

# An adaptive Sparse Grid Approach for Many-Body Systems

Karen Fischhuber

Born 15th January 1992 in Lahnstein, Germany

October 29, 2018

Master's Thesis Mathematics

Advisor: Prof. Dr. Michael Griebel

Second Advisor: Prof. Dr. Alexander Schweitzer

MATHEMATISCHES INSTITUT

MATHEMATISCH-NATURWISSENSCHAFTLICHE FAKULTÄT DER  
RHEINISCHEN FRIEDRICH-WILHELMS-UNIVERSITÄT BONN



# Abstract

In this work we use least squares regression to approximate the high-dimensional Born-Oppenheimer potential energy surface. Therefore we will combine the well-known many-body expansion approach with an adaptive sparse grid approach to optimally reduce the regression search set. Besides finding a good search space for the predictor we also will explore how to choose a good dataset to train the predictor, using active learning and the D-Optimality principle.

By first deriving principles of how to reach the stated goals in theory we then know what to look for in the practical numerical experiments. We will show that a more careful choice of the training data and search space is a promising approach to greatly reduce the cost of an approximation. We will evaluate the methods on the W-14 data set containing diverse atomic environments of elemental tungsten and the MD-17 molecule dynamics data set.



# Contents

<b>1</b>	<b>Introduction</b>	<b>1</b>
1.1	Using Machine Learning to Approximate the PES . . . . .	3
1.2	An adaptive Sparse Grid Approach for Many Body Systems . .	5
<b>2</b>	<b>Penalized Least Squares Regression</b>	<b>9</b>
2.1	Derivation of the Optimization Problem . . . . .	10
2.1.1	Least Squares Regression with given Measure . . . . .	10
2.1.2	Least Squares Regression with given Samples . . . . .	12
2.1.3	Least Squares Regression on a finite dimensional Search Set . . . . .	13
2.1.4	Penalized Least Squares Regression . . . . .	13
2.2	Bayesian Interpretation . . . . .	17
2.3	Stability Analysis in the Unpenalized Case . . . . .	20
2.4	Related Work . . . . .	23
<b>3</b>	<b>Sparse Tensor Product Construction</b>	<b>25</b>
3.1	A Tensor Product Construction of the Function Space . . . . .	25
3.1.1	Construction in One Dimension . . . . .	26
3.1.2	Construction in Multiple Dimensions . . . . .	27
3.1.3	General Approximation Operator . . . . .	28
3.2	Choice of an Index Set . . . . .	33
3.3	Related Work . . . . .	34
<b>4</b>	<b>Active Learning</b>	<b>37</b>
4.1	From Passive Learning to Active Learning . . . . .	38
4.2	Variance Reduction . . . . .	40
4.3	The Meaning of the Fisher Information . . . . .	41
4.4	Application to the Linear Regression Model . . . . .	44
4.4.1	Cramér-Rao for Linear Regression . . . . .	44
4.4.2	Minimizing the Cramér-Rao bound . . . . .	46
4.4.3	Query Strategy with fixed Training Size . . . . .	47
<b>5</b>	<b>The Born-Oppenheimer Potential Energy Surface</b>	<b>49</b>
5.1	The Potential Energy Surface . . . . .	49
5.2	The Atomic Decomposition Ansatz . . . . .	53
5.3	Physical and Computational Assumptions on the Potential Func- tion . . . . .	55

5.4	The Many-Body Expansion . . . . .	55
5.5	Descriptors of Local Atomic Environments . . . . .	59
<b>6</b>	<b>Approximating the PES</b>	<b>63</b>
6.1	Constructing a Regression MLIP $\hat{V}_{\mathcal{D}_n, V_p}$ . . . . .	64
6.1.1	Choosing a Descriptor . . . . .	64
6.1.2	Choosing the Search Set . . . . .	65
6.1.3	K-Body Regression Potential . . . . .	70
6.2	Choosing an Optimal Search Set $V_{\text{opt}}$ . . . . .	73
<b>7</b>	<b>Assessment and Validation</b>	<b>77</b>
7.1	Group Structure for used Physical Applications . . . . .	77
7.1.1	Solids with Periodic Boundary . . . . .	77
7.1.2	Molecules . . . . .	80
7.2	Comparing the different Basis Functions . . . . .	81
7.3	Remarks on the Implementation . . . . .	82
7.4	Data Sets . . . . .	83
7.4.1	The Tungsten Data Set . . . . .	83
7.4.2	The Molecular Dynamics (MD) Data Set . . . . .	85
<b>8</b>	<b>Numerical Results</b>	<b>89</b>
8.1	Results on Tungsten Data . . . . .	89
8.1.1	Varying the Penalization Norm . . . . .	89
8.1.2	Incorporating the Regularization Parameter . . . . .	96
8.1.3	Active Learning . . . . .	96
8.2	Results on the Molecular Dynamics (MD) Data Set . . . . .	100
8.2.1	A Sparse Tensor Product Approach . . . . .	101
<b>9</b>	<b>Conclusion</b>	<b>107</b>

# Chapter 1

## Introduction

In the area of material science we study the characteristics of materials. This includes finding stationary states and understanding chemical reactions. A primary goal is the design and discovery of new materials. As physical systems tend towards states that minimize their energies, calculating energy as function of that state is an important step. The macroscopic behavior of these materials are dictated by their nanoscale dynamics which are described by quantum dynamics.

In 1926, Schrödinger laid the foundation for quantum mechanical development with the publication of the Schrödinger equation (SE). Where Newtonian theory deals with classical mechanics, Schrödinger's describes changes over time of quantum mechanical systems, including systems containing nuclei and electrons. The Schrödinger equation is only analytically solvable for the hydrogen and hydrogen-like atoms. Taking just a single particle, the Schrödinger equation is already a complex second order differential equation in three dimensions. One can simplify things by treating the nuclear system with classical mechanics. But even in this case, simple systems containing only three particles can have no closed form solution.

*Schrödinger  
Equation*

A first simplification of the Schrödinger equation was the Born Oppenheimer approximation. In 1927, Born and Oppenheimer decoupled the behaviour of electrons and nuclei to simplify the approach of Schrödinger. Due to the enormous difference in mass and movement of electrons and nuclei, the nuclei are regarded motionless compared to the fast electrons. For a fixed configuration of slow nuclei, the Schrödinger equation is solely solved for the electrons. Although of lower dimension, the so-called electronic Schrödinger equation (eSE) is still in general not solvable. If one only treats the time-independent case, both the SE and the eSE are eigenvalue problems. Each eigenvalue encodes the energy of a physical state uniquely described by its eigenfunction - the corresponding wave function. As a further simplification, we are only interested in the physical state with the lowest energy, the ground state. Assigning each fixed set of nuclear positions to its corresponding electronic ground state energy forms the so-called *Born Oppenheimer potential energy surface* (BO-

*Born  
Oppenheimer  
Approximation*

PES). The Born Oppenheimer PES is defined on a high-dimensional domain and a closed form method to evaluate one point at the BO-PES does not exist, thus a wealth of approximation methods have evolved.

### *QM/MM Methods*

The technical advances of the last decades unleashed a new wave of possibilities and research. As a consequence, a plurality of approximate simulation methods evolved. Computational quantum chemistry was born. In molecular dynamics (MD), one roughly distinguishes between two classes of interatomic interaction models; the more accurate, but slow quantum mechanical (QM) methods, and the less accurate but fast molecular mechanical (MM) methods. QM methods are derived from exact quantum mechanical equations. Although in practice they offer only an approximation of the exact solution of the Schrödinger equation, they often operate as benchmarks for further considerations. Moreover, the errors are largely systematic and well documented. Examples for QM methods are electronic structure methods (ESM) such as Density Functional Theory (DFT) and Hartree-Fock (HF). ESM methods require the calculation of integrals, which in Empirical Quantum Mechanical Methods (EQM) are replaced by an empirical term. MM methods, however, operate in a simplified setting. They use classical mechanics instead of quantum mechanical equations to model molecular systems. Thus, atoms are simply treated as particles interconnected with springs. The idea of combining QM computations and MM simulations to form an seamless coupling of both, is called a hybrid approach or multiscale methods (QM/MM). In multiscale methods, one aims to split the problem and use QM methods where necessary and MM methods elsewhere. This groundbreaking idea was introduced in 1976, and resulted in the awarding of the Nobel price in Chemistry in 2013 to M. Karplus, M. Levitt and A. Warshel for their “development of multiscale models for complex chemical systems” [3].

### *Locality Assumptions*

Describing molecules has a much longer history than the material models, and in fact there are a number of critical requirements that differentiate material models from molecules. A potential describing materials needs to describe the forming and breaking of ionic and covalent bonds. Comprehensive models also cover multiple phases of the material, e.g. solid and liquid, and even phase transitions under changes of temperature and pressure. To model materials we often have to effectively breaking the system in smaller subsystems, which typically will still consist of tens to hundreds of atoms. A commonly used approach is the so-called (atomic decomposition). Here, we assume two particles to interact if and only if its distance is smaller than a prescribed threshold  $r_{\text{cut}} > 0$ . This way, the energy decomposes into energy contributions of  $r_{\text{cut}}$ -balls centered at each particle. I.e. for a system  $X = (X_1, \dots, X_M)$  containing  $M$  particles, we assume the potential function to decompose in the following way

$$V(X) \approx \sum_{i=1}^M V_{\text{atomic}}(B_{r_{\text{cut}}}(X_i)) \quad (1.0.1)$$

A common approach to further reduce the dimension are approaches based



on additivity models. Similar to the Analysis of Variance approach (ANOVA), these subdivide the subsystems dimension-wise. The energy is represented as a finite sum of contributions which depend on the positions of single nuclei, of pairs of nuclei, of triples of nuclei, and so on,

$$\begin{aligned}
 V_{\text{atomic}}(B_{\text{cut}}(X_i)) = & \sum_{\substack{A \subset B_{\text{cut}}(X_i) \\ |A|=1}} V_1(A) + \sum_{\substack{B \subset B_{\text{cut}}(X_i) \\ |B|=2}} V_2(B) \\
 & + \sum_{\substack{C \subset B_{\text{cut}}(X_i) \\ |C|=3}} V_3(C) + \dots
 \end{aligned} \tag{1.0.2}$$

This is called *many-body expansion* and there are many well-known methods building up on this, see [41] and references therein. The prevalent  $k$ -body potentials only consider sets of atoms up to size  $k$ . These approaches assume locality of electronic wave functions by assuming that the interaction energies between distinct, separate bodies decreases as the number of bodies being taken into account increases. The principal hope is that a high-dimensional system depends strongly only on few input variables. In general this assumption is not true, but it suffices in most cases in which relatively weaker interaction energies are considered. Commonly known examples of pair potentials are the Lennard-Jones 6-12 or the Morse potential. Though those potentials can be quickly and easily calculated for arbitrarily systems, they are only good approximations for the simplest closed shell systems. For more complex situations, such as strongly covalent systems like semiconductors, they appear completely inapplicable [64]. A many-body approach for greater body orders is rarely seen due to the rapid increase of the underlying dimension and therefore computational effort.

