radial arrangement of neighbouring atoms. All symmetry functions have the same cutoff $r_c = 6.0$. r_{ij}, r_s and r_c are in units of Å, while η have units Å^{-2}.

Angular symmetry functions

A set of initial angular symmetry functions can be made in a similar way. The angular distribution of neighbours is probed by varying ζ and λ at differing radii, i.e. for several values of η and/or r_c. We have found from empirical investigations that fixing η, while varying r_c yield the best results. In contrast to the radial parameters, the set of ζ parameters should not be equidistant. Instead, the values of ζ should increase exponentially to ensure that the symmetry functions are sufficiently dissimiliar, i.e. to *avoid strong correlations* (discussed further below). An example is shown in Figure 7.3.

Number of symmetry functions

How many symmetry functions should the initial set consist of? There is a trade-off between the ability of the symmetry function set to cover the configuration space and the efficiency of the calculation of the NN energy output. The key is to select enough functions to capture all regions of configuratiton space without having any redundancies.

Further, the number of functions must be large enough for the NNP to be able to distinguish distinct atomic structures that should yield different total energies. This distinction is more difficult for configurations of a large number of atoms, because e.g. removing one atom from an atomic environment containing
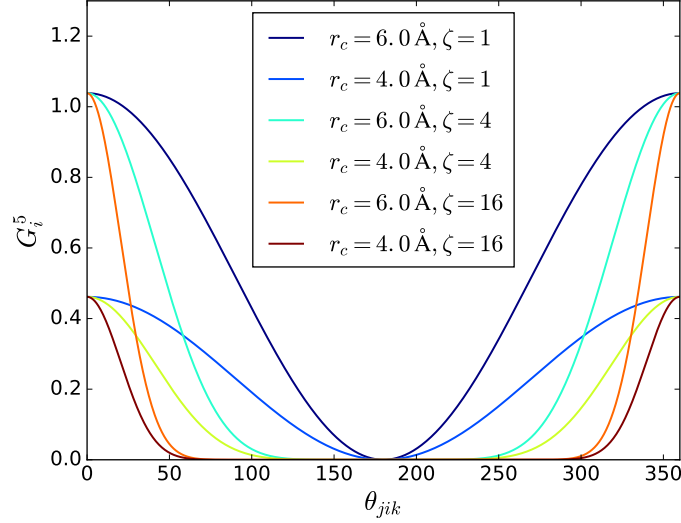
Figure 7.3: Example of an initial angular symmetry function set with $\eta = 0.01\,\text{Å}^{-2}$ and $\lambda = 1$. A corresponding set of functions with $\lambda = -1$ should also be added to make the set more complete. The ζ parameter is increased in a non-linear way to avoid excessively overlapping functions.

100 atoms only results in a small relative perturbation of the symmetry function values. Thus, we should obtain information on how many input atoms the eventual NNP is expected to experience in simulations.

In our case this is an easy task: We simply run MD simulations of the system under investigation with the empirical potential to be fitted, and measure coordination numbers. When the target potential is unknown, the information can be extracted from empirical knowledge of the system. The number of neighbouring atoms will obviously depend on the force cutoffs being used.

One could immidiately conclude that we should have at least as many radial and angular functions as there are atomic pairs and triplets respectively, but this is overkill considering that each one is a function of *all* the atoms inside the cutoff sphere. From experimenting, we have found that a rule of thumb is to use at least as many radial symmetry functions as the mean number of expected neighbours. The number of angular symmetry functions should in general be larger because there are more atomic triplets than pairs for a given ensemble of atoms.

7.2.2   Adjusting the set

When the initial set is assembled, we carry out various analyses to investigate whether our chosen symmetry functions adequately probe the configuration space defined by the data set. There are two requirements that serve as the guiding principles for the analyses.

For each symmetry function, the range of function values should be as large as possible

Each symmetry function assigns *one* numerical value to each training example. If the range of a symmetry function is too small, the difference between the function values for distinct structures will also be small. Consequentially, the NN will try to assign substantial energy changes to tiny changes in the symmetry function, which can lead to numerical instability.

This can be avoided by analyzing the distribution of values of each symmetry function on the data set. If the difference between the maximum and the minimum function value is below a certain threshold the symmetry function should be removed or its parameters adjusted. Specifically, symmetry functions which have a value of zero for all configurations should be avoided, as these will not contribute to the description of the chemical environments.

The set of values of two different symmetry functions on a given data set should not be strongly correlated

If there is a large correlation, the two symmetry functions are almost linearly dependent, and we have a redundancy of information. Consequentially, one of them should be removed, or the parameters adjusted to decrease the correlation. Removing functions in this way contributes to keeping the symmetry function set as small as possible.

To demonstrate how the correlations can be calculated, we remember that in section 3.5 the data set was represented as a matrix X, where each row is a training example, in this case the values of all the symmetry functions for a single atomic configuration. We can calculate the correlations between each pair of symmetry functions with the `NumPy` method `corrcoef`,

```
R = np.corrcoef(np.transpose(X))
```

The data set matrix X is transposed because `corrcoef` expects each *row* to represent all the samples of one variable (function), while each *column* is a single observation of all the variables. A matrix R is returned where the (i, j)-th element is the Pearson correlation coefficient between symmetry function i and j,

$$R_{ij} = \frac{C_{ij}}{\sigma_i \sigma_j} \qquad (7.4)$$

where C is the covariance matrix and σ_i is the standard deviation of the values of symmetry function i for the whole data set. Pair of functions which are strongly correlated can thus be identified and altered.

7.3   Setting hyperparameters

In section 3.5 and section 3.6 we presented the theory on how a MLP is trained by optimizing the weights and biases using an optimization algorithm of choice. At the end of subsection 4.2.1, we briefly mentioned that a number of so-called hyperparameters must be set prior to the actual weight optimization, such as the size of the network. Choosing hyperparameter values can be treated as a kind of *model selection.* One set of parameters defines one model, and the goal is to pick the most appropriate among a family of such models. The choice can have significant impact on the quality of the fit.

We distinguish between two kinds of hyperparameters: those that specify the structure of the network itself, and those that determine how the network is trained. The first type contains parameters that are an integral part of the NN and must be included in the evaluation of the eventual NNP, while the second are parameters that are only used during training. The former includes the symmetry function parameters and any other parameters related to transformations of the input data, the number of layers and nodes, and the choice of activation function(s). The second includes the learning rate, the mini-batch size, the number of training iterations, the choice of loss function and any additional parameters associated with the optimization algorithm and weight initialization.
In this section we will discuss common techniques for setting hyperparameters.

7.3.1   Preconditioning the input data

As stated in subsection 7.2.2, the individual symmetry functions may have very different ranges of values. For a given configuration, each node in the first hidden layer receives as input a weighted sum of all the symmetry function values. Symmetry functions with large absolute values will have a signicantly stronger impact on the output of these nodes than functions with small absolute values. Consequentially, the areas of configuration space covered by the latter functions may take a longer time for the NN to learn. The importance of the different symmetry functions can be balanced by normalizing the values of each symmetry function G_s [10],

$$G_{i,s}^{\mathrm{norm}} = \frac{2(G_{i,s} - G_s^{\mathrm{min}})}{G_s^{\mathrm{max}} - G_s^{\mathrm{min}}} - 1 \tag{7.5}$$

where $G_{i,s}$ is the value of symmetry function number s in the i-th training example[2], and G_s^{min} and G_s^{max} are the smallest and largest values of function G_s in the data set respectively.

Further, convergence is usually faster if the average of each input variable to a NN over the data set is close to zero [38]. This is because the update of a weight to a node in the first hidden layer is proportional to $\delta y_0 = \delta G_s$ (3.63) where δ is

---

[2]Corresponding to element X_{ij} of the input data matrix.

the error of the node, and $y_0 = G_s$ is the value of an input node. Since all the input values G_s are always positive, the weight updates will all have the same sign, namely $\text{sign}(\delta)$. As a result, these weights can only increase or decrease *together* for a given input vector, leading to slow convergence. The mean of each symmetry function can be shifted to zero by subtracting the mean $\overline{G}_s$ over the data set from all function values,

$$G_{i,s}^{\text{shift}} = G_{i,s} - \overline{G}_s = G_{i,s} - \frac{1}{N} \sum_{j=1}^{N} G_{j,s} \tag{7.6}$$

where N is the number of training examples. Note that scaling (7.5) and shifting (7.6) can also be combined.

For reasons outlined in subsection 7.3.2, another operation that can speed up training is variance scaling [38]. This is combined with shifting of means, such that each symmetry function e.g. has zero-mean and unit-variance, called *standardization*,

$$G_{i,s}^{\text{stand}} = \frac{G_{i,s} - \overline{G}_s}{\sigma_s} \tag{7.7}$$

where σ_s is the standard deviation of symmetry function G_s.

Convergence speed vs evaluation speed

We have seen that a number of operations can be performed on the input data to speed up the fitting process. Note that the these can be applied to the target values as well [3]. However, for the construction of NNPs, the convergence speed is not a main issue. In contrast to e.g. object recognition with deep convolutional networks, where often thousands of NN parameters are needed [25], NNPs usually require only up to a few hundred weights. Even though the data set to be fitted can be quite extensive, the duration of the training is rarely more than 30 minutes to 1 hour before an acceptable error is achieved.

Further, there is a trade-off between the convergence speed and the evaluation speed of the resulting NNP. As for the symmetry transformations, the above operations become an integral part of the NNP, and must be performed when the potential is to be applied in simulations. Considering the importance of computational performance in MD simulations, an increase in training speed at the cost of evaluation speed is not to be preferred, as long as the *quality* of the fit is not influenced by these operations. This will be investigated in chapter 9.

---

[3]The observations on scaling and shifting of input variables are valid for most types of input and output data. Scaling of target values are more important for classification, where the output is binary (e.g. {-1,1}), than for regression.

7.3.2   Activation functions and weight initialization

In section 3.4 we stated that the sigmoid (3.15) and the hyperbolic tangent (3.16) are the most common activation functions for regression with neural networks. They are both *saturating* functions because they map real numbers into the restricted intervals [0,1] and [-1,1] respectively. This property keeps the outputs of neurons from diverging, but it also contributes to the problem of *vanishing gradients*. We know from subsection 3.6.3 that weight updates are proportional to the derivatives of neuron outputs y (3.49),

$$y' = f'(u) \tag{7.8}$$

where f is the activation function and u is the input. We see from Figure 3.7 that the sigmoid and hyperbolic tangent have very small gradients for all inputs except those in the small non-linear region around zero, corresponding to about [-4,4] and [-2,2] respectively. Large positive or negative inputs therefore result in vanishing gradients and tiny updates, slowing down learning.

Note that saturation occurs naturally if a network is absolutely correct on all training examples and hence does not need to change its weights. However, premature saturation is a common problem in neural network training, and may slow down or even prevent learning. It can be avoided by restricting neuron inputs u to the non-linear region of their activation functions, with the help of two different procedures.

First, scaling/standardization (subsection 7.3.1) of the input data helps to achieve this for the neurons in the first hidden layer. The distribution of inputs to the next layer depend on the choice of activation function, and here the sigmoid and the tangent differ: The former have mean 1/2, while the latter is symmetric around zero. Obviously, zero-mean is preferred, as any positive or negative bias inevitably have a larger risk of pushing the activations towards saturation. Thus, the hyperbolic tangent is generally preferred over the sigmoid, which have also been demonstrated empirically [6, 43].

Secondly, since the weights and biases also contribute to the sum u, they should be initialized as small numbers centered around zero. A common scheme is to set all biases to zero, and assign weight values according to some probability distribution with zero-mean and a suitable (small) variance. Assuming standardized input data, a frequently used heuristic is to initialize the weights W_{ij} at each layer by [39, 38],

$$W_l \in U \left[-\frac{1}{\sqrt{n_{l-1}}}, \frac{1}{\sqrt{n_{l-1}}} \right] \tag{7.9}$$

where $U[-a, a]$ is the uniform distribution in the interval $[-a, a]$, n_l is the size of layer l, and W_l is the weight matrix connecting layer $l-1$ and l. A disadvantage of this method is that a layer-dependent weight variance might result in saturated upper layers if the layer size changes through the network. Glorot and Bengio

[43] have suggested a modification of this scheme that gives rise to a variance that is equal through the whole network. The idea is to conserve the input signal when propagated through the network by preventing it from shrinking or growing too much,

$$W_l \in U \left[-\frac{\sqrt{6}}{\sqrt{n_{l-1} + n_l}}, \frac{\sqrt{6}}{\sqrt{n_{l-1} + n_l}} \right] \tag{7.10}$$

This is only valid for an activation function that is symmetric and has a unit derivative at 0 ($f'(0) = 1$), i.e. only for the hyperbolic tangent. Further, all input variables are assumed to have the same variance.