## 1.1 Using Machine Learning to Approximate the PES

Due to the technical evolution of the past decades, the ability to store an process data has increased rapidly. The name Machine Learning (ML) was coined in 1959 by Arthur Samuel [53], who used a search tree of board positions to teach a computer play the game of checkers. Since then, an increase of available computer memory and speed extended the application possibilities and Machine Learning became a very active research area. Machine learning tasks are typically classified into several broad categories ranging from strict supervised regimes to uncontrolled unsupervised methods. In supervised learning approaches, the problem is learning an input-output mapping from empirical data. Given a data set containing  $n$  values and targets  $\mathcal{D}_n = \{(x_i, y_i) : i = 1, \dots, n\}$ , one tries to predict the output for a new value  $x^*$  as precisely as possible. Introducing a set of possible mappings, the problem of regression is reformulated as an optimization problem based on the given data set. A *fitting* of the function then describes the process of finding the best parameters/coefficients in a prescribed regime. In contrast unsupervised

*Machine Learning*

learning methods are only given input data and attempt to model an underlying regularity. Usual approaches include clustering of the data, or reduction of the dimension of the problem while retaining as much information as possible.

#### MLIP

The earliest efforts to introduce Machine Learning techniques to the approximation of the high-dimensional PES date back to 1995 with the usage of neural networks to describe molecules and small molecule clusters [12]. Another approach to this approximation is the use of is kernel-based methods [54]. The key to its success is the choice of a kernel function, and through it the basis functions employed. In 2010, Bartók *et al.* introduced the Gaussian Approximation Potential (GAP) [5] under the assumption of the atomic decomposition ansatz (1.0.1). They used Gaussian Process Regression (GPR) [50] to approximate the PES. Here, one constructs a Gaussian process uniquely characterized by its covariance function, called the kernel function, onto the space of functions mapping an atomistic system to the real numbers. The GAP is then chosen as the expected function given a set of training data.

#### Active Learning

Nevertheless, machine learning approximations do not only depend on the allowed algebraic form, i.e. on the underlying approximating function space, but also on the training data used to fit them. Usually, machine learning methods improve their accuracy through increasing the number of fitting parameters, making the allowed algebraic form of the estimation more flexible. But during the fitting process, the method is usually not able to influence the choice of the training data; rather learns from what it gets. Unfortunately, those *passive learning* algorithms tend to be interpolative, i.e. they fail to give reasonable results in areas outside their training domain. One attempt is a proper choice of the training domain in a way that minimizes those uncertain areas and ensures interpolation over relevant areas. This makes the choice of an optimal training set to a problem of transferability. Thus, if the method would be able to detect extrapolatory areas and add those to the training set, we would be able to improve the approximative power not only by allowing a more flexible algebraic form, but also by allowing a more flexible choice of the training set. Since the method is actively influencing the selection of the training set, this approach is called *active learning*. An active learner may pose *query strategies*, which describes how to decide which data point to query next to add to the training data. Perhaps the simplest and most commonly used query framework is uncertainty sampling [46]. In this framework, an active learner queries the instances which it is least certain how to label. Another, more theoretically-motivated query selection framework is the query-by-committee (QBC) algorithm [59]. Another decision-theoretic approach aims to measure how much the model's generalization error is likely to be reduced and query the instance with minimal expected future error [51, 73]. Minimizing the expected future directly is expensive and in general not possible in a closed form. However, the variance reduction approach still minimizes the expected future error indirectly. The minimization of the model's variance is well studied in the case of linear regression models, which is in statistical literature known as optimal experimental design (OED) [26]. Here, variance minimization is

equivalent to the maximization of the incorporated Fisher information in the data. There are different approaches which aim to maximize the Fisher information, the most commonly used are the A-optimal, D-optimal [17] or E-optimal [27] approaches.

## 1.2 An adaptive Sparse Grid Approach for Many Body Systems

Usually, machine learning interatomic potentials (MLIPs) are based on a partitioning of the interatomic interaction energy into individual contributions of small groups of atoms. A *fitting* of the potential then describes the process of finding the best parameters/ coefficients describing the potential function based on a previously chosen set of data. In summary, each MLIP has three major components:

- (1) A way to map the physical system to a set of numbers used as the input for the MLIP, coming in the form of a *descriptor*,
- (2) The search set of possible candidates, and
- (3) Training data helping us to choose the right candidate.

In this thesis, we aim to construct a MLIP based on a penalized least squares regression method using a search set based on orthogonal polynomials. Based on that, we will study two approaches to reduce the complexity while preserving the approximation accuracy. First, we will present an optimization algorithm of the search set using an adaptive sparse grid approach. Afterwards, we will perform an optimizing of the training data using the D-optimality approach of the sparse grid literature.

The starting point of our investigation will be the many-body expansion (1.0.2). Assuming that two particles in a system interact only if they are closer than a prescribed threshold, the whole system decomposes in areas which are closed under interaction. The many-body expansion describes a further decomposition of those areas in lower dimensional parts, i.e. into single atoms, pairs of atoms and so on. We could of course let this sum continue up to the full system, but in general we will only consider the first  $K$  sums with  $K \ll M$ , describing the incorporation of all sets of atoms up to size  $K$ . As a first step we will use a descriptor which describes each set of particles by all *pairwise distances* which is invariant under translation and rotation of the three-dimensional space. *Many Body Decomposition*

We have to constrain the search space of possible candidates to a finite dimensional space. This space will be spanned by modified *orthonormal polynomials*. In our considerations we include Legendre, Chebyshev and Laguerre polynomials. Each of these classes is an  $L^2$ -orthonormal base for a specific measure. In order to obtain a physically reasonable search space, we will modify these *Search Set*

polynomials. For example, they should be continuous under particles entering and leaving an area which is closed under interaction. Additionally, the energy potential should assign the same energy to nuclear systems which can be converted into each other by a rotation or translation in space. Nevertheless, the dimension-wise decomposition of the system in (1.0.2) needs to be reflected in the search space as well. To deal with the different dimensional terms of the system, we use *tensors* of the above one-dimensional basis functions. Linear regression using yet another class of basis functions was introduced by Shapeev in 2016 with his Moment Tensor Potentials (MTP) [60]. Connections exist also to the Gaussian approximation potential (GAP) [5] which is equivalent to an linear regression over an infinite dimensional search set. Nevertheless, our approach is based on a more simplified setting, since we use a further decomposition by (1.0.2).

*Sparse  
Tensor  
Product  
Construction*

Beside being a simplification, the many-body expansion also opens up the possibility to treat smaller subsystems in more depth than larger ones. Instead of approximating all terms  $V_2, \dots, V_K$  by the same search set, we are able to vary the accuracy. Following the idea of [34], we will present an adaptive sparse tensor product construction to choose a search space in an optimal way. We start with the coarsest approximation, or equivalently smallest search set possible by only approximating pairs of particles by constants. Progressively, we will enlarge the search space in an optimal way, and retrain the penalized least squares regression model to decide in which direction to refine next. This way, we will be able to greatly reduce the involved degrees of freedom while obtaining the same error as on a *full grid*. Instead of using a machine learning approach, in [34] the energy is exactly calculated by ESM methods.

*Active Learning*

As a third step, we use an active learning approach to choose the training data in an optimal way. Based on a chosen search set and a penalized least squares regression, we are able to define a *information measure* which measures the information one data point carries. This way, we are in the position to sort the available data points by their worth and eventually formulate queries which data point to incorporate next. In a D-optimal approach this is done by choosing the data points in such a way, that the variance of the model is minimized, i.e. such that we are the most certain which approximation to choose. In the case of linear regression, minimizing the model's variance is equivalent to maximizing its Fisher information. Thus, we formulate a query strategy which step by step retrains the model and decides which data point to incorporate next. A similar approach is presented in [49]. Here, the D-optimality approach is formulated in connection with moment tensor potentials (MTP).

The main contributions of this thesis are:

- Modifying orthonormal polynomials to be suitable candidates for potential functions
- Combining the many body decomposition with an Tikhonov regularized least squares regression to find optimal potential function candidates.

## 1.2. AN ADAPTIVE SPARSE GRID APPROACH FOR MANY BODY SYSTEMS 7

- In combination with the least squares method, using an adaptive sparse grid like approach to find an optimal sparser search space.
- Using an Active Learning approach to choose a suitable training set which maximizes its incorporated information.

This thesis is structured as follows. In chapter 2 we will introduce the concept of a penalized least squares regression on a finite dimensional search space. Therefore, we will first formulate the problem in a theoretical manner based on a probability measure and then turn to the case where we are only given a finite number of samples. Afterwards, we will introduce the concept of Tikhonov regularization. We will explain the impact of a regularization onto the search set, which leads to a shrinking of the search set to a bounded ball therein. Also we will explain the one-to-one correspondence of a chosen regularization and the choice of a prior distribution of the weights in a Bayesian setting. We conclude this chapter with a stability result in the general unstable unpenalized case, and perform an error analysis.

In chapter 3 we explain the construction of a sparse tensor product in general terms. We explain how a sparsification of a method can be achieved by introducing a benefit and cost function.

In chapter 4 we introduce the general concept of active learning. We explain the idea behind the variance reduction approach and relate it to other approaches. We explain the D-optimality criterion and its relation to the Fisher information. Moreover, we analyze the approach in depth for the linear regression model.

In chapter 5 we introduce the Born-Oppenheimer potential energy surface in more detail and explain its complexity. We introduce two simplifying assumptions on which the further considerations will be based on, the atomic decomposition and the many-body expansion.

In chapter 6, we apply the mathematical theory of chapters 3 to 4 to the PES. Given a set of data, we explain how to use these concepts to construct a MLIP.

In chapter 7 we present the data sets used in our numerical investigation, including a data set of tungsten and small molecules, and discuss peculiarities we faced in the implementation phase.

The presentation of the numerical results is done in chapter 8.



## Chapter 2

# Penalized Least Squares Regression

In supervised machine learning, regression is used for estimating the relationship among variables. The roots of regression date back to the beginning of the 19th century. To determine the orbits of bodies around the sun using astronomical observations, Legendre (1805) [45] and Gauss (1809) [30] published the earliest form of regression, the method of least squares. The term 'regression' was first coined by Francis Galton in 1886 [29] to describe a biological phenomenon, namely that the heights of descendants of tall ancestors tend to regress down towards a normal average, which is also known as regression toward the mean. Although the term regression for Galton initially only had this biological meaning, it became used as a general term for the study of the relationships between variables. Initial conceptualizations of regression date back to the 19th century, but it was really the technological revolution in the 20th century that started a wave of vast research. On the primal basis of regression laid 200 years ago, many areas have emerged to date, such as support vector machines [54] or artificial neural networks [9].

*History*

In this section we focus on least-squares regression on a finite dimensional search set with Tikhonov regularization. Given a finite number of samples driven by a generally unknown probability measure, we aim to approximate the so-called *regression function* describing its relationship. The problem of finding the optimal approximation translates to a quadratic functional minimization problem over a fixed search set. An application of the Chernoff inequalities for random matrices gives us conditions for the well-posedness of the problem. Under these conditions, we perform an error analysis for the unregularized case.

*Topic*

The section is structured as follows. In subsection 2.1 we derive the optimization problem. First, in 2.1.1 we will generally introduce the least squares regression problem in a theoretical manner with respect to an underlying probability measure. In 2.1.2 we generalize the regression problem to the practical case, where only samples of the probability measure are given in form of a data

*Structure*

set. These first two subsections follow the structure of [23]. In subsection 2.1.3 we will turn to the special case of least squares regression over a finite dimensional search set, and introduce Tikhonov regularization in 2.1.4. In section 2.2 we analyze the one-to-one correspondence between regularization and the choice of the prior distribution in a Bayesian setting. In section 2.3 we represent a probabilistic bound on the unique solvability of the regression problem on a finite dimensional search space in the unregularized case. The stated well-posedness result excerpts from [21, 22], in which the case of orthonormal basis functions is treated and [13, 14], which is a generalization of the prior to general basis sets. We will give an overview of related results in subsection 2.4.

For a comprehensive introduction to the field of regression we refer to [37].

## 2.1 Derivation of the Optimization Problem

In the area of functional regression, we are interested in approximating a  $d$ -dimensional function

$$f_\rho : X \rightarrow Y,$$

where  $X \subset \mathbb{R}^d$  and  $Y$  is a suitable measure space. In the following, the first step will be the definition of those measure spaces. Afterwards,  $f_\rho$  will be defined based on a probability measure  $\rho$  on the product space  $X \times Y$ .

### 2.1.1 Least Squares Regression with given Measure

*Measure  
Spaces*

Let  $(X, \Sigma_{\text{Leb}, X})$  be the measure space with a domain  $X \subset \mathbb{R}^d$  and Lebesgue algebra  $\Sigma_{\text{Leb}, X}$ . The Lebesgue algebra contains all sets in  $X$  which are measurable with respect to the Lebesgue measure. The second measure space is given by an Hilbert space  $(Y, \langle \cdot, \cdot \rangle_Y)$  and the induced Borel  $\sigma$ -algebra  $\Sigma_{\text{Borel}, Y}$ .

*Regression  
Function*

To introduce a certain connection between those spaces, let  $\rho$  be a probability measure on the product space  $X \times Y$  along with its product  $\sigma$ -algebra  $\Sigma_{\text{Leb}, X} \otimes \Sigma_{\text{Borel}, Y}$ . Denote by

$$\rho_X(\cdot) := \rho(\cdot, Y)$$

the corresponding marginal measure on  $X$ . For a random variable  $z = (x, y)$  with law  $\rho$ , we aim to approximate the corresponding  $d$ -dimensional *regression function*,

$$\begin{aligned} f_\rho : (X, \Sigma_{\text{Leb}, X}) &\longrightarrow (Y, \Sigma_{\text{Borel}, Y}) \\ x &\longmapsto \mathbb{E}_\rho(y|x). \end{aligned} \tag{2.1.1}$$

For each  $x \in X$ ,  $f_\rho(x)$  is the average of the  $y$  coordinate of  $\{x\} \times Y$ . From now on we assume the regression function  $f_\rho$  of  $\rho$  to be bounded, i.e. it exists a  $L < \infty$  such that

$$|f_\rho(x)| < L \tag{2.1.2}$$

for  $\rho$  almost every  $x \in X$ . Note that in general we do not know the underlying probability measure  $\rho$ , nevertheless we will assume the knowledge for now



and generalize the approach in the following subsection. The goal is to ‘learn’, i.e. to find the best approximation  $f$  of the regression function  $f_\rho$ .

To quantify the prediction quality of an estimation  $f$ , we need to introduce a notion of an estimation error. Thus we define the error functional by

$$\mathcal{E}(f) := \int_{X \times Y} \|f(x) - y\|_Y^2 d\rho(x, y). \quad (2.1.3)$$

*The  
Error  
Functional*

It measures the mean squared error in the prediction of  $y$ . One can show that  $f_\rho$  is a minimizer of  $\mathcal{E}$  which reassures us that we are picking the right functional to minimize. Formally the following proposition states that the conditional expectation  $f_\rho(x)$  is the  $L^2$ -projection of the random variable  $Y(x, y) = y$  to the subspace of random variables which are measurable w.r.t. their first variable and consequently has to be a function only of  $x$ .

**Proposition 2.1: (Proposition 1.8, [23])**

For every  $f \in L_{\rho_x}^2(X; Y)$  it holds that

$$\mathcal{E}(f) = \mathcal{E}(f_\rho) + \|f - f_\rho\|_{L_{\rho_x}^2(X; Y)}^2, \quad (2.1.4)$$

and  $\mathcal{E}(f_\rho)$  is a notion of the variance of the noise

$$\mathcal{E}(f_\rho) = \text{Var}_\rho(f_\rho(x) - y|x), \quad (2.1.5)$$

which is finite by (2.1.2)

This proposition shows that  $f_\rho$  is the unique minimizer of the the error functional in  $L_{\rho_x}^2$  as for each  $g \neq f_\rho$  the RHS in (2.1.4) is strictly greater than  $\mathcal{E}(f_\rho)$ . By definition  $f_\rho(x)$  is the mean of  $y|x$ . Being a random variable,  $y|x$  can fluctuate around that mean and this fluctuation can be measured as the variance of the random variable. Equation (2.1.5) states that the mean variance is exactly given by  $\mathcal{E}(f_\rho)$  and thus  $\mathcal{E}(f_\rho)$  quantifies how good the values of  $y|x$  will actually be predicted on average by  $f_\rho(x)$ .

It is important to note that minimizing the  $L^2$  norm is a choice. Generally one could also try to minimize

$$\int_{X \times Y} \|f(x) - y\|_Y^p d\rho(x, y). \quad (2.1.6)$$

*Other  
Error  
Functionals*

for other  $p$ . If  $p = 1$  this minimization would result in the conditional median instead of the conditional mean as for  $p = 2$ . For  $p \rightarrow 0$  the optimizer would converge to the conditional mode. This is due to  $\|f(x) - y\|_Y^p$  converging to unity if  $f(x) \neq y$  regardless of how big the actual error  $f(x) - y$  is. So the best solution would be to set  $f(x)$  to the most probable value of  $y$  - the mode of  $y|x$ .

In this thesis we will restrict ourselves the case  $p = 2$ .

In practice there are two problems with the minimization problem (2.1.3):

- (A) The actual distribution of our data  $\rho$  is unknown.
- (B) The space  $L_2(X, Y; \rho_X)$  which we want to optimize over is infinite dimensional.

We deal with these two problems in the following two sections. (A) is treated in section 2.1.2 and section 2.1.3 considers (B).

## 2.1.2 Least Squares Regression with given Samples

*Empirical  
Error  
Functional*

As previously mentioned, in practice we may not have access to either  $\rho$  or any information about  $f_\rho$ . But we are given a data set, which can be regarded as independent samples of  $\rho$ ,

$$\mathcal{D}_n := \{z_i = (x_i, y_i), i = 1, \dots, n\} \subset (X \times Y)^n.$$

Using this, we can define an approximation  $\rho_{\text{empirical}} = \frac{1}{n} \sum_{i=1}^n \delta_{(x_i, y_i)}$ . Under some mild assumptions one can show that if one would have access to an infinite amount of data,  $\rho_{\text{empirical}}$  would converge to  $\rho$  in some sense. Replacing  $\rho$  by the empirical measure  $\rho_{\text{empirical}}$  in (2.1.3), we obtain the *empirical error functional*

$$\mathcal{E}_{\mathcal{D}_n}(f) := \frac{1}{n} \sum_{i=1}^n \|f(x_i) - y_i\|_Y^2. \quad (2.1.7)$$

Analogously to the convergence of  $\rho_{\text{empirical}}$  to  $\rho$ , under mild assumptions and an infinite amount of data  $\mathcal{E}_{\mathcal{D}_n}$  would converge to  $\mathcal{E}$ . When taking  $(x_i, y_i)$  as independent samples from  $\rho$ , (2.1.7) takes the form of a Monte Carlo approximation of (2.1.3). Thus, bounds for the accuracy of this estimation will come in terms of lower bounds on the number of samples  $n$ .

Our problem is now to solve the *empirical regression problem*:

*Empirical  
Regression  
Problem*

$$\text{Find } f_{\mathcal{D}_n} := \underset{f \in L_{\rho_X}^2(X; Y)}{\text{argmin}} \mathcal{E}_{\mathcal{D}_n}(f). \quad (2.1.8)$$

An important difference between minimizing (2.1.3) and (2.1.7) is that (2.1.7) depends on point evaluations of  $d$ -dimensional functions in  $L_{\rho_X}^2(X; Y)$ . However, point evaluation is not well-defined for all of these functions. To ensure well-definedness we need some additional smoothness properties which can for example be obtained by requiring the function to be in a Sobolev space. By Sobolev's embedding theorem, the Sobolev spaces  $H^\beta$  for  $\beta \geq \frac{d}{2}$  are continuously embedded in the continuous functions  $C^0(\mathbb{R}^d)$ , for which point evaluations are well-defined (note that  $\beta$  grows with the dimension of the space). In the next section we will restrict the search set to a space which only contains continuous functions.