7.3.3   Network architecture and overfitting

There is no standard and accepted method for finding the optimal size of a NN for a specific data set. In this work, the input layer is a set of symmetry functions that acts as descriptors of atomic configurations. The size of this set thus decides the number of neurons in the input layer, which is discussed in subsection 7.2.1.

The output layer only consists of *one* neuron that represents the total energy of an atomic configuration. Additionaly, we can have an arbitrary number of hidden layers and nodes. Many nodes means many parameters that can be tuned to the data set, resulting in a better fit. However, if the number of neurons is too large, the network will *overfit* the reference data. In this case, the network fits the data set very well, but interpolation between the training points is poor. Thus, the NNP becomes less predictive, and will not perform well in simulations.

Regularization is a number of processes from statistical learning theory that deals with ill-posed optimization problems and overfitting. A well-known method to prevent the latter is *early stopping*. Here, the reference data is randomly split into a *training set*, which is employed for the optimization of the weights, and a *test set*, whose error is monitored during the fit but which is not used to adjust the weights. A typical test set size is 10-20 % of the reference data. If the errors of these two sets are similar, we know that the NN generalizes to data that has not been used in its training. On the other hand, if the test set error is significantly larger than the training set error, overfitting is occuring, and training should be stopped. The most commmon error measure is the root-mean-square-error (RMSE),

$$\text{RMSE(NN)} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (Y_i - y_i)^2} \tag{7.11}$$

where y_i is the NN prediction and Y_i is the reference value, as before. This expression is almost equivalent to the cost function (3.19) used in the optimation of the NN.
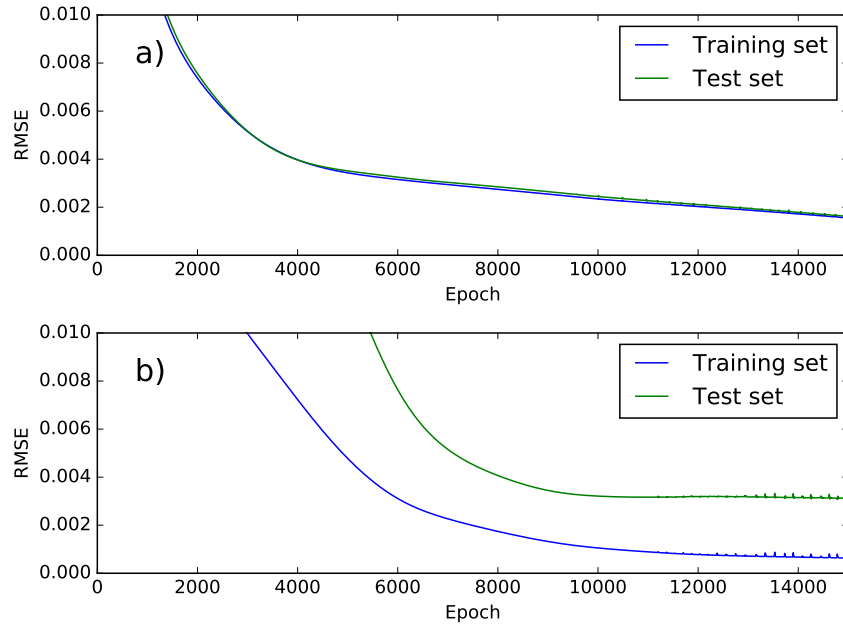
Figure 7.4: Demonstration of overfitting. Both figures show the RMSE of a NN evaluated on both the training set and the test set. The training set consists of 5578 training examples from a Si Stillinger-Weber simulation. In a), the NN has been trained on the whole training set, while in b) only the first 300 training examples were used. The test set is identical in both cases. We clearly see that overfitting occurs in b).

Figure 7.4 displays a typical case of overfitting. Here we have sampled a training set of 5578 configurations and energies from a Si Stillinger-Weber simulation. Then, we have trained two NNs with the same set of hyperparameters, first on the full training set a), then on only the first 300 training examples b). The test set is identical in both cases. We observe that the training set error and the test set error are very similar in a), while in b) there are a significant deviation between the two. The latter is clearly a case of overfitting caused by an incomplete training set.

The early stopping method automatically determines another hyperparameter, namely the number of epochs to train. However, it does not give us the optimal NN architecture. Overfitting can also be caused by an incomplete data set, and reducing the number of neurons is not necessarily the best solution. Often, more data should be added instead, and the current NN size may fit the extended data set better.

Early stopping is in other words not necessarily a reliable method to decide if a NN is of a suitable size or not, but primary a tool to prevent overfitting. The only way to find the optimal architecture is by training several networks of different sizes and choose the one that yields the lowest error while simultaneously not

overfitting. The problem with this strategy is that the possibilities are very many. We can however make a few observations to narrow down the search. According to Larochelle et al. [40], having the same number of nodes in each hidden layer givse the best results. Further, in the literature we have studied concerning NNPs, e.g. [37, 20, 41], only one or two hidden layers are used, and we know from the universal approximation theorem (subsection 3.3.1) that one hidden layer is sufficient to represent any function. For typical AI tasks, deeper architectures are required [42], but in this thesis we will limit ourselves to maximum three hidden layers.

Other regularization teqhniques

There exist other regularization techniques in addition to early stopping that reduce overfitting. *Weight decay* or *L2 regularization* adds an extra term to the cost function, called the *regularization term*, which is the sum of squares of all weights in the network [44],

$$\Gamma = \Gamma_0 + \frac{\lambda}{2N} \sum_w w^2 \qquad (7.12)$$

where Γ_0 is the original, unregularized cost function, N is the size of the training set and λ is the *regularization parameter*. The idea is to penalize large weights, and the modified cost function can be viewed as a compromise between finding small weights and minimizing the original cost function. The parameter λ determines the relative importance of these two elements. Keeping the weights small makes the output of the NN more resistant to outliers and noise in the input data, thus putting more emphasis on the frequent input patterns. The hope is that this will force the NN to see "the bigger picture", and hence generalize better.

Another more recent technique to reduce overfitting is *dropout* [45]. Here, the key idea is to randomly remove neurons from the NN during training. This prevents nodes from co-adapting too much and also decreases the size of the network. We will not go into further detail on this subject here.

We have decided to employ early stopping only to prevent overfitting, as more training data is easily obtained from MD simulations with empirical potentials, making overfitting not a pressing issue.

7.3.4   Optimizer parameters

There are several optimization hyperparameters that need to be decided, namely the learning rate, the mini-batch size (if we employ mini-batch gradient descent) and parameters inherent to the update rule of choice. We restrict ourselves to optimization algorithms that use adaptive learning rates, thus only an *initial* learning rate needs to be set. According to Bengio [39], a default value of 0.01 is

appropriate, but a set of different values should be tested (subsection 7.3.6), as the optimal value may depend strongly on the data set and the NN architecture.

In subsection 3.6.1 we discussed different types of gradient descent learning, and declared that mini-batch gradient descent usually is the preferred scheme. The main reason for this is that data sampled from molecular dynamics simulations, even with a good sampling algorithm, often contain redundant information: many configurations are similar because the system tends to stay near equilibrium[4]. Thus, batch learning, where the weights are updated only after the complete training set have been fed to the network, is wasteful because it recomputes nearly the same outputs several times before updating the weights [25].

However, to employ mini-batch gradient descent, we have to determine another hyperparameter, namely the mini-batch size. According to Bengio [39], 32 is a good default value. Further, it is argued that the mini-batch size should mainly impact the training time and not so much the eventual error of the fit. Therefore, we will not discuss this parameter any further, and refer to subsection 7.3.6 for a discussion on how to optimize hyperparameters.

Some optimizers, like Adam (3.32), contain other hyperparameters in addition to a learning rate. Each of these algorithms has a set of default values that are well justified, and we have decided not to alter these.

7.3.5   Cost function

We only use the mean-square error (3.19) as a cost function in this thesis. An obvious weakness of using this is that we only fit the reference function Y (the potential) itself, not its gradient Y′ (the forces). If a small error is achieved for the potential fit, we would expect that the error is low also for the gradient, but this is not assured. The force error is of vital importance, considering that the forces determine the dynamics of a simulation.

Pukrittayakamee et al. [22] have suggested an extension to the standard mean-square error, that enables simultaneous fitting of both a function and its gradient (the CFDA algorithm). This is achieved by penalizing both the error in the potential and in the forces, and has also been employed by Behler in a modifed form [10],

$$\Gamma = \frac{1}{N} \sum_{i=1}^{N} \left[(E_i^{\mathrm{ref}} - E_i^{\mathrm{NN}})^2 + \frac{\beta}{3N_i} \sum_{j=1}^{3N_i} (F_{i,j}^{\mathrm{ref}} - F_{i,j}^{\mathrm{NN}})^2 \right] \qquad (7.13)$$

As before, the first sum is over N training examples, where E_i^{NN} is the energy prediction of the NN for training example i and E_i^{ref} is the i-th reference energy. For training example i, there are $3N_i$ force components corresponding to a

---

[4]This is of course not the case if the intended application of the NNP is for non-equilibrium molecular dynamics.

configuration of N_i atoms. We have that

$$F_{i,j}^{\text{ref}} = \frac{\partial E_i^{\text{ref}}}{\partial r_{i,j}} \tag{7.14a}$$

$$F_{i,j}^{\text{NN}} = \frac{\partial E_i^{\text{NN}}}{\partial r_{i,j}} \tag{7.14b}$$

where $r_{i,j}$ is the j-th position component of training example i. The NN still has only one output neuron that represents the total energy of a training example. The forces (7.14a, 7.14b) are computed as described in subsection 4.2.2. To minimize (7.13) we need its gradient, which means that second derivatives of the NN are required. This seems like a daunting task, but can be performed quite easily in TensorFlow. The command

```
tf.gradients(y, x)
```

computes the symbolic partial derivative of TF tensor `y` with respect to TF tensor `x`. In general, each tensor in a computational graph can be differentiated with respect to any other tensor in the same graph.

We will not go into further detail on how the CFDA algorithm can be embedded in the training process. Computing second derivatives of the NN during training are computationally demanding, and since we have obtained good results without directly fitting the forces as well, we will stick to the ordinary mean-square cost function.

7.3.6   Hyperparameter space exploration

In this section we have discussed various tricks of the trade regarding hyperparameters for training neural networks. We realize that there are a substantial number of choices to be made, corresponding to a large parameter space, and that many (if not all) of these choices are ipmortant for a successful outcome. Even though the above discussions narrows down the search for suitable hyperparameters, it is still a difficult task.

An alternative is to treat hyperparameter selection as an optimization problem, i.e. to train the NN for many different parameter sets and choose the set that yields the best fit. To perform this task, it is common to split the data set into three parts: A training set, a test set, and a *validation set*. The first two sets have the same purpose as before (subsection 7.3.3), while the validation set is used to test the model during hyperparameter optimization. This is done so that we do not overfit the hyperparameters to the test set, which is only used to assess the performance of the final NN model of choice.

In the following we will describe various hyperparameter optimization schemes.

Coordinate descent

In coordinate descent, we keep all hyperparameters fixed except for one. Then we adjust this single parameter and train the NN for each value to minimize the validation error .

Grid search

Grid search is the most simple and brute-force way to optimize hyperparameters: We simply try every hyperparameter setting over a specified range of values and choose the set with the lowest validation error. This is obviously very computationally expensive, and is often infeasible in practice.

Random search

Instead of performing a grid search, we can sample the hyperparameter space randomly. This is usually a better option because a grid search may take a long time to reach sensible combinations of parameters, and because only a few parameters tend to be vital for the result.

7.3.7   Summary

We have in this section discussed all hyperparameters that needs to be chosen to construct a NNP. It must be stressed that many of the above recommendations for hyperparameter choices may not be suitable for all problems, and that the only way to find unbiased, optimal values is by an extensive search in hyperparameter space. However, the quality of the fit may be insensitive to some parameters, which makes their determination less important. We refer to chapter 9 for a demonstration on how a combination of qualitative arguments and optimization can be performed to find suitable parameters for a specific case.

7.4   Applying the neural network potential

The quality of a NNP is measured by its test set error, but that does not necessarily mean it performs well in MD simulations. To properly validate the NNP, we have to apply it in simulations, measure thermodynamic quantites, and see if we reproduce known, experimental results.

In this thesis we use the TensorFlow (TF) Python API to train neural networks, while LAMMPS, written in C++, is employed to run MD simulations for speed purposes. We therefore need a way to transfer the NNP from Python to C++.

There are built-in functions in TF that automatically saves computational graphs (architecture of network plus nodes for training) and variables (weights

and activation functions) as binary files. These files can then be read and the NN evaluated using the TF C++ API. This API is however quite under-developed and the documentation is very sparse.