### 2.1.3 Least Squares Regression on a finite dimensional Search Set

To overcome the problem of well-defined point evaluations and the existence of an unique solution of the empirical regression problem (2.1.8), we restricted ourselves to a finite dimensional search set. Let  $V_k \subset L^2_{\rho_X}(X; Y)$  be a  $N_k$  dimensional function space with a given basis set  $\{\phi_1, \dots, \phi_{N_k}\}$ ,

$$V_k := \text{span}\{\phi_1, \dots, \phi_{N_k}\}.$$

Here, the subscript  $k \in \mathbb{N}$  is a notion for the approximal power or 'finess' of the function space. One can think of a finite element approximation, where  $k$  would denote the grade of fineness of the underlying grid and  $N_k$  would denote the number of corresponding grid points or basis functions. To ensure the existence of point evaluations, we further assume that the function space  $V_k$  can be continuously embedded in the continuous functions  $C(X; Y)$ . The regression problem constrained to the search set  $V_k$  reads

$$\text{Find } f_{\mathcal{D}_n, V_k} := \underset{f \in V_k}{\text{argmin}} \mathcal{E}_{\mathcal{D}_n}(f). \quad (2.1.9)$$

Due to the finite dimensionality of the search set, we are now able to present the problem by a system of equations. Since  $f_{\mathcal{D}_n, V_k} \in V_k$ , there exist coefficients  $c \in \mathbb{R}^{N_k}$  such that  $f_{\mathcal{D}_n, V_k} = \sum_{j=1}^{N_k} c_j \phi_j$ . Thus, we formulate an equivalent problem, which searches the optimal coefficients instead of the optimal function,

$$\begin{aligned} c_{\mathcal{D}_n, V_k} &:= \underset{c \in \mathbb{R}^{N_k}}{\text{argmin}} \frac{1}{n} \sum_{i=1}^n \left\| \sum_{j=1}^{N_k} c_j \phi_j(x_i) - y_i \right\|_Y^2 \\ &= \underset{c \in \mathbb{R}^{N_k}}{\text{argmin}} \frac{1}{n} \|A \cdot c - y\|_{Y^n}^2 \end{aligned} \quad (2.1.10)$$

where

$$A = (\phi_j(x_i))_{\substack{i=1, \dots, n \\ j=1, \dots, N_k}} \in \mathbb{R}^{n \times N_k} \quad (2.1.11)$$

and  $y = (y_1, \dots, y_n)^T$ .

### 2.1.4 Penalized Least Squares Regression

To ensure the existence of an unique solution which, we introduce a *regularization term*. For a norm  $\|\cdot\|_{\Gamma}$  on the search space  $V_k$ , we formulate the corresponding penalized regression problem by

$$f_{\mathcal{D}_n, V_k, \Gamma} := \underset{f \in V_k}{\text{argmin}} \mathcal{E}_{\mathcal{D}_n}(f) + \|f\|_{\Gamma}. \quad (2.1.12)$$

If there exists a matrix  $\Gamma \in \mathbb{R}^{N_k \times N_k}$  such that the prescribed norm can be written in terms of the  $L^2$  norm, the matrix is called a *Tikhonov matrix* and the corresponding regularization is termed *Tikhonov regularization*. Assume the existence of such a matrix for the prescribed norm. For all  $f \in V_k$  let

$$\|f\|_{\Gamma} = \|\Gamma \cdot c\|_{L^2(X; Y)}, \quad (2.1.13)$$

where  $f = \Phi \mathbf{c}$  for a coefficient vector  $\mathbf{c} \in \mathbb{R}^{N_k}$  and  $\Phi = (\phi_1, \dots, \phi_{N_k})$ . This way we are only considering situations where the norm of a function is equivalent to the  $L^2$  norm of the function after changing the basis with respect to  $\Gamma$ .

**Example 2.1** (Tikhonov Matrix of a Sobolev Norm Regularization). Let  $\phi_1, \dots, \phi_k \in L^2_{\rho_X}(X; Y)$  be linearly independent basis functions and

$$V_k := \text{span}\{\phi_1, \dots, \phi_k\}$$

the considered search space which we assume to be closed under differentiation, i.e. for  $\Phi := (\phi_1, \dots, \phi_k)$ , there exists a matrix  $D \in \mathbb{R}^{k \times k}$  such that  $\nabla \Phi = D \cdot \Phi$ . Take a function  $f = \sum_{i=1}^k c_i \phi_i = \Phi \mathbf{c} \in V_k$  and  $s \in \mathbb{N}$ . Then, the  $H^s$  Sobolev norm of  $f$  can be written as

$$\begin{aligned} \|f\|_{H^s}^2 &= \sum_{k=0}^s \|\nabla^k f\|_2^2 = \sum_{k=0}^s \|\nabla^k \Phi \mathbf{c}\|_2^2 = \sum_{k=0}^s \|D^k \Phi \mathbf{c}\|_2^2 \\ &= \mathbf{c}^T (M + D^T M D + \dots + D^T M^s D) \mathbf{c}, \end{aligned}$$

where  $M := (\langle \phi_i, \phi_j \rangle_{L^2(X)})_{i,j}$  denotes the mass matrix based on the basis set of  $V_k$ . This norm is equivalent to the expression

$$\|f\|_{H^s}^2 = \mathbf{c}^T (M + D M D^T)^s \mathbf{c}.$$

Thus, we can write

$$\|f\|_{H^s}^2 = \|\Gamma \mathbf{c}\|_2^2 \tag{2.1.14}$$

with the matrix

$$\Gamma = (M + D^T M D)^{s/2},$$

which is called the corresponding Tikhonov matrix of the Sobolev norm.  $\Gamma$  is well-defined, since  $(M + D^T M D)$  is positive semi-definite. Therefore, we are also able to define Sobolev norms with real valued order  $s \in \mathbb{R}$  by (2.1.14).

**Example 2.2** (Sobolev Norm Regularization and a Fourier basis). Let  $X = [0, 2\pi]$  and  $Y = \mathbb{C}$ . Consider the basis functions  $\phi_t(x) := e^{itx}$  spanning the search space  $V_k$  by

$$V_k := \text{span}\{\phi_0, \phi_1, \dots, \phi_k, \phi_{-1}, \dots, \phi_{-k}\}.$$

The basis functions are pairwise orthonormal with respect to the scalar product  $\langle \cdot, \cdot \rangle_{L^2(X)}$  and thus  $M = I$ . Moreover it holds that  $D = \text{diag}(1, \dots, k)$ . If  $f \in V_k$ , then  $f$  can be written as linear combination with respect to the basis functions,  $f = \Phi \mathbf{c} \in V_k$ . Analogously to the above example  $s \in \mathbb{R}$ , the  $H^s$  Sobolev norm can be written as

$$\|f\|_{H^s}^2 = \|\Gamma \mathbf{c}\|_2^2,$$

where  $\Gamma = \text{diag}((1 + 1^2)^{s/2}, \dots, (1 + k^2)^{s/2})$ .

Let  $\Gamma$  be a Tikhonov matrix in the sense of (2.1.13). The penalized least squares regression problem (2.1.12) can equivalently be written as

$$\text{Find } c_{\mathcal{D}_n, V_k, \Gamma} := \operatorname{argmin}_{c \in \mathbb{R}^{N_k}} \frac{1}{n} \|Ac - y\|_{Y^n}^2 + \|\Gamma c\|_2^2. \quad (2.1.15)$$

Obviously, for  $\Gamma$  equal to the zero matrix we are back in the unpenalized case. The solution of the least squares problem with Tikhonov regularization can be computed by solving a  $N_k \times N_k$  linear system:

*Normal  
Equations*

### Lemma 2.1: Normal Equations

Let  $(\phi_1, \dots, \phi_{N_k})$  be an arbitrary basis of  $V_k$  and  $\|\cdot\|_Y = \|\cdot\|_2$ . Then the normal equations for the least squares problem with Tikhonov regularization with respect to a matrix  $\Gamma$  in (2.1.15) can be written as

$$(A^T A + \Gamma^T \Gamma) \hat{c} = A^T y, \quad (2.1.16)$$

where  $A$  is given as in (2.1.11) and  $y = (y_1, \dots, y_n)^T \in \mathbb{R}^n$ .

*Proof.* The objective is to minimize

$$\begin{aligned} S(c) &:= \|Ac - y\|_2^2 + \|\Gamma c\|_2^2 \\ &= (Ac - y)^T (Ac - y) + (\Gamma c)^T (\Gamma c) \\ &= c^T A^T A c - y^T A c - c^T A^T y + y^T y + c^T \Gamma^T \Gamma c \\ &= c^T A^T A c - 2y^T A c + y^T y + c^T \Gamma^T \Gamma c. \end{aligned}$$

The last equality holds, since  $y^T A c \in \mathbb{R}$  and in one dimension the transpose of a value is the value itself. Differentiating  $S(c)$  with respect to  $c$  gives

$$\frac{\partial S(c)}{\partial c} = 2c^T A^T A - 2y^T A + 2c^T \Gamma^T \Gamma.$$

Equating to zero to satisfy the first-order conditions results in the normal equations

$$(A^T A + \Gamma^T \Gamma) \hat{c} = A^T y.$$

□

The additive Tikhonov penalization is a usual way to obtain existence and uniqueness of the minimizer, since an invertible Tikhonov matrix can make the matrix on the LHS of (2.1.16) invertible. Another way to understand the Tikhonov regularization is that we are shrinking the search set of (2.1.9) to a bounded ball in  $V_k$  with respect to a norm  $\|\cdot\|_\Gamma$  instead of minimizing over the whole space  $V_k$ . Denote by  $\|f\|_\Gamma := \|\Gamma c\|_2$  the induced norm with respect to the Tikhonov matrix  $\Gamma$ , where  $c$  are the coefficients of  $f$  with respect to the basis set  $\phi_1, \dots, \phi_{N_k}$ . When minimizing over a closed ball of  $V_k$ , we obtain existence of a minimizer by an application of the generalized Weierstrass theorem, see e.g. [70]. Here, we use the fact that the ball is convex and closed and that

*Tikhonov  
Penalization  
and a Ball in  $V_k$*

the minimization problem is convex and sequentially lower semicontinuous. Nevertheless we may get several solutions of the problem. But there exists an unique solution with the least  $\|\cdot\|_\Gamma$  norm. When considering the Lagrange formulation of the constrained regression problem over a bounded ball and least norm, we end up with (2.1.15). Thus, the larger the entries of the penalization matrix  $\Gamma$  become, the more rigorous the induced norm gets and the fewer functions of  $V_k$  we will allow. When  $\Gamma = 0$ , the problem is equivalent to the unpenalized regression problem and hence to the minimization over the whole space  $V_k$ . For the unpenalized case, the existence and uniqueness of the problem is not assured. In the next subsection we investigate under which assumption, we can assume existence and uniqueness of a solution, even in the unpenalized case.