Another alternative that we have developed is to manually save the weights and biases as a text file, along with the necessary hyperparameter values, such as choice of activation function and symmetry function parameters. This file can then be used to reconstruct the NN in the same way that it is made in Python: by representing the connection weights between layers as matrices and the biases as vectors (subsection 3.3.2 and subsection 6.2.2). The activation functions must be implemented manually. A linear algebra library can be utilized for speed and simplicity; we have chosen to use Armadillo[5].

To find out which alternative that is the most computationally efficient, we compare the time usage between evaluating a NN with TensorFlow and Armadillo in C++. We also compare with TF in Python to investigate if any differences between TF and Armadillo in C++ is caused by an under-developed C++ TF API, or if the cause is inherent to the TF implementation of neural networks. The evaluation time is measured for a range of different NN sizes, and for each size an average of 50 evaluations have been computed. The results are shown in Figure 7.5.

Figure a) and b) display the ratio $T_{\text{TFC}}/T_{\text{ARMA}}$ for NN sizes up to 200, the former with all data points visible, while the latter is zoomed in to give a more detailed picture. We observe that the time difference is huge for small networks, and that even for NNs of size 200 the ratio is still about 5. From c), we decude that the ratio seems to converge to 1 for large networks, but the convergence rate is surprisingly low. As we see from d), the performance of the built-in TF functions is similar in C++ and Python, thus the time deviation can not be caused by an under-developed C++ API.

One reason for this phenomenon may be that the `Session` environment in TF has a lot of overhead. Further, a NN size of 3000 is small compared to e.g. state-of the art convolutional networks, which are usually trained with GPUs. Considering that we will not need NNs containing more than 100-200 neurons, a manual implementation of NNs with Armadillo is definitively the best choice.

7.5   Summary

In this chapter we have discussed the details of the NNP construction steps, specifically how to sample an adequate training set, how to determine hyperparameters, and how to apply a NNP in simulations. The construction process will be exemplified in two specific cases. In chapter 8, we will train a NN reproduce the truncated and shifted Lennard-Jones potential (2.22). The quality of the fit
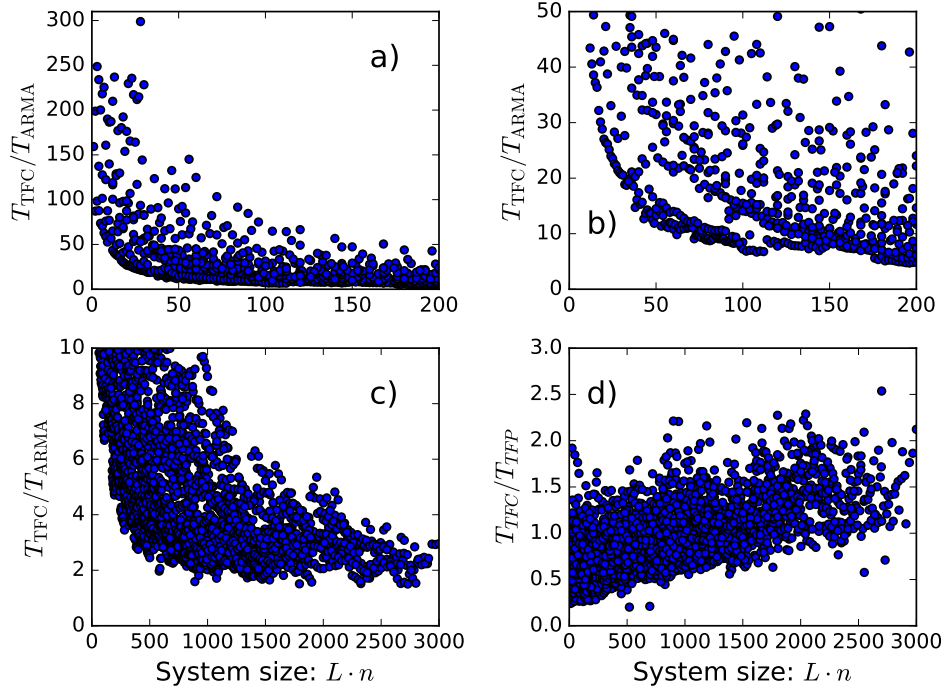
---

[5]Armadillo

Figure 7.5: Scatter plot of time usage when evaluating untrained NNs with random weights and sigmoid activation functions using the TF Python API (T_{TFP}), the TF C++ API (T_{TFC}) and Armadillo (T_{ARMA}). L is the number of hidden layers, n is the number of nodes in each hidden layer. All the NNs have one input and one output. The time has been calculated by averaging over 50 evaluations for each NN architecture.

will be analyzed, but the NNP will not be applied in simulations. In chapter 9, we will construct a NNP for solid-state Si and apply it in simulations to measure the radial distribution function and various mechanical properties.

Chapter 8

Lennard-Jones validation

As a first test of our implementation of the NNP construction process, we train a neural network to reproduce the truncated and shifted Lennard-Jones potential (2.22). The fitting is performed with a NN that has one input and one output, resulting in a NNP that computes the potential energy of a central atom i in the presence of a single neighbour j of the same element at a distance r_{ij}. This one-dimensional NNP does not face the same symmetry issues as its many-dimensional counterpart: having only one input removes the problem that the energy should be invariant to the order of atoms, and the relative coordinate r_{ij} is invariant to a translation or rotation of a diatomic system. Thus, no symmetry functions are needed.

The configuration space of this NNP is simply $r_{ij} \in [a, b]$, where a and b are the smallest and largest interatomic distance respectively that the NNP is expected to experience in a simulation. Suitable values of a and b are found by running a MD simulation with the potential that is to be fitted at the temperature intended for the NNP application of the NNP, and measure the minimum and maximum pair-wise distance. To sample this configuration space, we simply draw random numbers from a uniform distribution with a and b as endpoints,

```
inputTrain = np.random.uniform(a, b, trainSize)
inputTest  = np.random.uniform(a, b, testSize)
```

while the reference energies are calculated by employing the analytical expression for the target function (2.22). The hyperparameters used are listed in Table 8.1. These have not been optimized, but suitable values have been determined on the basis of the discussion in section 7.3. The input data has not been transformed in any way.

After 5000 epochs, we obtain a test set RMSE of 0.896 meV. This is a good result, considering that no hyperparameters were optimized. However, fitting a one-dimensional target function is not a great challenge, despite its non-linear behaviour. The test set RMSE and the training set RMSE follow each other very closely, and no overfitting occurs.

Hyperparameters			
Network structure	Inputs	1	
	Outputs	1	
	Hidden layers	1	
	Hidden nodes	10	
	Activation	Sigmoid	
Training	a	$3.2\,\text{Å}$	
	b	$8.5125\,\text{Å}$	
	Training set size	10000	
	Test set size	1000	
	Mini-batch size	200	
	Number of epochs	5000	
	Optimizer	Adam	
	Initial learning rate	0.001	
	Weight initialization	$U \in [-0.1, 0.1]$	
	Bias initialization	$U \in [-0.1, 0.1]$	

Table 8.1: Hyperparameters used in the fitting of the truncated and shifted Lennard-Jones potential (2.22). The meaning of each parameter is found in the text. "Hidden nodes" signifies the number of nodes in each hidden layer.

8.1   Error in configuration space

It can be useful to visualize the error distribution in configuration space by evaluating the trained NN on all data points in in the test set, and compare with the target function,

```
...
# evaluate trained network on this interval and find absolute error
inputTest = np.sort(inputTest)
energiesNN = sess.run(NNprediction, feed_dict={x: inputTest})
energiesLJ = LennardJones(inputTest)
energyError = energiesLJ - energiesNN
```

where `x` is a `tf.placeholder` as described in subsection 6.2.2. The absolute error $E_{\text{LJ}} - E_{\text{NN}}$ is plotted as a function of distance r_{ij} in Figure 8.1. We notice that the error oscillates around zero, and that the amplitude is larger for smaller distances. This is in accordance with the observations made in subsection 7.1.2: The regions of configuration space where the gradient of the target function is large, is more difficult to fit, and should therefore have a higher density. The training set is in this case uniformly distributed, thus the error is larger for small inter-atomic distances where strong repulsive forces are at work. Nevertheless,
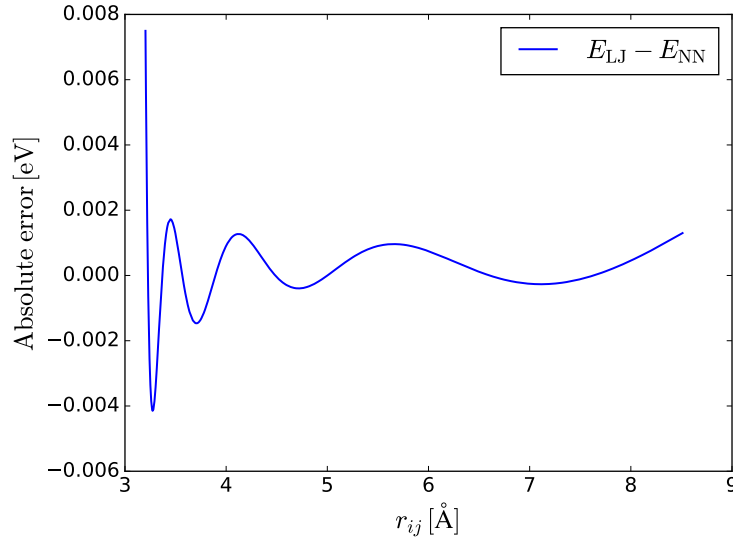
Figure 8.1: Absolute error of a NN trained to reproduce the truncated and shifted Lennard-Jones potential (2.22). The plot is as a visualization of the error distribution in the configuration space $r_{ij} \in [a, b]$ defined by the test set. The employed hyperparameters are listed in Table 8.1.

the NN energy prediction follows the shape of the LJ potential very closely for most values of r_{ij}.

To obtain the forces in a simulation, we must calculate the gradient of the NN. In subsection 7.3.5 we mentioned that a small energy RMSE does not necessarily imply a small force RMSE. In this one-dimensional example, it is easy to measure the gradient error. This is done in a same manner as for the energy above, except now we are comparing the gradients of the NN and the target function. We also calculate the RMSE of the gradient on the test set,

```
...
derivativeLJ = LennardJonesDerivative(inputTest)

# define a NN gradient operation and differentiate
gradientOP = tf.gradients(NNprediction, x)
derivativeNN = sess.run(gradientOP, feed_dict={x: inputTest} )[0]

gradientError = derivativeLJ - derivativeNN
gradientRMSE = np.sqrt(np.sum(gradientError**2)/testSize)
```

We obtain a gradient RMSE of 24.3 meV, which is about one order of magnitude larger than for the function itself. This function-gradient RMSE ratio is also reported by others [54, 55].

The absolute error plot is shown in Figure 8.2. Again we have an oscillating error over the training interval, but with larger amplitudes. The gradient error
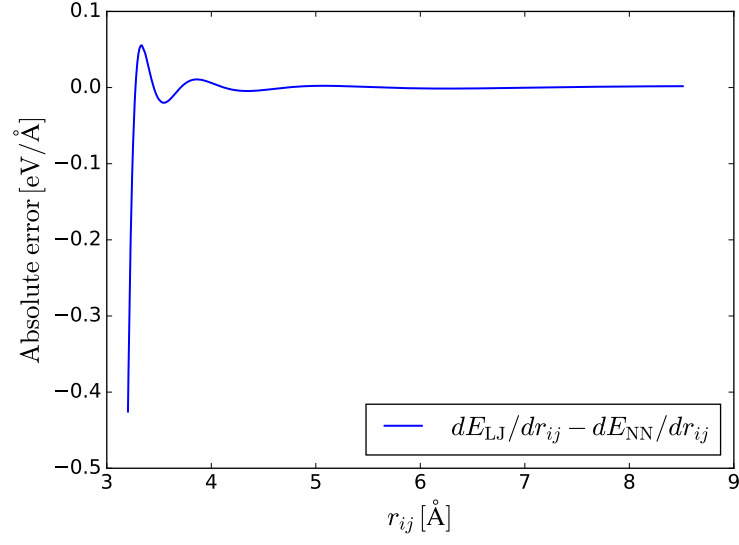
Figure 8.2: Absolute error of the gradient of a NN trained to reproduce the truncated and shifted Lennard-Jones potential (2.22). Comparing with Figure 8.1, we see that the graph is of the same shape, but with larger oscillations. The employed hyperparameters are listed in Table 8.1.

at $r_{ij} = a$ is about 50 times larger than the energy error at this point, which demonstrates the importance of having a higher density of training points in certain regions.

Chapter 9

Neural network potential for Si

In this chapter we perform a demonstration of the complete NNP construction process, including the Behler-Parrinello method. The goal is construct a potential for solid-state Si by sampling atomic configurations and energies from Stillinger-Weber MD simluations. The eventual NNP will be validated by computing the radial distribution function and several mechanical properties, and compare with SW.

9.1 Sampling initial set

To assemble the initial data set, we need to decide on the specific application of the NNP that we want to develop. We choose to construct a potential that can be used to simulate solid-state Si in a temperature range of $T \in [0, 500]$ K. The system is prepared in a diamond cubic state with a lattice spacing of 5.431 Å, which is the crystalline structure of Si. We perform several simulations of the microcanocical ensemble with initial temperatures in the above range and a time step of $\Delta t = 0.001$ ps. The SW potential is parametrized according to [30], with a force cutoff of $r_c = 3.771\,18$ Å. Periodic boundary conditions are imposed and the sampling algorithm (7.2) is applied in each simulation, resulting in 4406 structures and corresponding energies. Note that if an *ab inito* NNP is to be constructed, the reference energies must be calculated with a quantum mechanical method, which is outside the scope of this work.

9.2 Initial symmetry function set

To get an idea of how many symmetry functions that are needed, we run a MD simulation as described above and measure the minimum, maximum and average coordination numbers as a function of time,

```
# coordination number of each atom
```

```
compute neighbours all coord/atom cutoff 3.77118

# global coordination numbers
compute minNeigh all reduce min c_neighbours
compute maxNeigh all reduce max c_neighbours
compute aveNeigh all reduce ave c_neighbours

thermo_style custom step c_minNeigh c_maxNeigh c_aveNeigh
```

After 1e5 time steps, the minimum and maximum coordination number of any atom were 4 and 12 respectively, with an average of around 6. Thus, according to the discussion in subsection 7.2.1, 6-8 radial symmetry functions should be sufficient, while the angular symmetry functions should be at least twice as many. Note that this information must be obtained through emipirical knowledge of the system when constructing *ab inito* potentials.

We have decided to use G_i^2 as the radial function, since this is the better choice (subsection 4.2.1), and G_i^5 as the angular counterpart because of computational efficiency. The initial set thus consists of 8 equidistant G_i^2 functions and 16 G_i^5 functions, with parameters listed in Appendix B.

The initial radial and angular symmetry functions are shown in Figure 9.1, as a function of inter-atomic distance r_{ij} and triplet angle θ_{jik} respectively. Only the angular part of G_i^5 is displayed. We have also calculated the radial and angular distribution of neighbouring atoms in the data set, displayed as dotted lines. This is a convenient way to visualize how well the symmetry set represents the atomic environment: functions that has no overlap with the area defined by the dotted lines does not contribute to the description.

9.3 Fitting the initial data set

According to the iterative scheme in subsection 7.1.1, the next step in the construction process is to create a preliminary NNP by fitting the initial data set. We do not perform a systematic search for optimal hyperparmameters at this intermediate stage. We have decided to employ the sigmoid activation function (3.15) for the remainder of the construction process because it is more computationally efficent compared to the hyperbolic tangent, especially its derivative,

$$f'(x) = f(x)[1 - f(x)] \tag{9.1}$$

Also, we have not experienced any significant improvement with the use of the hyperbolic tangent over the sigmoid. Further, we have chosen the Adam optimizer for all training sessions, as this has been proven to outperform other schemes [18]. All the employed hyperparameters are listed in Table 9.1. The symmetry-transformed training data has been shifted so that each symmetry function has zero-mean, and the minimum and maximum value of each symmetry function
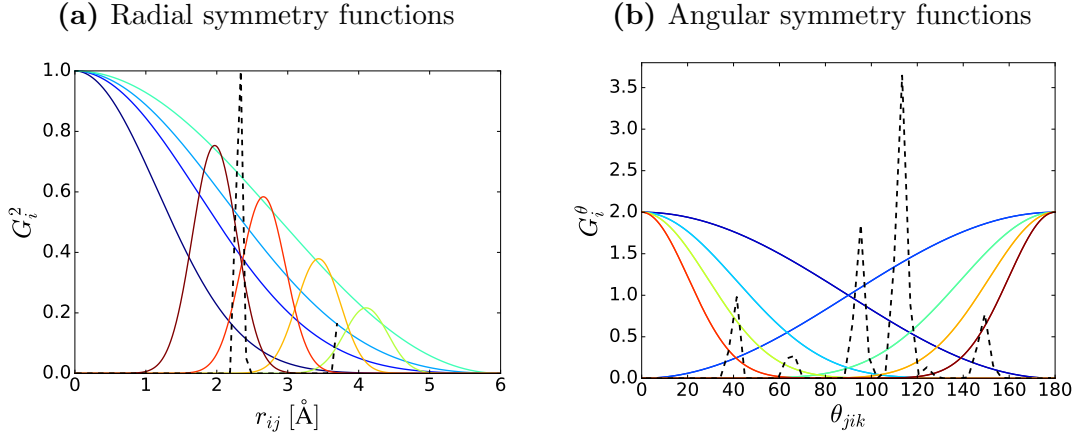
(a) Radial symmetry functions

(b) Angular symmetry functions

Figure 9.1: Initial set of 8 G_i^2 and 16 G_i^5 symmetry functions for constructing a Si NNP. Only the angular part G_i^θ of G_i^5 is displayed. Further, only 8 of these are visible because pairs of G_i^5 functions with identical angular parts overlap. The dotted lines are the radial and angular distribution of neighbours defined by the training set. The corresponding symmetry parameters are not included.

Hyperparameters		
Network structure	Inputs	24
	Outputs	1
	Hidden layers	1
	Hidden nodes	10
	Activation	Sigmoid
Training	Training set size	3900
	Test set size	440
	Mini-batch size	100
	Number of epochs	20125
	Test set RMSE tolerance	1.0 meV
	Optimizer	Adam
	Initial learning rate	0.05
	Weight initialization	$U \in [-0.1, 0.1]$
	Bias initialization	Zeros

Table 9.1: Hyperparameters used in the fitting of the initial data set sampled from Stilling-Weber Si simulations. The data set is split into a training set and a test set, where the latter is 10 % of the total number of structures.

over the training set is stored for detecting extrapolating structures in the next step. Overfitting is detected by the following test,

```python
if testRMSE / trainRMSE > 10:
    print 'Overfitting is occuring, training ends'
    break
```

Several preliminary runs showed the test set RMSE to flatten out at about 1.0 meV, and we therefore decided to stop training if an error below this threshold was obtained, which occured after 20125 epochs. The final test set RMSE was 0.992 meV.

9.4   Iterative sampling

After the initial fit, we apply the preliminary NNP in MD simulations to find new relevant structures not present in the current training set. The Si system is prepared in the same manner as the Stillinger-weber runs above. We use the minimum and maximum symmetry function values stored in the previous step to detect extrapolating configurations according to the procedure described in subsection 7.1.1. Each time a structure leads to a symmetry value outside the range of the above extreme values, we sample that structure, and add it to the training set,

```cpp
bool extrapolation = 0;
for (int s=0; s < m_numberOfSymmFunc; s++) {
  if (symmVector[s] < m_minmax[s][0]) {
    m_minmax[s][0] = inputVector[s];
    extrapolation = 1;
  }
  else if (symmVector[s] > m_minmax[s][1]) {
    m_minmax[s][1] = inputVector[s];
    extrapolation = 1;
  }
}
```

where `m_minmax` is a vector containing the minimum and maximum values of each symmetry function. After a certain number of new structures is added, the NN fit is refined by training on the extended data set. The NNP simulations are carried out for several temperatures $T \in [0, 500]\,\mathrm{K}$. This iterative sampling process is continued until no new extrapolating structures are found, which occurs after three iterations. A total of 578 structures are added, resulting in an extended data set of 4984 training examples.

Next, we employ the multiple-NN method (subsection 7.1.1) to identify new structures that are *within* the range of existing symmetry values, but located in regions of configuration space that are poorly represented in the training set.
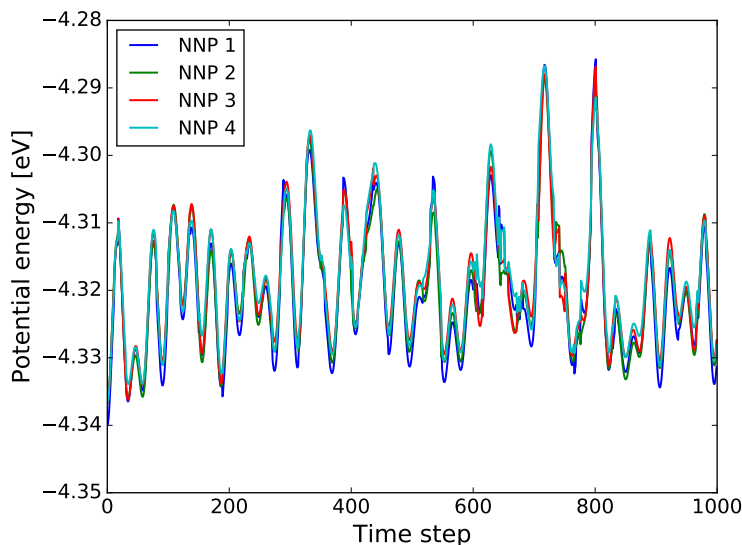
Figure 9.2: Example plot of the multiple-NN method. Four NNs with different architectures have been fitted to the same data set to similar accuracies. Then, several configurations are generated with one of these NNPs, and the potential energy predictions of each NNP for these structures are compared. The plot shows the energies of a single atom.

First, we train four NNs with different architectures on the complete training set to similar accuracies. Then, we run several MD simulations with one of these NNPs to generate a large number configurations. The atomic coordinates are stored with the command

```
dump configs all custom 1 configs.txt id x y z vx vy vz
```

The same hyperparameters as in Table 9.1 are used, except that the RMSE tolerance is changed to 5 meV. Lastly, we run pseudo-simulations with the other three NNPs on the configurations above,

```
rerun configs.txt dump x y z vx vy vz
compute peAtom all pe/atom
dump potEnergy chosen custom 1 potEnergy.txt id c_peAtom
```

and compare the predicted potential energies of the atoms in group `chosen`, which are randomly selected. Any structures for which the energy disagreement is significant, is added to the data set. An example of a potential energy comparison for a single atom is shown in Figure 9.2. The simulations are carried out at an initial temperature of $T = 300$ K, and the first 1000 time steps of one of these

simulations are displayed. For each structure, we measure

$$\Delta E = E_{\max} - E_{\min} \tag{9.2}$$

where ΔE is the difference between the maximum $E_{\max}$ and minimum $E_{\min}$ energy prediction of the four NNPs on that structure. Any configuration for which,

$$\Delta E > \epsilon \tag{9.3}$$

is added to the training set, where ϵ is a constant of choice. We have found that setting ϵ equal to the RMSE tolerance, $\epsilon = 5$ meV, is an adequate choice.

After adding 1595 new structures with the multiple-NN scheme, resulting in a final data set of 6579 training examples, we consider the sampling as complete.

9.5 Fitting the final data set

Now that the iterative sampling scheme is finished, we are ready to construct the final NNP. To investigate if the symmetry function set needs to be adjusted, we calculate the range of each symmetry function on the training data. We find that the G_i^5 functions with a cutoff of $R_c = 3.0$ Å have a very small range of values, thus we change the cutoff to $R_c = 4.0$. Other than that, no adjustments are made.

To find the optimal NN architecture, we perform a grid search of different number of hidden layers L and nodes in each hidden layer n_l. For each architecture (L, n_l) we perform a training session of 40000 epochs, and for every session we store the smallest RMSE obtained, together with the number of epochs and the time elapsed to reach this RMSE. We use the same hyperparamaters as in Table 9.1 (with various architectures), except that no RMSE tolerance is set, and the mini-batch size is changed to 200 to decrease the error oscillations. We assume that no more than two layers are needed to fit the Si potential, and the maximum number of nodes has been set to 32.

The results are listed in Table 9.2. We immediately observe that all RMSEs lie in a range of about 1.5-2.5 meV, except for the two cases where $n_l = 4$. This tells us that the quality of the fit is not very sensitive to the NN size. Some of the two-layer NNs converge at an earlier epoch, but they do not produce significantly better fits, and since each epoch takes a longer time, they do not necessarily converge more *quickly*. Ideally, we should have trained each NN several times and calculated averages, but the above analysis is sufficient to conclude that the choice of NN size is not important to the quality of the fit. For computational efficiency of the eventual NNP, we therefore decide to continue using 1 layer and 10 nodes.