*Review*

Let us recap our progress so far. In subsection 2.1.1 we defined the regression function  $f_\rho : X \rightarrow Y$  by  $f_\rho(x) := \mathbb{E}_\rho[y|x]$  with respect to a given probability measure  $\rho$  on  $X \times Y$ . To approximate the regression function, we defined the general regression problem by

$$\hat{f} := \operatorname{argmin}_{f \in L_{\rho_X^2}(X;Y)} \int_{X \times Y} \|f(x) - y\|_Y^2 d\rho(x, y).$$

Since we do not have more information about  $\rho$  other than a finite number of samples  $\mathcal{D}_n = \{(x_i, y_i), i = 1, \dots, n\}$  we defined in subsection 2.1.2 the empirical regression problem given by the empirical measure instead of  $\rho$ ,

$$f_{\mathcal{D}_n} := \operatorname{argmin}_{f \in L_{\rho_X^2}(X;Y)} \frac{1}{n} \sum_{i=1}^n \|f(x_i) - y_i\|_Y^2.$$

To ensure well-posedness and well-definedness of the empirical regression problem, we restrict ourself in subsection 2.1.3 to a finite dimensional search set  $V_k := \operatorname{span}\{\phi_1, \dots, \phi_{N_k}\} \subset L_{\rho_X^2}(X;Y)$ . We assumed the search set to be continuously embedded in the continuous functions in order to ensure the existence of point evaluations. The empirical regression problem on  $V_k$  reads

$$f_{\mathcal{D}_n, V_k} := \operatorname{argmin}_{f \in V_k} \frac{1}{n} \sum_{i=1}^n \|f(x_i) - y_i\|_Y^2.$$

Since  $f_{\mathcal{D}_n, V_k} = \sum_{i=1}^{N_k} \hat{c}_i \phi_i$  we can rewrite the above as

$$\hat{c} = \operatorname{argmin}_{c \in \mathbb{R}^{N_k}} \|Ac - y\|_{Y^n}^2,$$

with  $A$  as in (2.1.11). In subsection 2.1.4 we introduced a Tikhonov regularization to ensure the solvability of the empirical regression problem. We did this by adding a penalization term  $\|\Gamma c\|_2^2$ . Furthermore we proved that the solution of the minimization problem is now given by the solution of the normal

equations 2.1.

Outlook

We also saw one interpretation of how Tikhonov regularization works. There is another way to interpret such a regularization, which we will treat in the following section.

## 2.2 Bayesian Interpretation

In the previous section, we introduced least squares regression with an regularization term. As the choice of the norm of the error functional has a further statistical meaning, so has the regularization term. In the following, we will give a short introduction to Bayesian statistics so that we can interpret the Tikhonov regularization in that setting. A very short summary of how to treat an inference problem in the Bayesian way is as follows:

Bayesian  
Statistics

1. First, we make some assumptions on a parametrized model which underlies the generation of samples. We assume that  $x$  is deterministic, so what we need to model is  $y|x, w$ . That is, given a parameter  $w$  for our model and an "input"  $x$ , how will  $y$  be distributed?
2. We model  $w \in \mathbb{R}^m$  as a random variable and assume an underlying distribution. For example, we could assume that  $w \sim \mathcal{N}(0, C)$  with a covariance matrix  $C$ .
3. We need to assemble independent samples  $(x_i, y_i)$ . Here, we ignore the distribution of  $x$  for simplicity.
4. Since an exact model of  $y|x$  is given, we can calculate how probable the outcome  $(x_i, y_i)$  was. This probability distribution is called the *likelihood*  $p(y|w, x)$ .
5. Using Bayes' formula, one calculates the *posterior probability*  $p(w|y, x)$  of  $w$ .

We will now walk through the above steps in the special case of *Bayesian linear regression* in one dimension. We assume that  $y$  is distributed as

$$y = f_w(x) + \epsilon \sim \mathcal{N}(f_w(x), \sigma^2),$$

Bayesian  
Linear  
Regression

where  $\epsilon \sim \mathcal{N}(0, \sigma^2)$ . We assume  $f_w$  to be given by

$$f_w(x) = w^T x.$$

Thus, we assume that when we know  $x$ , then  $y$  is given as a linear function of  $x$  with an added random term  $\epsilon$ . This concludes step 1. For step 2 we assume that  $w \sim \mathcal{N}(0, C)$  for some covariance matrix  $C$ . The choice of this prior distribution is a crucial one and should reflect our assumptions on  $w$ .

In the following, given samples from our distribution, we want to calculate the

Likelihood

likelihood distribution. We assume that, given the matrix  $A$ , where  $A_{i,j} = (x_i)_j$  is the  $j$ th entry of the  $i$ th training example. We also assume that  $y_i$  are conditionally independent given  $A, w$ . Therefore we can write the probability of all the samples occurring as the product of the probabilities of all the samples  $(x_i, y_i)$ :

$$\begin{aligned} p(y|A, w) &= \prod_{i=1}^n p(y_i|x_i, w) \\ &= \prod_{i=1}^n \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - w^T x_i)^2}{2\sigma^2}\right) \\ &= \mathcal{N}(A^T w, \sigma^2 I). \end{aligned}$$

*Maximum Likelihood Estimator*

One way, other than the Bayesian way, to proceed from here would be to maximize this likelihood. This corresponds to the idea that we want to choose the parameter  $w$  which fits our data the best, and consequently choose  $w$  to maximize the probability of our data appearing. In this case, the estimator would be called the maximum likelihood estimator (MLE):

$$\begin{aligned} w_{MLE} &= \max_w p(y|A, w) \\ &= \min_w -\log(p(y|A, w)) \\ &= \min_w \sum_{i=1}^n (w^T x_i - y_i)^2. \end{aligned}$$

Therefore, maximizing the likelihood corresponds to minimizing the mean squared error without any regularization. When picking the parameter that fits the data the best, this estimator often overfits. Overfitting means that the parameter is influenced by the noise in the data one used to calculate it. Instead of being good at generally predicting  $y|x$  it is good at predicting  $y_i$  given  $x_i$  for our few training examples. This is not very surprising as we obtained the estimator by optimizing its ability to predict the training data  $(x_i, y_i)$ .

*Bayesian Interpretation of an Unpenalized Regime*

Therefore, we can interpret what it means to solve the unregularized regression problem, assuming that it is solvable. Likewise, we assume a additive white noise and pick the parameter  $c$  that makes the observed data as likely as possible.

*Bayes' Theorem*

Let us return to the Bayesian setting and remember that we are now at step 5. Instead of only maximizing the likelihood of step 4, we combine it to get the posterior distribution using Bayes' theorem,

$$p(w|y, x) = \frac{p(y|w, x)p(w)}{p(y|x)}.$$

Here, we omit the treatment of  $x$  on the RHS to simplify things. Since we know everything on the right hand side, we can exactly calculate the left hand side and get that the posterior distribution of the weight vector again has Gaussian distribution,

$$w|y, x \sim \mathcal{N}(\mu, \Sigma)$$



with expectation

$$\mu = (A^T A + \sigma^2 C^{-1})^{-1} (A^T \mathbf{y})$$

and covariance matrix

$$\Sigma = \left( \frac{1}{\sigma^2} A^T A + C^{-1} \right)^{-1},$$

see for example [9], section 2.3.3. This posterior distribution expresses our belief of  $w$ , after we incorporated the data into our prior beliefs which were given through the prior distribution. This is again a distribution over different values of  $w$ . One way to make a prediction would be to take the mean  $\mu$  of this distribution.

Taking a closer look at  $\mu$ , we see that if we set  $\Gamma^T \Gamma = \sigma^2 C^{-1}$ ,  $\mu$  will exactly be the solution of our normal equation. So, we can interpret what the Tikhonov regularization is doing in this framework. Instead of just maximizing the likelihood as before, we now assumed that the prior distribution of  $\mathbf{c}$  is given by  $\mathcal{N}(0, \frac{1}{\sigma^2} (\Gamma^T \Gamma)^{-1})$  and then use the mean of the posterior distribution  $\mu$  as an estimator. The covariance matrix gives us a description of the assumptions we are making: In which directions do we have a big or small prior variance (translating to not so much and much prior certainty about the parameter value)? What are the covariances between the different components of the vector (how much do we think these are coupled to each other)? Knowing that using a Tikhonov regularization with Tikhonov matrix  $\Gamma$  is the same as assuming a prior distribution with covariance  $\frac{1}{\sigma^2} (\Gamma^T \Gamma)^{-1}$  gives us a new way to think about which assumptions we are making. Also, one can use it to analyze which assumptions will make your posterior broader or narrower. This tells us how well we think  $\mu$  will actually work as an estimator 'most of the time' as the covariance matrix tells us how much  $w$  will typically deviate from its mean.

*Bayesian  
Interpretation of  
a Tikhonov  
Regularization*

We with a few general remarks about Bayesian statistics. We chose both the prior and the likelihood as Gaussian distributions and developed a corresponding posterior which was also Gaussian distributed. In this case, we were even in the position to write down the mean and covariance of the posterior in a closed form. Given a likelihood, a prior is said to be a *conjugate* prior if the posterior will belong to the same class as the prior distribution. In that case, the posterior is often given in a closed form. In general, the posterior could be very complicated, and more involved methods such as Markov Chain Monte Carlo are necessary to sample the posterior measure and obtain its mean. Sometimes, one can also use the *maximum a posteriori* (MAP) estimator which corresponds to a maximization of the posterior distribution. This gives a different result than directly maximizing the likelihood which leads to the form of the MLE. In the presented Gaussian case,  $\mu$  is not only the mean of the distribution but also the maximum a posteriori estimate, since the mode and the mean of a Gaussian random variable coincide. It is also interesting to note that if we wanted to model  $L^1$  regularization in a Bayesian setting we would take the same likelihood as here, but take a Laplace prior. The MAP estimator would then correspond to an  $L^1$  regularization. For a comprehensive overview of the

*Remarks*

Bayesian approach, we refer to [50].

### 2.3 Stability Analysis in the Unpenalized Case

We saw previously, that the solution of the least squares regression problem with a Tikhonov regularization on a finite dimensional search set

$$V_k := \text{span}(\{\phi_1, \dots, \phi_{N_k}\})$$

is given by the solution of the corresponding normal equations,

$$G \cdot \hat{c} = A^T y, \quad (2.3.1)$$

where  $G := (A^T A + \Gamma^T \Gamma)$  with respect to a given Tikhonov matrix  $\Gamma$ . The regression problem is therefore uniquely solvable if the matrix  $G$  is invertible. In the penalized case, this is already ensured if the matrix  $\Gamma$  is invertible. In the unpenalized case,  $\Gamma = 0$ , the solvability of the normal equations is more interesting. But, it can be shown that if the number of samples are strictly greater than a lower bound which depends on the search set  $V_k$ , the normal equations are solvable with an high probability. Although we consider the existence of a penalization term in our setting, we will sketch the idea of the proof.

*Chernoff  
Inequality*

In the following we assume the unpenalized case, i.e.  $G = A^T A$ . Since the matrix  $A$  and thus the solvability of (2.3.1) depends on the specific choice of the samples  $x_i \sim \rho_X$ , we are receiving a probabilistic bound with respect to  $\rho_X$  on the solvability. Therefore, we will have to make sure that the number of samples and the characteristic of the search set are in a certain relationship to each other. In [21,22] solely the case of  $\rho_X$ -orthonormal basis sets are treated. In this case the knowledge of the dimension of  $V_k$  was sufficient to check the well-posedness condition. For arbitrary basis sets we have to define a more general characteristic. While [21,22] characterizes  $V_k$  with a notion  $K(N_k)$ , following [14] we define in the general case,

$$K(\phi_1, \dots, \phi_{N_k}) := \sup_{x \in X} \sum_{i=1}^{N_k} |\phi_i(x)|^2.$$

Additionally, we denote in the following by

$$M := \left( \langle \phi_i, \phi_j \rangle_{\rho_X} \right)_{i,j=1, \dots, N_k}$$

the mass matrix with respect to the prescribed basis functions with respect to the scalar product induced by  $\rho_X$ . Note that for an  $\rho_X$ - orthonormal basis set, the mass matrix would be equal to the identity matrix. The following theorem will tell us that given a  $\theta > 0$  and sufficient samples, then the unpenalized least squares regression problem is uniquely solvable.