The hyperparameters of the final NN fit are listed in Table 9.3. As for the initial fit, the training data is shifted so that each symmetry function has zero-

Layers	Nodes	RMSE	Epoch	Time
$L = 1$	4	4.445	37035	575
	8	2.250	37305	570
	12	2.303	37980	622
	16	2.201	39780	630
	20	1.860	36180	617
	24	1.928	37305	621
	28	2.407	39375	697
	32	2.214	38700	672
$L = 2$	4	2.947	39960	750
	8	1.933	36180	671
	12	1.450	37350	766
	16	1.791	32265	633
	20	1.492	24840	546
	24	2.118	37620	819
	28	1.455	37350	895
	32	2.008	14895	344

Table 9.2: Results of a grid search for finding the optimal NN architecture to fit the final Si training set. Each NN architecture is trained for 40000 epochs, and the smallest RMSE obtained during each session are listed, together with the number of epochs and time elapsed to reach this RMSE. The RMSEs are in units of meV, while the time is measured in seconds.

Hyperparameters		
Network structure	Inputs	24
	Outputs	1
	Hidden layers	1
	Hidden nodes	10
	Activation	Sigmoid
Training	Training set size	5922
	Test set size	657
	Number of epochs	11000
	Optimizer	Adam
	Initial learning rate	0.05
	Weight initialization	$U \in [-0.1, 0.1]$
	Bias initialization	Zeros

Table 9.3: Hyperparameters used in the fitting of the final training set for construction of the Si potential.

mean. This, in combination with zero-centered (small) weights and biases, prevent pre-mature saturation of the sigmoid activation function, which is of great importance for the quality of the fit.

To investigate whether offline or online learning produce the best fit, we run several training sessions employing both. We find that offline learning consequently results in smaller RMSEs, thus we decide to use this method for the final fit. This is slower than online learning, but since we work with quite small data sets, convergence speed is not an issue, as we can see from Table 9.2.

The NN is trained for 40000 epochs, and for every 1000 epochs the current NN parameters and RMSE are stored. When training is finished, the NN with the smallest RMSE is selected as the final NNP. The best fit was obtained after 11000 epochs, with an RMSE of 0.864 meV. This result is similar to other studies employing the Behler-Parrinello method for NN interpolation [37, 55, 54] .

We have only fitted energies, and not forces, thus the exact force RMSE is not easily available. This would have required the computation of Stillinger-Weber forces for all configurations sampled from NNP simulations, and vice versa. However, a good estimate can be obtained by running a SW simulation to generate a large number of configurations, and store the total force on a single atom at each time step. Then, we run a pseudo-simulation with the NNP on these configurations, and compare the computed forces on the same atom. The resulting force RMSE is 41.2 meV, which is also similar to other studies [55, 54]. As for the Lennard-Jones NNP (chapter 8), the RMSE of the forces is between 1-2 orders of magnitude larger than that for the energy.

We also wish to investigate if there is any correlation between the force error and the magnitude of the force. This can be done by binning the set of force errors calculated above, and computing the RMSE of each bin, shown in Figure 9.3. The data points are placed at the bin centers. We clearly see that there is a correlation, and that the error of the strongest forces are significantly larger than for small forces. This tells us that the large potential gradient-regions of configuration space do not have a sufficiently high density in the training set, despite the use of the sampling algorithm (7.2) for the initial data set.

9.6   Applying the NNP

Calculating fitting errors is not sufficient for assessing the applicability of a NNP. To properly validate the potential we have constructed, we need to evaluate its accuracy in MD simulations. For validation, we have chosen to confirm that the crystal structure of Si and various mechanical properties are reproduced by the NNP by comparing with Stillinger-Weber results.
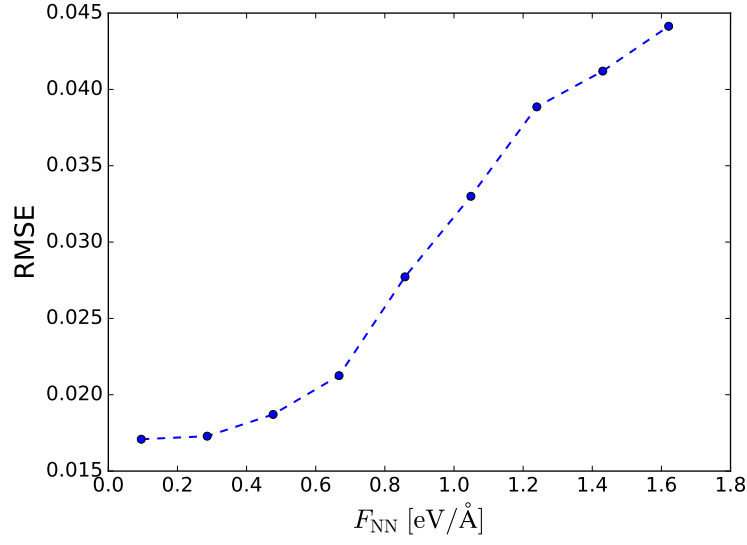
Figure 9.3: The RMSE of the Si NNP force predictions on a single atom as a function of the total force on the atom, demonstrating a clear correlation. The force errors were calculated based on 3000 configurations generated by Stillinger-Weber simulations.
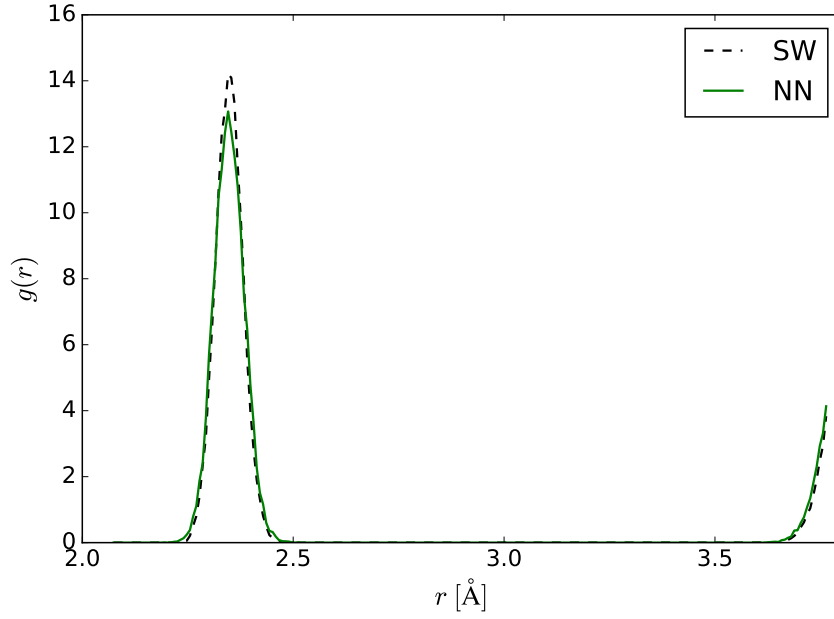

Figure 9.4: Comparison of the radial distribution function of a Stillinger-Weber (SW) simulation and a NNP simulation of a Si crystal at temperature of $T \approx 150$ K. The crystal structure of Si is reproduced by the NNP, with a small difference attributed to the increased kinetic energy of the atoms due to the energy error.

9.6.1   Radial distribution function

The radial distribution function $g(r)$, (or pair correlation function) of an atomic system is a measure of how particle density varies as a function of distance from a reference atom. If a system is randomly distributed, all distances are equally probable, and $g(r)$ is unity for all distances. Any devations from unity signifies an ordered structure. The typical shape of $g(r)$ for a crystal lattice is a flat region for small distances due to repulsive forces, followed by a series of peaks corresponding to bond lengths.

To measure $g(r)$, we preare a system consisting of 1000 Si atoms on a diamond cubic lattice with a lattice constant of $b = 5.431$ Å. The simulation box is cubic with sides $5L$, where L is the length of one unit cube. The initial temperature is $T = 300$ K, and NVE-integration is performed, creating a system trajectory consistent with the microcanonical ensemble. The system is equilibrated for 10 ps, with a time step of $\Delta t = 0.001$ ps. Then, $g(r)$ is calculated every 100 time steps for 1000 time steps, and an average of these 10 measurements are obtained.

The above simulation is executed with both the NNP and the SW potential. The resulting distributions are depicted in Figure 9.4 for $r \in [2.0, r_c]$, with a cutoff of $r_c = 3.77118$ as above. We observe that the distinct crystal structure of Si is reproduced, but the peak centered at $r \approx 2.3$ Å is slightly wider and less tall than than that for SW. This difference is attributed to the error of the energy fit, which leads to a slight increase in the kinetic energy of the atoms, resulting in larger vibrations.

9.6.2   Mechanical properties

To further validate the NNP, we measure the bulk modulus, the shear modulus and the Poisson ratio of the Si crystal in NNP simulations. Bulk modulus K is the coefficient that relates stress to strain under uniform compression,

$$\frac{F}{A} = K\frac{\Delta V}{V_0} \tag{9.4}$$

where F/A is the uniform pressure (stress), and $\Delta V/V_0$ is the induced relative volume change (strain). A material with a large bulk modulus experiences a small volume change when pressure is applied. Note that the material only changes volume, not shape, under a uniform compression.

A force that is applied tangentially (or transversely) to one face of an object is called a *shear* stress. Shear stress changes the shape of an object, but not volume. Shear modulus G is the coefficient that relates shear stress to shear strain.

$$\frac{F}{A} = G\frac{\Delta x}{y} \tag{9.5}$$

where $\Delta x/y$ is the shear strain.

When a material experiences a compressive force, it tends to expand in directions perpendicular to the direction of compression. Conversely, if the material is stretched, it tends to *contract* in the perpendicular direction. The Poisson ratio ν is the ratio of relative contraction to relative expansion, or vice versa. More generally, it is the ratio of transverse strain ϵ_{trans} to longitudinal strain ϵ_{long},

$$\nu = -\frac{\epsilon_{\text{trans}}}{\epsilon_{\text{long}}} \tag{9.6}$$

These quantities can be obtained by employing Hooke's law. For an anisotropic material like Si, this is written as a tensor equation

$$\sigma_{ij} = c_{ijkl}\epsilon_{kl} \tag{9.7}$$

where c_{ijkl} is a fourth rank *stiffness* tensor relating stress σ_{ij} and strain ϵ_{kl}. However, the cubic symmetry of Si allows us to specify the stiffness tensor as a matrix (tensor of rank two) with only three independent components c_{11}, c_{12} and c_{44}, called *elastic constants*. The bulk modulus, shear modulus and Poisson ratio is then calculated by the below formulas [56].
Bulk modulus:

$$B = \frac{c_{11} + 2c_{12}}{3} \tag{9.8}$$

Shear modulus:

$$S = c_{44} \tag{9.9}$$

Poisson ratio:

$$\frac{1}{1 + \frac{c_{11}}{c_{22}}} \tag{9.10}$$

To measure c_{11}, c_{12} and c_{44} in a simulation, we employ a set of example scripts contained in the LAMMPS distribution[1]. These are slightly modified so that our NNP is employed to compute energies and forces instead of SW. The elastic constants are obtained by deforming the simulation box and calculating the induced stress.

The results are listed in Table 9.4. We observe that the SW values are reproduced with small relative errors, demonstrating that also the mechanical properties of the Si crystal are modelled correctly with the NNP. Since SW has been used to compute reference energies for the NNP fit, we do not compare with experimental values, as this would only assess the accuracy of SW, not the NNP.

---

[1]https://github.com/johnands/lammps/tree/master/examples/ELASTIC

	NNP	Analytic SW [57]	Relative error
Bulk modulus	103.0	101.4	1.58 %
Shear modulus	53.6	56.4	5.22 %
Poisson ratio	0.348	0.335	3.88 %

Table 9.4: Various mechanical properties measured from NNP simulations of Si. The analytical values for Stillinger-Weber are included, along with the relative error. The bulk modulus and shear modulus have units of GPa, while the Poisson ratio is unitless.

Part III

Conclusions and future work

Chapter 10

Summary and conclusion

In this thesis, we developed a method for constructing high-dimensional neural network potentials. This method was applied for the construction of an interatomic potential for solid-state Si, and the radial distribution function, bulk modulus, shear modulus and Poisson ratio were measured for benchmarking.

To find atomic structures that were relevant for the intended application of the potential, we implemented a custom `compute` in LAMMPS that sampled atom-centered configurations and total potential energies according to a modified sampling algorithm. This `compute` was used to sample configurations from Stillinger-Weber simulations, which constituted the initial training data for the construction of the Si potential.

We implemented the Behler-Parrinello method to deal with the fixed dimensionality and symmetry problems of neural networks. A set of symmetry functions was created for the initial Si training data. This set was compared with the radial and angular distribution of atoms in the training data to evaluate its ability to adequately probe the atomic environments. The symmetry functions were also used in an iterative sampling scheme to find new relevant structures outside the range of existing symmetry values. Further, the multiple-NN method was employed to detect regions of configurations space poorly represented in the training set.

We developed a set of Python scripts to train neural networks to fit the sampled data sets. The machine learning library TensorFlow was used for this purpose. Various properties of the fitting procedure were investigated by training a one-dimensional neural network to reproduce the truncated and shifted Lennard-Jones potential. We found that the NN were able to accurately fit the functional form of the LJ potential, and obtained a RMSE of 0.896 meV. The absolute error was plotted as a function of inter-atomic distance, demonstrating that the error was largest for small distances.

Further, TensorFlow was used in combination with the BP method to construct a potential for Si by fitting data sets sampled from both Stillinger-Weber and preliminary NNP simulations. The fit was gradually refined with the iterative

sampling scheme mentioned above. A final RMSE of 0.864 meV was obtained, which was comparable with other studies.

Lastly, we applied the eventual Si NNP in MD simuations to evaluate its accuracy. The radial distribution function was measured, and was found to be in good agreement with Stillinger-Weber. The small deviations were attributed to an increase in the kinetic energy of the atoms due to the fitting error. Further, various mechanical properties of the NNP Si system were calculated, specifically the bulk modulus, the shear modulus, and the Poisson ratio. The results had relative errors in the range 1.5-5.2 % when compared to the analytic Stillinger-Weber values.

Properties of a high-dimensional neural network potential

When developing a new PES, it is useful to define a set of properties that the "ideal" potential should possess. These may serve as guiding principles during the construction process, and is also a good starting point for comparing different types of potentials. Behler [1] lists 10 requirements that the ideal potential should meet. We will now evaluate our NNP in light of these requirements.

1. *The potential should be accurate, i.e. produce results that are in good agreement with experiments. Ideally, it should be constructed using ab inito methods.*

   The NNP can be fitted to *ab inito* data to arbitrary accuracy, and can therefore be considered an *ab inito* potential.

2. *There should be ways to systematically detect situations where the potential is not suffientently accurate, and improve it without too much effort.*

   The NNP can be systematically improved by employing the iterative scheme proposed in section 7.1.

3. *The potential should be unbiased and applicable to all types of systems and chemical bonding.*

   The NNP has no pre-defined functional form, and its applications are unlimited as long as the production of *ab inito* reference energies is feasible. However, for systems containing several different chemical elements, the symmetry function vectors grow very large, and the symmetry function sets get increasingly difficult to set up.

4. *The potential should enable the breaking and making of bonds.*

   The NNP is able to model the breaking and making of bonds as long as the reference set contains configurations that are relevant for these processes.

5. *The potential should be high-dimensional, i.e. depend on all degress of freedom of the system.*

   The NNP is high-dimensional because it depends on all the degrees of freedom of the chemical environment of each atom.

6. *The construction of the potential should be as automated as possible.*

   The NN parameters, and hence the functional form of the NNP, is automatically optimized. However, the determination of symmetry parameters and other hyperparmameters is to a certain degree empirical. Further, training data can be automatically produced by MD simulations, but the physical conditions of each simulation and the sampling of configurations must be carefully considered.

7. *The potential should be predictive, i.e. be reliable for atomic configurations that have not been used in the construction.*

   The NNP is generally not able to extrapolate, and is thus reliable only for configurations that are within the configuration space spanned by the training set.

8. *The potential should be fast to evaluate.*

   The NNP is several orders of magnitude faster to evalute than e.g. DFT or Hartree-Fock potentials, and thus enables MD simulations with *ab inito* accuracy on larger scales of space and time. It is however notably slower compared to empirical potentials like Lennard-Jones or Stillinger-Weber.

9. *The construction of the potential should not require much computation time.*

   Training does not require much computation time, but calculating reference energies for thousands of configurations is time consuming.

10. *To compute forces, analytic derivatives of the potential should be available.*
    Analytic derivatives of both the NN and the symmetry functions are readily available.

We observe that our NNP fulfills the majority of the 10 requirements stated above. None of the potentials currently available, including NNPs, are ideal. However, NNPs are an excellent tool for multi-scale physics. They connect the realms of quantum mechanics and molecular dynamics by enabling *ab inito* simulations on a larger scale of space and time than what is feasible for convential *ab inito* MD. This property, in addition to the unbiased and transferable nature of the Behler-Parrinello approach, signify that research in this field will be increasingly important in the future.

Chapter 11

Prospects and future work

An important aspect of the construction of NNPs is to employ a suitable *ab inito* method to calculate reference energies. In this thesis, we have computed target energies with common empirical potentials, and thus not constructed *ab inito* potentials. The most obvious extension to the current work is therefore to employ methods like Hartree-Fock or Density Functional theory to obtain energies. However, there is also much room for improvement in other areas.

11.1 Force accuracy

We saw in chapters chapter 8 and chapter 9 that the force RMSE is generally 1-2 orders of magnitude larger than that for the energy. The forces determine the dynamics of a simulation, and their accuracy is therefore of vital importance. One way to improve the force RMSE is to employ the modified cost function (7.13) to fit both energies and forces. To do this, we must compute reference forces in addition to energies, which is obviously more computationally demanding. However, according to Behler [10], adding forces to the training data reduces the required number of reference energies, leading to a smaller data set.

Further, we have repeatedly observed that the fitting error increases for large forces and energies, which can be avoided by including more high-gradient configurations in the training data. We employed a sampling algorithm (7.2) for this purpose, and although this reduced the problem, it did not remove it entirely. Improved sampling algorithms could therefore be an improvement to this work.

11.2 Other systems

We have constructed a NNP for a system consisting of only one chemical element, but the Behler-Parrinello (BP) method is applicable to systems containing several elements. The BP scheme is high-dimensional, which makes it suitable

for studying complex systems where many degrees of freedom must be included to provide an accurate description of the dynamics. The BP method can also be extended to include electrostatic contributions [10], and water H_2O is therefore an interesting candidate.

NNPs are also able to model the breaking and making of bonds; studying chemical reaction dynamics is therefore a possibility.

11.3   Optimization

For NNPs to serve as a multi-scale physics tool, they must be much more computationally efficient than *ab inito* methods. Ideally, their computational efficiency should be comparable to that of empirical potentials. We have not performed any proper performance analysis in this thesis, but based on several simulations with the Si NNP, we have found that Stillinger-Weber is between 1-2 orders of magnitude faster. Consequentially, the NNP is several orders of magnitude faster than *ab inito* methods[1].

However, optimizing the evaluations of the symmetry functions and the network may contribute to a significant speed-up. We have employed Armadillo to evaluate and differentiate NNPs, which may not the most efficient approach. Further, the size of the NN and the symmetry function set may be reduced by performing a more thorough hyperparmeter optimization.

---

[1]lammps.sandia.gov/bench.html

Appendices

Appendix A

Symmetry function derivatives

According to (4.13), we need the derivatives of the symmetry functions defined in subsection 4.2.1 to obtain forces during simulations. Analytic derivatives have been obtained with SymPy, shown here for G_i^5,

```python
from sympy.utilities.codegen import codegen
from sympy import *

# variables
xj, yj, zj, xk, yk, zk =  symbols('xj yj zj xk yk zk')

# parameters
eta, Rc, zeta, Lambda = symbols('eta Rc zeta Lambda')

# substitution parameters
Rj, Rk, Rj2, Rk2, RjDotRk, CosTheta, Fcj, Fck, dFcj, dFck = \
    symbols('Rj Rk Rj2 Rk2 RjDotRk CosTheta Fcj Fck dFcj dFck')

rj = sqrt(xj**2 + yj**2 + zj**2)
rk = sqrt(xk**2 + yk**2 + zk**2)
fcj = 0.5*cos(pi*rj/Rc) + 0.5
fck = 0.5*cos(pi*rk/Rc) + 0.5
rjDotrk = (xj*xk + yj*yk + zj*zk)
cosTheta = rjDotrk / (rj*rk)

F1 = 2**(1-zeta) * (1 + Lambda*cosTheta)**zeta
F2 = exp(-eta*(rj**2 + rk**2))
F3 = fcj*fck
G5 = F1*F2*F3

dG5dxj = diff(G5, xj)
dG5dxj = dG5dxj.subs(xj**2 + yj**2 + zj**2, Rj2)
dG5dxj = dG5dxj.subs(xk**2 + yk**2 + zk**2, Rk2)
dG5dxj = dG5dxj.subs(sqrt(Rj2), Rj)
dG5dxj = dG5dxj.subs(sqrt(Rk2), Rk)
```

```
dG5dxj = dG5dxj.subs(xj*xk + yj*yk + zj*zk, RjDotRk)
dG5dxj = dG5dxj.subs(RjDotRk/(Rj*Rk), CosTheta)
dG5dxj = dG5dxj.subs(0.5*cos(pi*Rj/Rc) + 0.5, Fcj)
dG5dxj = dG5dxj.subs(0.5*cos(pi*Rk/Rc) + 0.5, Fck)
dG5dxj = dG5dxj.subs(sin(pi*Rj/Rc)/(Rc), -2*dFcj/pi)

dG5dxk = diff(G5, xk)
dG5dxk = dG5dxk.subs(xj**2 + yj**2 + zj**2, Rj2)
dG5dxk = dG5dxk.subs(xk**2 + yk**2 + zk**2, Rk2)
dG5dxk = dG5dxk.subs(sqrt(Rj2), Rj)
dG5dxk = dG5dxk.subs(sqrt(Rk2), Rk)
dG5dxk = dG5dxk.subs(xj*xk + yj*yk + zj*zk, RjDotRk)
dG5dxk = dG5dxk.subs(RjDotRk/(Rj*Rk), CosTheta)
dG5dxk = dG5dxk.subs(0.5*cos(pi*Rj/Rc) + 0.5, Fcj)
dG5dxk = dG5dxk.subs(0.5*cos(pi*Rk/Rc) + 0.5, Fck)
dG5dxk = dG5dxk.subs(sin(pi*Rk/Rc)/(Rc), -2*dFck/pi)

print codegen(("dG5dxj", simplify(dG5dxj)), "C", "file")[0][1]
print codegen(("dG5dxk", simplify(dG5dxk)), "C", "file")[0][1]
```

However, calculating the derivatives by hand allows us to write the expression for the derivatives in a more compact form, speeding up the force computations. Thus, we compute the derivatives of all symmetry functions except G_i^3 in the following. The SymPy expressions serve as a validation of our calculations.

Notation and symmetries

In some cases we only show the deriviative with respect to r_{ij}. Each component α can be calculated by

$$\frac{\partial r_{ij}}{\partial \alpha} = \frac{\alpha}{r_{ij}} \tag{A.1}$$

The following notation applies,

$$r_{ij} = \left(x_{ij}^2 + y_{ij}^2 + z_{ij}^2 \right)^{1/2} = \left((x_j - x_i)^2 + (y_j - y_i)^2 + (z_j - z_i)^2 \right)^{1/2} \tag{A.2}$$

From subsection 4.2.2 we have the identities,

$$\frac{\partial r_{ij}}{\partial x_{ij}} = \frac{\partial r_{ij}}{\partial x_j} \tag{A.3a}$$

$$\frac{\partial r_{ij}}{\partial x_{ij}} = -\frac{\partial r_{ij}}{\partial x_i} \tag{A.3b}$$

Derivatives of cutoff function and radial symmetry functions

Derivative of the cutoff function (4.4):

$$\frac{\partial f_c(r)}{\partial r} = -\frac{1}{2}\frac{\pi}{r_c} \sin\left(\frac{\pi r}{r_c} \right) = M(r) \tag{A.4}$$

This expression is valid for all combinations of indicies (i, j, k). The derivatives w.r.t. to individual coordinates are

$$\frac{\partial f_c(r_{ij})}{\partial x_{ij}} = M(r_{ij})\frac{x_{ij}}{r_{ij}} \tag{A.5}$$

and

$$\frac{\partial f_c(r_{ik})}{\partial x_{ik}} = M(r_{ik})\frac{x_{ik}}{r_{ik}} \tag{A.6}$$

while

$$\frac{\partial f_c(r_{ij})}{\partial x_{ik}} = \frac{\partial f_c(r_{ik})}{\partial x_{ij}} = 0 \tag{A.7}$$

according to (A.3a). Also,

$$\frac{\partial f_c(r_{jk})}{\partial x_{ij}} = -M(r_{jk})\frac{x_{jk}}{r_{jk}} \tag{A.8}$$

and

$$\frac{\partial f_c(r_{jk})}{\partial x_{ik}} = M(r_{jk})\frac{x_{jk}}{r_{jk}} \tag{A.9}$$

The derivative of each term in G_i^1 (4.5) is simply equal to (A.4).
Derivative of G_i^2 (4.6):

$$\frac{\partial G_i^2}{r_{ij}} = \exp(-\eta(r_{ij} - r_s)^2)\left[2\eta(r_s - r_{ij}) + \frac{\partial f_c}{\partial r_{ij}}\right] \tag{A.10}$$