**Theorem 2.1: Well-Posedness**

Let  $n \geq N_k$  and for some  $\theta > 0$  let

$$K(\phi_1, \dots, \phi_{N_k}) \leq c \cdot \frac{n}{\log(n)} \cdot \frac{\lambda_{\min}(M)}{1 + \theta}, \quad (2.3.2)$$

where  $c = |\log(\frac{e^{0.5}}{(1.5)^{1.5}})| > 0$ . Then, with probability at least  $1 - 2n^{-\theta}$  the solution of the unpenalized regression problem exists and is unique. More specific, we get

$$\mathbb{P} \left( \frac{1}{2} \cdot \mu_{\min} < \lambda_{\min} \left( \frac{1}{n} G \right) \leq \lambda_{\max} \left( \frac{1}{n} G \right) < \frac{3}{2} \cdot \mu_{\max} \right) \geq 1 - 2n^{-\theta},$$

with  $\mu_{\min} = \lambda_{\min}(M)$  and  $\mu_{\max} = \lambda_{\max}(M)$ .

Under the additional assumption that the basis is orthonormal, i.e.  $M = I$ , we are in the special case of [21]. The key idea of the proof is to interpret the left hand side, the matrix  $G$ , of the normal equation (2.3.1) as the sum of random matrices rather than as evaluations of random samples. With that, the proof is based on an application of the Chernoff inequalities. Given a sequence of independent, equally distributed random matrices, the Chernoff inequalities give us an upper bound for the probability that the sum of the sequence deviates from its mean. In the following, we denote by  $\lambda_{\min}(R)$  the minimal and by  $\lambda_{\max}(R)$  the maximal eigenvalue for a matrix  $R \in \mathbb{R}^{m \times m}$ .

**Theorem 2.2: ([67], p.417.) Chernoff Inequality for Random Matrices**

Consider a finite sequence  $\{R_i\}$  of independent, random, self-adjoint, positive semidefinite,  $m \times m$  matrices satisfying

$$\lambda_{\max}(R_i) \leq L.$$

Define

$$\mu_{\min} := \lambda_{\min} \left( \sum_{i=1}^n \mathbb{E}(R_i) \right) \quad \text{and} \quad \mu_{\max} := \lambda_{\max} \left( \sum_{i=1}^n \mathbb{E}(R_i) \right).$$

almost surely. Then

$$\mathbb{P} \left\{ \lambda_{\min} \left( \sum_{i=1}^n R_i \right) \leq (1 - \delta) \mu_{\min} \right\} \leq \left[ \frac{e^{-\delta}}{(1 - \delta)^{1-\delta}} \right]^{\mu_{\min}/L} \quad \text{for } \delta \in [0, 1]$$

and

$$\mathbb{P} \left\{ \lambda_{\max} \left( \sum_{i=1}^n R_i \right) \geq (1 + \delta) \mu_{\max} \right\} \leq \left[ \frac{e^{\delta}}{(1 + \delta)^{1+\delta}} \right]^{\mu_{\max}/L} \quad \text{for } \delta \geq 0.$$

Therefore, if we are able to write  $G$  as a sum of a finite sequence of random

matrices fulfilling the assumptions of the Chernoff inequality, we are able to bound the probability of  $G$  to be smaller than  $(1 - \delta)\mu_{\min}$  or  $(1 - \delta)\mu_{\max}$ . The proof of theorem 2.1 works along this lines.

### Theorem 2.3: Stability of the Solution

Assume the requirements of the previous Theorem 2.1 hold true. Then, the solution  $f_{\mathcal{D}_n, V_k, \Gamma} = \sum_{j=1}^{N_k} \hat{c}_j \phi_j$  of (2.3.1) fulfills

$$\|f_{\mathcal{D}_n, V_k, \Gamma}\|_{L_{\rho_X}^2(X; Y)} < \frac{1}{\sqrt{n}} \cdot \frac{\sqrt{6}\lambda_{\max}(M)}{\lambda_{\min}(M)} \cdot \|y\|_{l_2}$$

with probability at least  $1 - 2n^{-\theta}$ .

This theorem has a very short proof by basically writing everything out and using theorem 2.1. The following theorem gives an upper probabilistic error bound on the overall error. Let  $\omega > 0$ . We define the truncation operator  $\tau_\omega : L_{\rho_X}^\infty(X; Y) \rightarrow L_{\rho_X}^\infty(X; Y)$  by  $\tau_\omega = P_\omega \circ f$ , where  $P_\omega : Y \rightarrow Y$  is given by

$$P_\omega(x) := \begin{cases} x & , \text{if } \|x\|_Y < \omega \\ \text{sign}(x) \cdot \omega & , \text{otherwise.} \end{cases} \quad (2.3.3)$$

### Theorem 2.4: Expected regression error

Take  $\Gamma = 0$ . Let  $n \geq N_k$  and let the assumption of Theorem 2.1 hold, i.e.

$$K(\phi_1, \dots, \phi_{N_k}) \leq c \cdot \frac{n}{\log(n)} \cdot \frac{\lambda_{\min}(M)}{1 + \theta} \quad (2.3.4)$$

where  $c = |\log(\frac{e^{0.5}}{(1.5)^{1.5}})| \approx 0.04696 > 0$  and some  $\theta > 0$ . Then

$$\mathbb{E}_{\rho_X^n} [\mathcal{E}(\tau_r \circ f_{\mathcal{D}_n, V_k}) - \mathcal{E}(\hat{f})] \leq (1 + \varepsilon(n)) \inf_{f \in V_k} \|f - \hat{f}\|_{L_{\rho_X}^2(X; Y)} + 8r^2 n^{-\theta}, \quad (2.3.5)$$

where  $\varepsilon(n) := \frac{4|\log(c)|}{1 + \theta} \cdot \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)} \cdot \frac{1}{\log(n)}$  which tends to zero for  $n \rightarrow +\infty$ , if the condition number of the mass matrix  $\kappa(M) = \frac{\lambda_{\max}(M)}{\lambda_{\min}(M)}$  is bounded. Here, the expectation is taken with respect to the product measure  $\rho^n = \rho \times \dots \times \rho$ .

**Example 2.3.** • Trigonometric polynomials and uniform measure. Take  $X = [-\pi, \pi]$  and consider for odd  $m = 2p + 1$  the space  $V_m$  spanned by trigonometric polynomials of degree  $p$

$$L_k(x) := e^{ikx}, \text{ for } k = -p, \dots, p.$$

Assuming  $\rho_X$  to be the uniform measure, this is an orthonormal basis with respect to  $L^2(X, \rho_X)$ . In this example,

$$K(L_{-p}(x), \dots, L_p(x)) = \sup_{x \in [-\pi, \pi]} \sum_{k=-p}^p |\phi_k(x)|^2 = \sum_{k=-p}^p 1 = 2p + 1 = m.$$

Therefore, the expected regression error of an unpenalized least squares regression based on  $n$  randomly drawn sample points can be bounded in the sense of theorem 2.4, if

$$m \sim \frac{n}{\log(n)}.$$

- Legendre polynomials and uniform measure. Let  $X = [-1, 1]$  and consider  $V_m = \mathbb{P}_{m-1}$ , the space of algebraic polynomials of degree  $m - 1$ . Let be  $\rho_X$  the uniform measure. Then the Legendre polynomials  $(L_k)_k$  is an orthogonal basis with respect to  $L^2(X, \rho_X)$  and

$$\|L_k\|_{L^\infty([-1,1])} = |L_k(1)| = \sqrt{2k - 1},$$

and thus

$$K(L_0, \dots, L_{m-1}) = \sup_{x \in [-1,1]} \sum_{k=0}^{m-1} |\phi_k(x)|^2 = \sum_{k=0}^{m-1} (2k + 1) = m^2.$$

Therefore, the expected regression error of an unpenalized least squares regression based on  $n$  randomly drawn samples can be bounded in the sense of theorem 2.4, if

$$m \sim \sqrt{\frac{n}{\log(n)}}.$$

## 2.4 Related Work

The ideas described in this section can be developed for a more general class of learning algorithms known as *empirical risk minimization* (ERM) or *structural risk minimization* algorithms [61]. Here, the least squares error functional we introduced in (2.1.3) is replaced by an error functional dependend on a *loss function*  $\psi : Y \times Y \rightarrow [0, \infty)$ , which is a notion of distance in the 'image space'  $Y$  and therefore must fulfill that  $\psi(y, y) = 0$  for every  $y \in Y$ . Given a loss function, the general error functional is defined by

$$\mathcal{E}^\psi(f) := \int_{X \times Y} \psi(f(x), y) d\rho(x, y).$$

A simple example for  $\psi$  would be a metric; however many commonly used loss functions are not metrics, for example the triangle inequality does not hold [54, 68]. With  $\psi$  equal to the squared Euclidean distance, we are back in the previously introduced case. In subsection 2.1.3 we restricted ourself to a

finite dimensional search set, nevertheless in [23] a general setting for infinite dimensional Banach spaces is treated. Considering a Banach space  $(H, \|\cdot\|_H)$  which is continuously embedded in the continuous functions and restricting ourself only to functions which satisfy a upper norm bound, then existence of a solution can be proven for strictly convex loss functions with the help of the generalized Weierstrass theorem. Moreover, there exists a unique solution of the problem with the least norm in  $H$ . Assuming a boundedness for the loss function (*M-boundedness*), error bounds for the corresponding overall error, bias and sampling error are derived. If  $H$  an Hilbert space, then it follows by the continuous embedding into the continuous functions, that it is also a *reproducing kernel Hilbert space* (RKHS). In this case we can represent the solution according to the *representer theorem* with the help of the kernel function evaluated at the sample points. This allows us to break the problem down to finite dimensions [54, 69]. Unregularized least squares regression on finite dimensional search spaces with orthogonal basis functions in the noise-free and noisy case are extensively observed in [21] and lead to inherently better rates than the general analysis for Banach spaces. Here, the *M-boundedness* of the loss function is substituted by considering the *truncated least squares regression*, as we considered in the previous subsection. In [22] the approach is generalized to *weighted least squares regression* in the noisy and noise-free case. A generalization of the statements to general basis functions can be found in [13, 14].