Derivatives of angular symmetry functions

G_i^4 (4.8):

$$G_i^4 = F_1(\theta)F_2(r_{ij}, r_{ik}, r_{jk})F_3(r_{ij}, r_{ij}, r_{jk}) \tag{A.11}$$

where

$$F_1(\theta) = 2^{1-\zeta}(1 + \lambda\cos\theta_{ijk})^\zeta \tag{A.12}$$
$$F_2(r_{ij}, r_{ik}, r_{jk}) = \exp[-\eta(r_{ij}^2 + r_{ik}^2 + r_{jk}^2)] \tag{A.13}$$
$$F_3(r_{ij}, r_{ij}, r_{jk}) = f_c(r_{ij})f_c(r_{ik})f_c(r_{jk}) \tag{A.14}$$

and where

$$\cos\theta = \frac{x_{ij}x_{ik} + y_{ij}y_{ik} + z_{ij}z_{ik}}{r_{ij}r_{ik}} \tag{A.15}$$

Using the product rule:

$$\frac{\partial G_i^4}{\partial x_{ij}} = \frac{\partial F_1}{x_{ij}}F_2F_3 + F_1\left(\frac{\partial F_2}{\partial x_{ij}}F_3 + F_2\frac{\partial F_3}{\partial x_{ij}}\right) \tag{A.16a}$$

$$\frac{\partial F_1}{x_{ij}}F_2F_3 + F_1\frac{\partial F_2}{\partial x_{ij}}F_3 + F_1F_2\frac{\partial F_3}{\partial x_{ij}} \tag{A.16b}$$

We have

$$\frac{\partial F_1}{\partial x_{ij}} = \frac{\partial F_1}{\partial \cos\theta}\frac{\partial \cos\theta}{\partial x_{ij}} \tag{A.17}$$

where

$$\frac{\partial F_1}{\partial \cos\theta} = \lambda\zeta(1 + \cos\theta)^{\zeta-1} = K \tag{A.18}$$

and

$$\frac{\partial \cos\theta}{\partial x_{ij}} = \frac{x_{ik}}{r_{ij}r_{ik}} - \frac{x_{ij}\cos\theta}{r_{ij}^2} \tag{A.19}$$

To get the corresponding expression for x_{ik}, simply substitute $x_{ij} \to x_{ik}$ and $r_{ij} \to r_{ik}$. Further,

$$\frac{\partial F_2}{\partial x_{ij}} = \frac{\partial F_2}{\partial r_{ij}}\frac{\partial r_{ij}}{\partial x_{ij}} = (-2\eta r_{ij}\frac{x_{ij}}{r_{ij}} - 2\eta r_{jk}\frac{x_{jk}}{r_{jk}})\exp[-\eta(r_{ij}^2 + r_{ik}^2 + r_{jk}^2)]$$
$$= -2\eta F_2 x_{ij} - 2\eta F_2 x_{jk} \tag{A.20}$$

The corresponding expression for x_{ik} is

$$\frac{\partial F_2}{\partial x_{ik}} = -2\eta F_2 x_{ik} + 2\eta F_2 x_{jk} \tag{A.21}$$

Lastly,

$$\frac{\partial F_3}{\partial x_{ij}} = \frac{\partial f_c(r_{ij})}{\partial x_{ij}}f_c(r_{ik})f_c(r_{jk}) + f_c(r_{ij})f_c(r_{ik})\frac{\partial f_c(r_{jk})}{\partial x_{ij}} \tag{A.22a}$$

$$= f_c(r_{ik})\left(M(r_{ij})\frac{x_{ij}}{r_{ij}}f_c(r_{jk}) - f_c(r_{ij})M(r_{jk})\frac{x_{jk}}{r_{jk}}\right) \tag{A.22b}$$

The corresponding expression for x_{ik} is

$$\frac{\partial F_3}{\partial x_{ik}} = f_c(r_{ij})\left(\frac{\partial f_c(r_{ik})}{\partial x_{ik}}f_c(r_{jk}) + f_c(r_{ik})\frac{\partial f_c(r_{jk})}{\partial x_{ik}}\right) \tag{A.23a}$$

$$= f_c(r_{ij})\left(M(r_{ik})\frac{x_{ik}}{r_{ik}}f_c(r_{jk}) + f_c(r_{ik})M(r_{jk})\frac{x_{jk}}{r_{jk}}\right) \tag{A.23b}$$

Then, taking (A.16b) into account:

$$\frac{\partial G_i^4}{\partial x_{ij}} = \left(\frac{x_{ik}}{r_{ij}r_{ik}} - \frac{x_{ij}\cos\theta}{r_{ij}^2}\right)KF_2F_3 -$$
$$2\eta F_1 F_2 F_3 x_{jk} - 2\eta F_1 F_2 F_3 x_{ij} + \tag{A.24}$$
$$F_1 F_2 f_c(r_{ik})\left(M(r_{ij})\frac{x_{ij}}{r_{ij}}f_c(r_{jk}) + f_c(r_{ij})M(r_{jk})\frac{x_{jk}}{r_{jk}}\right)$$

and for x_{ik}:

$$
\begin{aligned}
\frac{\partial G_i^4}{\partial x_{ik}} =& \left(\frac{x_{ij}}{r_{ij}r_{ik}} - \frac{x_{ik}\cos\theta}{r_{ik}^2} \right) K F_2 F_3 + \\
& 2\eta F_1 F_2 F_3 x_{jk} - 2\eta F_1 F_2 F_3 x_{ik} + \\
& F_1 F_2 f_c(r_{ij}) \left(M(r_{ik})\frac{x_{ik}}{r_{ik}} f_c(r_{jk}) - f_c(r_{ik}) M(r_{jk})\frac{x_{jk}}{r_{jk}} \right)
\end{aligned}
\tag{A.25}
$$

or in terms of x_{ij} and x_{ik} and x_{jk}:

$$
\begin{aligned}
\frac{\partial G_i^4}{\partial x_{ij}} =& x_{ij} \left(-\frac{\cos\theta}{r_{ij}^2} K F_2 F_3 - 2\eta F_1 F_2 F_3 + \right. \\
& \left. F_1 F_2 M(r_{ij}) f_c(r_{ik}) f_c(r_{jk})\frac{1}{r_{ij}} \right) + \\
& x_{ik}\frac{K F_2 F_3}{r_{ij}r_{ik}} - x_{jk}\left(F_1 F_2 M(r_{jk}) f_c(r_{ik}) f_c(r_{ij}) + 2\eta F_1 F_2 F_3 \right)\frac{1}{r_{jk}}
\end{aligned}
\tag{A.26}
$$

and

$$
\begin{aligned}
\frac{\partial G_i^4}{\partial x_{ik}} =& \frac{x_{ik}}{r_{ik}} \left(-\frac{\cos\theta}{r_{ik}^2} K F_2 F_3 - 2\eta F_1 F_2 F_3 + \right. \\
& \left. F_1 F_2 M(r_{ik}) f_c(r_{ij}) f_c(r_{jk})\frac{1}{r_{ik}} \right) + \\
& x_{ij}\frac{K F_2 F_3}{r_{ij}r_{ik}} + x_{jk}\left(F_1 F_2 M(r_{jk}) f_c(r_{ij}) f_c(r_{ik}) - 2\eta F_1 F_2 F_3 \right)\frac{1}{r_{jk}}
\end{aligned}
\tag{A.27}
$$

The derivative of G_i^5 (4.9) is found in a similar way,

$$
G_i^5 = F_1(\theta) F_2(r_{ij}, r_{ik}) F_3(r_{ij}, r_{ij})
\tag{A.28}
$$

where

$$
F_1(\theta) = 2^{1-\zeta}(1 + \lambda\cos\theta_{ijk})^\zeta
\tag{A.29}
$$
$$
F_2(r_{ij}, r_{ik}) = \exp[-\eta(r_{ij}^2 + r_{ik}^2)]
\tag{A.30}
$$
$$
F_3(r_{ij}, r_{ij}) = f_c(r_{ij}) f_c(r_{ik})
\tag{A.31}
$$

The derivative of F_1 is the same, while for F_2 we obtain

$$
\frac{\partial F_2}{\partial x_{ij}} = -2\eta F_2 x_{ij}
\tag{A.32}
$$

and

$$
\frac{\partial F_2}{\partial x_{ik}} = -2\eta F_2 x_{ik}
\tag{A.33}
$$

For F_3,

$$\frac{\partial F_3}{\partial x_{ij}} = \frac{\partial f_c(r_{ij})}{\partial x_{ij}} \frac{x_{ij}}{r_{ij}} f_c(r_{ik}) = M(r_{ij}) f_c(r_{ik}) \frac{x_{ij}}{r_{ij}} \tag{A.34}$$

and

$$\frac{\partial F_3}{\partial x_{ik}} = f_c(r_{ij}) \frac{\partial f_c(r_{ik})}{\partial x_{ij}} \frac{x_{ik}}{r_{ik}} = f_c(r_{ij}) M(r_{ik}) \frac{x_{ik}}{r_{ik}} \tag{A.35}$$

so that

$$\frac{\partial G_i^5}{\partial x_{ij}} = \left(\frac{x_{ik}}{r_{ij} r_{ik}} - \frac{x_{ij} \cos\theta}{r_{ij}^2} \right) K F_2 F_3 - 2\eta F_1 F_2 F_3 x_{ij} + \\ F_1 F_2 M(r_{ij}) f_c(r_{ik}) \frac{x_{ij}}{r_{ij}} \tag{A.36}$$

and

$$\frac{\partial G_i^5}{\partial x_{ik}} = \left(\frac{x_{ij}}{r_{ij} r_{ik}} - \frac{x_{ij} \cos\theta}{r_{ik}^2} \right) K F_2 F_3 - 2\eta F_1 F_2 F_3 x_{ik} + \\ F_1 F_2 M(r_{ik}) f_c(r_{ij}) \frac{x_{ik}}{r_{ik}} \tag{A.37}$$

In terms of x_{ij} and x_{ik}:

$$\frac{\partial G_i^5}{\partial x_{ij}} = x_{ij} \left(-\frac{\cos\theta}{r_{ij}^2} K F_2 F_3 - 2\eta F_1 F_2 F_3 + \\ F_1 F_2 M(r_{ij}) f_c(r_{ik}) \frac{1}{r_{ij}} \right) + x_{ik} \frac{K F_2 F_3}{r_{ij} r_{ik}} \tag{A.38}$$

and

$$\frac{\partial G_i^5}{\partial x_{ik}} = x_{ik} \left(-\frac{\cos\theta}{r_{ik}^2} K F_2 F_3 - 2\eta F_1 F_2 F_3 + \\ F_1 F_2 M(r_{ik}) f_c(r_{ij}) \frac{1}{r_{ik}} \right) + x_{ij} \frac{K F_2 F_3}{r_{ij} r_{ik}} \tag{A.39}$$

Appendix B

Symmetry function parameters

Here we list the symmetry parameters employed in the construction of the Si NNP in chapter 9.

r_c	η	r_s
6.0	0.3	0
6.0	0.1	0
6.0	0.05	0
6.0	0.005	0
6.0	5.0	4.2
6.0	5.0	3.5
6.0	5.0	2.7
6.0	5.0	2.0

Table B.1: Symmetry function parameters for G_i^2 employed in the fitting of the Si NNP in chapter 9. The same parameters were used in the initial fit and all subsequent fits. r_c and r_s have units Å, while η is in units of Å^{-2}.

η	r_c	ζ	λ
0.01	6.0	1	1
0.01	3.0	1	1
0.01	6.0	1	-1
0.01	3.0	1	-1
0.01	6.0	2	1
0.01	3.0	2	1
0.01	6.0	2	-1
0.01	4.0 (3.0)	2	-1
0.01	6.0	4	1
0.01	4.0 (3.0)	4	1
0.01	6.0	4	-1
0.01	4.0 (3.0)	4	-1
0.01	6.0	16	1
0.01	4.0 (3.0)	16	1
0.01	6.0	16	-1
0.01	4.0 (3.0)	16	-1

Table B.2: Symmetry function parameters for G_i^5 employed in the fitting of the Si NNP in chapter 9. For parameters that were adjusted between the initial and final fits, the values of the initial fit are enclosed in parentheses. r_c have units Å, while η is in units of Å^{-2}. ζ and λ are unitless.

Bibliography

[1] J. Behler. "Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations". In: *Phys. Chem. Chem. Phys.* 13. (2011), 17930-17955. DOI: 10.1039/c1cp21668f.

[2] J. Behler. "Atom-centered symmetry functions for constructing high-dimensional neural network potentials". In: *The Journal of Chemical Physics* 134. (2011), 074106. DOI: 10.1063/1.3553717.

[3] S. A. Dragly. "Bridging quantum mechanics and molecular dynamics with artificial neural networks". MA thesis. University of Oslo (2014).
URL: https://www.duo.uio.no/handle/10852/41843.

[4] K. Hornik, M. Stinchcombe, and H. White. "Multilayer Feedforward Networks are Universal Approximators". In: *Neural Networks* 2. (1989), 359. DOI: 10.1016/0893-6080(89)90020-8

[5] R. Rojas. *Neural Networks: A Systematic Introduction.* Springer, 1996.

[6] B. Karlik and A. V. Olgac. "Performance Analysis of Various Activation Functions in Generalized MLP Architectures of Neural Networks". In: *IJAE* 1. (2011), 111-122.

[7] Y. LeCun, Y. Bengio, and G. Hinton. "Deep learning". In: *Nature* 521. (2015), 436-444. DOI: 10.1038/nature14539

[8] X. Glorot, A. Bordes, and Y. Bengio. "Deep Sparse Rectifier Neural Networks". In: *JMLR* 15. (2011), 315-323.

[9] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. "Learning representations by back-propagating errors". In: *Nature* 323. (1986), 533-536. DOI: 10.1038/323533a0

[10] J. Behler. "Constructing High-Dimensional Neural Network Potentials: A Tutorial Review". In: *International Journal of Quantum Chemistry* 115. (2015), 1032-1050. DOI: 10.1002/qua.24890

[11] Tuckerman, Berne, and Martyna . "Reversible multiscale molecular dynamics". In: *J. Chem. Phys.* 97. (1992), 1990.

[12] S. Plimpton. "Fast Parallel Algorithms for Short-Range Molecular Dynamics". In: *Journal of Computational Physics* 117. (1995) 1-19. DOI: 10.1006/jcph.1995.1039.

[13] M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, R. Jozefowicz, Y. Jia, L. Kaiser, M. Kudlur, J. Levenberg, D. Man, M. Schuster, R. Monga, S. Moore, D. Murray, C. Olah, J. Shlens, B. Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Vigas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu and X. Zheng. "TensorFlow: Large-scale machine learning on heterogeneous systems". *Google Research* (2015). arXiv: 1603.04467

[14] M. Abadi, P. Barham, J. Chen, Z. Chen, A. Davis, J. Dean, M. Devin, S. Ghemawat, G. Irving, M. Isard, M. Kudlur, J. Levenberg, R. Monga, S. Moore, D. G. Murray, B. Steiner, P. Tucker, V. Vasudevan, P. Warden, M. Wicke, Y. Yu, and X. Zheng. "TensorFlow: A System for Large-Scale Machine Learning". *Google Brain* (2016). arXiv: 1605.08695

[15] N. Qian. "On the momentum term in gradient descent learning algorithms" In: *Neural Networks* 12(1). (1999) 145-151. DOI: 10.1016/S0893-6080(98)00116-6

[16] J. Duchi, E. Hazan, and Y. Singer. "Adaptive Subgradient Methods for Online Learning and Stochastic Optimization". In *The Journal of Machine Learning Research* 12. (2011) 2121-2159. URL: http://jmlr.org/papers/v12/duchi11a.html

[17] M. D. Zeiler. "Adadelta: an adaptive learning rate method". (2012). arXiv: 1212.5701

[18] D. P. Kingma and J. Ba. "Adam: A Method for Stochastic Optimization". (2014). arXiv: 1412.6980

[19] J. Ischtwan and M. A. Collins. "Molecular potential energy surfaces by interpolation". In: *The Journal of Chemical Physics* 100. (1994), 8080. DOI: 10.1063/1.466801

[20] L. M. Raff, M. Malshe, M. Hagan, D. I. Doughan, M. G. Rockley, and R. Komanduri. "Ab inito potential-energy surfaces for complex, multichannel systems using modified novelty sampling and feedforward neural networks". In: *The Journal of Chemical Physics* 122. (2005), 084104. DOI: 10.1063/1.1850458

[21] L. M. Raff, R. Komanduri, M. Hagan, and S. T. S. Bukkapatnam *Neural networks in chemical reaction dynamics*. Oxford University Press, 2012.

[22] A. Pukrittayakamee, M. Malshe, M. Hagan, L. M. Raff, and R. Narulkar, S. Bukkapatnum and R. Komanduri. "Simultaneous fitting of a potential-energy surface and its corresponding force fields using feedforward neural networks". In: *The Journal of Chemical Physics* 130. (2009) 134101. DOI: 10.1063/1.3095491

[23] W. S. McCulloch and W. Pitts. "A logical calculus of the ideas immanent in nervous activity". In: *Bulletin of Mathematical Biophysics* 5. (1943), 115. DOI: 10.1007/BF02478259

[24] D. Kriesel. "A Breif Introduction to Neural Networks" Available at http://www.dkriesel.com

[25] Y. LeCun, P. Haffner, L. Bottou, and Y. Bengio. "Object recognition with Gradient-Based Learning". In: *Shape, Contour and Grouping in Computer Vision. Lecture Notes in Computer Science.* 1681. (1999) 319-345. DOI: 10.1007/3-540-46805-6_19

[26] F. Rosenblatt. "The Perceptron: A probabilistic model for information storage and organization in the brain." In: *Psychological Review* 65. (1958), 386-408.

[27] S. Ruder. "An overview of gradient descent optimization algorithms." (2016). arXiv: 1609.04747

[28] R. Dawes, D. L. Thompson, A. F. Wagner, and Michael Minkoff. "Interpolating moving least-squares methods for fitting potential energy surfaces: A strategy for efficient automatic data point placement in high dimensions". In: *The Journal of Chemical Physics* 128. (2008), 084107. DOI: 10.1063/1.2831790

[29] J. E. Jones. "On the Determination of Molecular Fields - II. From the Equation of State of a Gas." In: *Proceedings of the Royal Society of London. Series A* 106. (1924), 463. DOI: 10.1098/rspa.1924.0082

[30] F. H. Stillinger and T. A. Weber. "Computer simulation of local order in condensed phases of silicon." In: *Physical Review B* 31. (1985), 5262. DOI: 10.1103/PhysRevB.31.5262

[31] V. Molinero, E. B. Moore. "Water modeled as an intermediate element between carbon and silico." In: *J. Phys. Chem. B* 113 (13). (2008), 4008-4016. DOI: 10.1021/jp805227c

[32] P. Vashishta, R. K. Kalia, J. P. Rino, and I. Ebbsjö. "Interaction potential for SiO$_2$: A molecular-dynamics study of structural correlations." In: *Physical Review B* 41. (1990), 12197. DOI: 10.1103/PhysRevB.41.12197

[33] P. Vashishta, R. K. Kalia, A. Nakano, J. P. Rino. "Interaction potential for silicon carbide: A molecular dynamics study of elastic constants and vibrational density of states for crystalline and amorphous silicon carbide". In: *Journal of Applied Physics* 101. (2007), 103515.

[34] D. Frenkel and B. Smit. "Understanding molecular simulation: from algorithms to applications." Academic Press, 2001.

[35] P. M. Agrawal, L. M. Raff, M. T. Hagan, and R. Komanduri. "Molecular dynamics investigations of the dissociation of SiO$_2$ on an *ab inito* potential energy surface obtained using neural network methods." In: *The Journal of Chemical Phyics* 124. (2006), 134306. DOI: 10.1063/1.2185638

[36] A. Bholoa, S. D. Kenny, and R. Smith. "A new approach to potential fitting using neural networks". In: *Nuclear Instruments and Methods in Physics Research Section B: Beam Interactions with Materials and Atoms* 255. (2007) 1. DOI: 10.1016/j.nimb.2006.11.040

[37] J. Behler and Michele Parrinello. "Generalized Neural-Network Representation of High-Dimensional Potential-Energy Surfaces". In: *Physical Review Letters* 98. (2007), 146401. DOI: 10.1103/PhysRevLett.98.146401

[38] Y. LeCun, L. Bottou, G. B. Orr, and K.-R. Mller. "Efficient backprop". In: G. Montavon, G. B. Orr, K.-R. Mller (Eds.), *Neural Networks: Tricks of the Trade* (2nd edition, pp. 9-48). Springer, 2012.

[39] Y. Bengio. "Practical Recommendations for Gradient-Based Training of Deep Architectures." In: G. Montavon, G. B. Orr, K.-R. Mller (Eds.), *Neural Networks: Tricks of the Trade* (2nd edition, pp. 437-478). Springer, 2012. arXiv: 1206.5533v2

[40] H. Larochelle, Y. Bengio, J. Louradour, and P. Lamblin. "Exploring Strategies for Training Deep Neural Networks". In: *Journal of Machine Learning Research* 10. (2009), 1-40. DOI: 10.1145/1577069.1577070

[41] J. B. Witkoskie, and D. J. Doren. "Neural Network Models of Potential Energy Surfaces: Prototypical Examples". In: *Journal of Chemical Theory and Computation* 1. (2005), 14-23. DOI: 10.1021/ct049976i

[42] Y. Bengio, and Y. LeCun. "Scaling learning algorithms towards AI". In: L. Bottou, O. Chapelle, D. DeCoste, and J. Weston (Eds.), *Large-scale kernel machines*. MIT Press, 2007.

[43] X. Glorot, and Y. Bengio. "Understanding the difficulty of training deep feedforward neural networks." In: *Proceedings of AISTATS* 9. (2010), 249-256.

[44] A. Krogh, and J. A. Hertz. "A simple weight decay can improve generalization". In: D. S. Lippman, J. E. Moody, and D. S. Touretzky (Eds.) *Advances in neural information processing systems vol. 4* (pp. 950-957). Morgan Kaufmann.

[45] N Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, R. Salakhutdinov. "Dropout: A Simple Way to Prevent Neural Networks from Overfitting." In: *Journal of Machine Learning Research* 15. (2014), 1929-1958.

[46] A. Krizhevsky, I. Sutskever, and G. Hinton. "ImageNet classification with deep convolutional neural networks." In: *Proc. Advances in Neural Information Processing Systems* 25. (2012), 10901098.

[47] G. Hinton, et al. "Deep neural networks for acoustic modeling in speechrecognition." In: *IEEE Signal Processing Magazine* 29, (2012), 8297.

[48] R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa. "Natural Language Processing (Almost) from Scratch". In: *Journal of Machine Learning Research* 12. (2011), 2493-2537. arXiv: 1103.0398

[49] T. Ciodaro, D. Deva, J. M. de Seixas, and D. Damazio. "Online particle detection with Neural Networks based on topological calorimetry information". In: *Journal of Physics: Conference Series*, 368. (2012), 012030.

[50] M. Helmstaedter, K. L. Briggman, S. C. Turaga, V. Jain, H. S. Seung, W. Denk. "Connectomic reconstruction of the inner plexiform layer in the mouse retina". In: *Nature* 500. (2013), 168-174. DOI: 10.1038/nature12346

[51] G. Carleo, and M. Troyer. "Solving the quantum many-body problem with artifical neural networks". In: *Science* 355. (2017), 602-606. DOI: 10.1126/science.aag2302

[52] M. Krenn, M. Malik, R. Fickler, R. Lapkiewicz, and A. Zeilinger. "Automated Search for new Quantum Experiments". In: *Physical Review Letters* 116. (2016), 090405. DOI: 10.1103/PhysRevLett.116.090405

[53] F. V. Prudente, P. H. Acioli, and J. J. S. Neto. "The fitting of potential energy surfaces using neural networks: Application to the study of vibrational levels of H_3^+". In: *The Journal of Chemical Physics* 109. (1998), 8801. DOI: 10.1063/1.477550

[54] S. K. Natarajan, and J. Behler. "Neural network molecular dynamics simulation of solid-liquid interfaces: water at low-index copper surfaces". In: *Phys. Chem. Chem. Phys* 18. (2016), 28704. DOI: 10.1039/c6cp05711j

[55] N. Artrith, and J. Behler. "High-dimensional neural network potentials for metal surfaces: A prototype study for copper". In: *Physical Review B* 85. (2012), 045439. DOI: 10.1103/PhysRevB.85.045439

[56] M. A. Hopcroft, W. D. Nix, and T. W. Kenny. "What is the Young's Modulus of Silicon?". In: *Journal of microelectromechanical systems* 19. (2010), 229-238.

[57] E. R. Cowley. "Lattice Dynamics of Silicon with Empirical Many-Body Potentials". In: *Physical review letters* 60. (1988), 2379-2381.

[58] C. D. Sherrill. "Frontiers in electronic structure theory". In: *The Journal of Chemical Physics* 132. (2010), 110902. DOI: 10.1063/1.3369628

[59] H. Eshet, R. Z. Khaliullin, T. D. Khne, J. Behler, and M. Parrinello. "*Ab inito* quality neural-network potential for sodium". In: *Physical Review B* 81. (2010), 184107. DOI: 10.1103/PhysRevB.81.184107