## Chapter 3

# Sparse Tensor Product Construction

In a variety of applications, such as interpolation, solving of PDEs or machine learning, one aims to find a best approximation of an unknown function within the possibilities. In the previous chapter we thematized the case of a least squares regression. Therefore, we established that the approximation accuracy crucially depends on the relation of the number of data points and the dimension of the search space. As the dimension of the unknown function grows, the complexity of the problem usually increases exponentially, sometimes even making problems in three dimensions already infeasible. This phenomenon is associated with the term *curse of dimensions*, which is first named in 1961 [8]. Since then, circumventing the curse of dimensions has been a much studied field and usually requires additional information about the function to be approximated. In this section we will describe a method using a *tensor-based subset splitting*, which can break the curse of dimensionality in certain situations. First, the high-dimensional problem is split into lower-dimensional sub-problems. If the function meets certain smoothness assumptions, it is enough to take only a portion of the sub-problems into account. This way, a greatly reduced complexity of the problem can be achieved with only a slightly reduced accuracy of the approximation. For a short introduction in the theory of Sparse Grids we refer to [32], for a comprehensive treatment of the topic see [16].

*Curse of  
Dimensions*

This chapter is structured as follows. In section 3.1 we will present the general idea of the sparse tensor product construction. In section 3.2 we thematize the optimal way to sparsen the function space, in order to maximize the benefit-cost ratio under a given workload. In section 3.3 we give an overview over already developed optimal sparse function spaces under specific smoothness assumptions.

*Structure*

### 3.1 A Tensor Product Construction of the Function Space

To open up the possibility to sparsen the function space, we have to ensure that the underlying function space is constructed by a tensor product. In subsec-

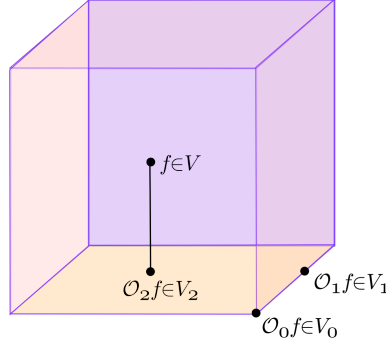


Figure 3.1: The first three elements of a nested sequence of linear subspaces  $(V_n)_{n \in \mathbb{N}_0}$ ,  $V_0 \subset V_1 \subset V_2$  and its corresponding sequence of operators  $\mathcal{O}_n : V \rightarrow V_n$ .

tion 3.1.1 we thematize the case of one dimensional functions and generalize in subsection 3.1.2 to the multi dimensional case. For a prescribed approximation operator, which maps a function onto its approximation in the so-constructed function space, we will introduce in subsection 3.1.3 a so-called *general* approximation operator. This operator will be able to approximate a function in only a fraction of the function space. We discuss in the next section, which fraction to be chosen in order to get the 'optimal' approximation.

### 3.1.1 Construction in One Dimension

*Discretization  
of the Function  
Space*

Let  $I \subseteq \mathbb{R}$  be a finite or infinite subspace of the real numbers and let be  $V$  an infinite dimensional Hilbert space containing functions  $I \rightarrow \mathbb{R}$  of interest. Consider a sequence of linearly independent functions  $\phi_n \in V$  for every  $n \in \mathbb{N}_0$ . To discretize the function space  $V$ , we define finite dimensional linear subspaces with respect to the basis set

$$\mathcal{B}_n := \{\phi_i : i = 0, \dots, g_n\}$$

by

$$V_n := \text{span}\{\mathcal{B}_n\}, \quad (3.1.1)$$

for  $n \in \mathbb{N}_0$  and an arbitrary strictly increasing sequence  $(g_n)_{n \in \mathbb{N}_0}$  of natural numbers. By construction these function spaces are strictly nested,

$$V_0 \subset V_1 \subset V_2 \subset \dots \subset V.$$

Moreover we assume that  $\bigcup_{n \in \mathbb{N}_0} V_n$  is dense in  $V$ .

*Approximation  
Operator*

The next step is to define an unique approximation operator, which maps any function  $f \in V$  onto its unique approximation in  $V_n$ ,

$$\mathcal{O}_n : V \rightarrow V_n : f \mapsto \mathcal{O}_n f. \quad (3.1.2)$$

Since we do not know the function  $f$ , we need a certain amount of additional information to define a unique operator. This information comes in the form



of evaluations of the function at specified values. The larger the dimension of the underlying function space  $V_n$ , the more function evaluations are usually involved to ensure uniqueness. The choice of this operator characterizes the problem one wants to solve. If one aims to interpolate the function  $f$ , then this operator would be an interpolation operator, uniquely defined by a sufficiently chosen set of interpolation points. If the goal is to approximate the integral of  $f$ , then this operator would be defined as a quadrature operator uniquely defined by a number of quadrature points. In the case of a linear regression operator mapping onto the search space, the uniqueness is ensured by a sufficient amount of data points and the introduction of a regularization term. Characterizing this operator really handles the problem. In this section we will only consider an unspecified uniquely defined operator to illustrate the main idea of the tensor-based subset splitting. Later on, we will focus exclusively on a least squares regression problem.

Let's briefly repeat what we have done up to this point. First, we started with a Hilbert space  $V$  consisting of one-dimensional functions over  $I \subseteq \mathbb{R}$ . We discretized this function space using a nested sequence of finite-dimensional function spaces  $(V_n)_{n \in \mathbb{N}_0}$  lying dense in  $V$ . Moreover we assumed the existence of a unique approximation operator  $\mathcal{O}_n : V \rightarrow V_n$  which we did not characterize further. In the following we construct the case of a multi dimensional function space. *Review*

### 3.1.2 Construction in Multiple Dimensions

The  $d$ -dimensional case comes natural due a product construction. This construction will allow us an independent view of each dimension and enables the sparse tensor product construction. From now on, we will denote with bold letters multi indices or elements in multiple dimensions. Let *Multi Dimensional Function Space*

$$V^d := V \otimes \dots \otimes V$$

be an infinite dimensional Hilbert space of  $d$ -dimensional functions. For a multi index  $\mathbf{k} \in \mathbb{N}_0^d$ , define the multivariate version of linearly independent functions  $\phi_{\mathbf{k}} : I^d \rightarrow \mathbb{R}$  as the tensor product of one dimensional functions chosen in (3.1.1) by

$$\phi_{\mathbf{k}} := \left( \bigotimes_{i=1}^d \phi_{k_i} \right) (x_i) = \prod_{i=1}^d \phi_{k_i}(x_{k_i}).$$

Again, these functions span strictly nested linear subspaces of  $V^d$  like in one dimension with the basis sets

$$\mathcal{B}_{\mathbf{k}} := \{\phi_{\mathbf{l}} : \mathbf{l} \in \mathcal{G}_{\mathbf{k}}\}$$

by

$$V_{\mathbf{k}} := \text{span}(\mathcal{B}_{\mathbf{k}}), \tag{3.1.3}$$

where the multivariate version of the index set reads

$$\mathcal{G}_{\mathbf{k}} := \{0, \dots, g_{k_1}\} \times \dots \times \{0, \dots, g_{k_d}\},$$

for a strictly increasing sequence  $(g_n)_{n \in \mathbb{N}_0}$  of natural numbers as in the one dimensional case. Equivalently to (3.1.3) one can write

$$V_{\mathbf{k}} = V_{k_1} \otimes \dots \otimes V_{k_d},$$

where  $V_{k_i}$  constructed as in the previous subsection.

*Multi  
Dimensional  
Approximation  
Operator*

The approximation operator in multiple dimensions is given by

$$O_{\mathbf{k}} : V^d \rightarrow V_{\mathbf{k}}, \quad f \mapsto O_{\mathbf{k}}f := \bigotimes_{i=1}^d O_{k_i}f_i. \quad (3.1.4)$$

In this subsection, we generalized the one dimensional previous situation into the multidimensional case. We considered the tensor product function space  $V^d = V \times \dots \times V$  and defined a multi-dimensional approximation operator  $O_{\mathbf{k}} : V^d \rightarrow V_{\mathbf{k}}$  by  $O_{\mathbf{k}}f := O_{k_1}f_1 \cdot \dots \cdot O_{k_d}f_d$  for  $f \in V^d$  for a *full grid space*  $V_{\mathbf{k}}$ . Our goal is to replace the full grid space  $V_{\mathbf{k}}$  with an optimally chosen function space, which is more sparse. For this we need to introduce a more flexible approximation operator which is able to operate on such sparse spaces. Therefore we define in the following the *general approximation operator*.

### 3.1.3 General Approximation Operator

*Hierarchical  
Representation  
of  $O_{\mathbf{k}}$*

The cost of constructing the approximation operator usually grows exponentially in the dimension and is defined on a product space or so-called *full grid space*. The idea of the tensor-based subset splitting approach is to define the approximation operator on a more flexible space instead of on the product space as we did in this subsection. First, we derive a hierarchical representation for the approximation operator  $O_{\mathbf{k}}$ . This is only possible due to the specific construction by tensor products in multi dimensions. The hierarchical representation opens the possibility of an *optimal* truncation, which correspondingly leads to a truncation of the function space it is defined on. By a truncation, if it is possible, we are turning from a full grid function space to a more sparse one. Here we call a truncation 'optimal', if the resulting general approximation operator achieves maximum accuracy at a given cost. We start with the hierarchical representation of the approximation operator  $O_{\mathbf{k}}$ . Therefore we write the operator in terms of an hierarchical sum of so-called *difference operators*. We will first consider the one dimensional case and then generalize to higher dimensions.

**Definition 3.1: Difference Operator**

In the one dimensional case, let  $(\mathcal{O}_n)_{n \in \mathbb{N}_0}$  be a given sequence of operators, where each  $\mathcal{O}_n : V \rightarrow V_n$  as defined in (3.1.2). Then we call

$$\Delta_n := \mathcal{O}_n - \mathcal{O}_{n-1} : V \rightarrow V_n,$$

the difference operator sequence with respect to  $(\mathcal{O}_n)_{n \in \mathbb{N}_0}$ .

In the  $d$ -dimensional case, let be  $(\mathcal{O}_{\mathbf{k}})_{\mathbf{k} \in \mathbb{N}_0^d}$  a given sequence of operators, where each  $\mathcal{O}_{\mathbf{k}} : V^d \rightarrow V_{\mathbf{k}}$  as defined in (3.1.4). For a multi index  $\mathbf{k} \in \mathbb{N}^d$  the corresponding difference operator reads

$$\Delta_{\mathbf{k}} := \Delta_{k_1} \otimes \cdots \otimes \Delta_{k_n} : V^d \rightarrow V_{\mathbf{k}}.$$

An application of the telescopic sum now gives the *hierarchical representation* of the approximation operator,

$$\begin{aligned} \mathcal{O}_{\mathbf{k}} &= \mathcal{O}_{k_1} \otimes \cdots \otimes \mathcal{O}_{k_d} \\ &= \sum_{i_1=0}^{k_1} \Delta_{i_1} \otimes \cdots \otimes \sum_{i_d=0}^{k_d} \Delta_{i_d} \\ &= \sum_{\mathbf{0} \leq \mathbf{i} \leq \mathbf{k}} \Delta_{\mathbf{i}}. \end{aligned} \quad (3.1.5)$$

Now, a truncation of this sum would also lead to an underlying sparser function space than  $V_{\mathbf{k}}$ .

Let us in the following define a more general approximation operator, which can be only defined by a subset of the above summands.

*Truncating  
the Hierarchical  
Representation  
of  $\mathcal{O}_{\mathbf{k}}$*

**Definition 3.2: Generalized Approximation Operator**

Consider an arbitrary index set  $\mathcal{I} \subseteq \mathbb{N}_0^d$ . Define the generalized approximation operator  $\mathcal{O}_{\mathcal{I}}$  by

$$\mathcal{O}_{\mathcal{I}} := \sum_{\mathbf{i} \in \mathcal{I}} \Delta_{\mathbf{i}}. \quad (3.1.6)$$

Note that If  $\mathcal{I} = \{\mathbf{i} \in \mathbb{N}_0^d : \mathbf{i} \leq \mathbf{k}\}$ , then  $\mathcal{O}_{\mathcal{I}} = \mathcal{O}_{\mathbf{k}}$ . With this definition we are able to take only a fraction of the difference operators into account, which would be needed for an approximation on a full grid.

Thus, the previously introduced operator splitting (3.1.5) induces a splitting of the entire function space  $V_{\mathbf{k}}$ . Let us for a second assume, we have chosen a arbitrary subset of difference operators of characterized by an index set  $\mathcal{I} \subseteq \{\mathbf{i} \in \mathbb{N}_0^d : \mathbf{i} \leq \mathbf{k}\}$ . The choice of this index set will later be chosen in an optimal way and will be the most significant choice we have to make in the following. Then, the generalized approximation operator now operates on a

*Image Space of  
the Truncation*

more sparse function space  $\mathcal{O}_{\mathcal{I}} : V \rightarrow V_{\mathcal{I}} \subseteq V_{\mathbf{k}}$ . The next important step will be the specification of the function space  $V_{\mathcal{I}}$ . Incidentally there are two ways to do so. The first one is to define a set of *hierarchical basis functions* which correlates in a way with the difference operators. Once an hierarchical basis is found, it is easy to define the function space  $V_{\mathcal{I}}$ . Nevertheless, in some cases the choice of an hierarchical basis is far away from trivial or it is not meaningful to define one for other reasons. Therefore, there is a second proceeding, called the *combination technique* [36], where the function space  $V_{\mathcal{I}}$  is constructed by an inclusion-exclusion-type combination of full grid spaces. Therefore, no basis set has to be chosen, but usually the numerical effort is greater. In the following we will concentrate on the first case.

*Hierarchical  
Representation  
of  $V_{\mathbf{k}}$*

Define the linear subspace of  $V_{\mathbf{k}}$  characterized by the difference operator by

$$W_{\mathbf{k}} := \text{Im}(\Delta_{\mathbf{k}}) \subset V_{\mathbf{k}}.$$

By (3.1.5) this gives us a hierarchical representation also of the linear function space  $V_{\mathbf{k}}$  for some  $\mathbf{k} \in \mathbb{N}_0^n$ ,

$$V_{\mathbf{k}} = \sum_{0 \leq \mathbf{i} \leq \mathbf{k}} V_{\mathbf{i}} = \bigoplus_{0 \leq \mathbf{i} \leq \mathbf{k}} W_{\mathbf{i}}. \quad (3.1.7)$$

For a visualization of the interrelations of operators and functional spaces see Figure 3.2. To characterize a function space  $V_{\mathcal{I}}$  which only depends on a fraction of the above summands, we introduce *hierarchical basis sets*. By [38], we call a basis set of  $V_{\mathbf{k}}$  an hierarchical basis set if it satisfies the following definition.

### Definition 3.3: System of Hierarchical Basis Sets

We call a sequence of basis sets  $\hat{\mathcal{B}}_{\mathbf{k}}$  for  $\mathbf{k} \in \mathbb{N}_0^d$  a *system of hierarchical basis sets* with respect to the operators  $(\mathcal{O}_{\mathbf{k}})_{\mathbf{k} \in \mathbb{N}_0^d}$ , if it satisfies the following properties:

1. The sequence is nested, i.e.  $\hat{\mathcal{B}}_{\mathbf{k}-\mathbf{e}_j} \subset \hat{\mathcal{B}}_{\mathbf{k}}$  for all  $\mathbf{k} \in \mathbb{N}_0^d$  and  $j \in \{1, \dots, d\}$ .
2.  $\hat{\mathcal{B}}_{\mathbf{k}}$  forms a basis set of  $V_{\mathbf{k}}$  for all  $\mathbf{k} \in (\mathbb{N}_0 \cup \{-1\})^d$ .
3.  $\hat{\mathcal{B}}_{\mathbf{k}} \setminus \hat{\mathcal{B}}_{\mathbf{k}-1}$  forms a basis set of  $W_{\mathbf{k}}$  for  $\mathbf{k} \in \mathbb{N}_0^d$ , where

$$\begin{aligned} \hat{\mathcal{B}}_{\mathbf{k}} \setminus \hat{\mathcal{B}}_{\mathbf{k}-1} &= \hat{\mathcal{B}}_{\mathbf{k}} \setminus \bigcup_{j=1}^d \hat{\mathcal{B}}_{\mathbf{k}-\mathbf{e}_j} \\ &\stackrel{!}{=} \hat{\mathcal{B}}_{\mathbf{k}} \setminus \bigcup_{\mathbf{i} \leq \mathbf{k}} \hat{\mathcal{B}}_{\mathbf{i}} \end{aligned}$$

*Admissible  
Index Set  $\mathcal{I}$*

The question of interest is now, how to choose the index set  $\mathcal{I}$  in to obtain the optimal approximation of  $f$  on  $V_{\mathcal{I}}$ . Gerstner and Griebel introduced in [31] the concept of *general Sparse Grids*, by concentrating on index sets, which are *admissible* or *downward closed*.

### 3.1. A TENSOR PRODUCT CONSTRUCTION OF THE FUNCTION SPACE 31

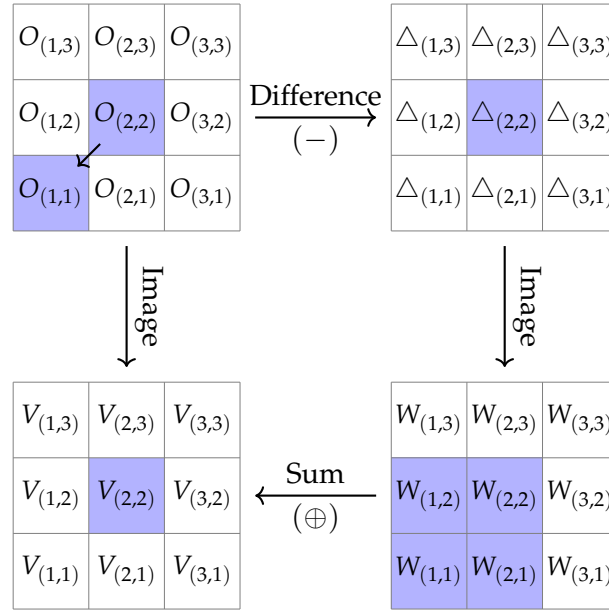


Figure 3.2: Visualization of the relationship of the introduced operators  $\mathcal{O}_{\mathbf{k}}, \Delta_{\mathbf{k}}$  (upper two grids) and function spaces  $V_{\mathbf{k}}, W_{\mathbf{k}}$  (lower two grids) in two dimensions ( $d = 2$ ).

#### Definition 3.4: admissible index set

An index set  $\mathcal{I}$  is called *admissible* if for all  $\mathbf{k} \in \mathcal{I}$ ,

$$\mathbf{k} - \mathbf{e}_j \in \mathcal{I}, \quad (3.1.8)$$

for  $1 \leq j \leq d, k_j > 1$ .

From now on we will consider only index sets which are assumed to have the admissibility property. Consider an admissible index set  $\mathcal{I} \subset \mathbb{N}_0^d$ , then the image space of the generalized approximation operator  $\mathcal{O}_{\mathcal{I}} : X^d \rightarrow V_{\mathcal{I}}$  can be written as

$$V_{\mathcal{I}} := \sum_{\mathbf{l} \in \mathcal{I}} V_{\mathbf{l}}.$$

Moreover a generalized version of (3.1.7) holds:

$$V_{\mathcal{I}} = \bigoplus_{\mathbf{i} \in \mathcal{I}} W_{\mathbf{i}}.$$

Hence, any function  $f \in V_{\mathcal{I}}$  and can be uniquely split into its contributions of the hierarchical function spaces

$$f(x) = \sum_{\mathbf{l} \in \mathcal{I}} f_{\mathbf{l}}(x), \text{ where } f_{\mathbf{l}} \in W_{\mathbf{l}}.$$

Let us shortly recap what we have done so far. To reduce the complexity of an

*Truncating the Hierarchical Representation of  $V_{\mathbf{k}}$*

*Review*

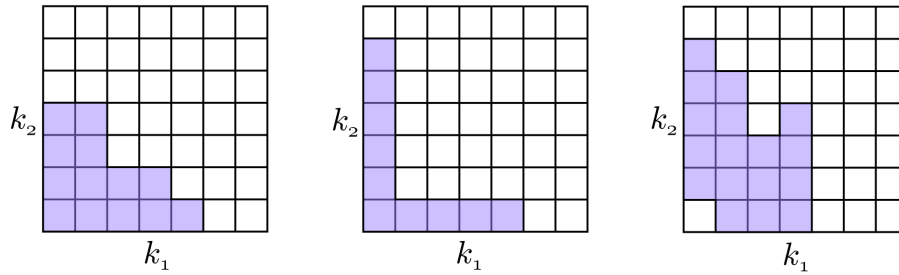


Figure 3.3: Index sets in two dimensions. The leftmost picture shows one admissible index sets with some relation between the dimensions, the middle one totally without. The rightmost picture shows a non-admissible index set.

approximation of a multidimensional function, we reduced from an approximation on a Full Grid space  $V_{\mathbf{k}} = \bigoplus_{i \leq \mathbf{k}} W_i$  to an approximation on a freely chosen Space  $V_{\mathcal{I}} = \bigoplus_{i \in \mathcal{I}} W_i \subseteq V_{\mathbf{k}}$ , depending on an admissible index set  $\mathcal{I}$ . The optimal choice of this index set is the crucial point of the Sparse Grid Approach. Given an effort that we are willing to pay for the approximation, we want to find the index set leading to the best possible approximation. In the next chapter we tackle the problem on how to choose an optimal index set, by specifying the vague notion ‘optimality’.

*Various  
Smoothness  
Assumptions*

There is already a broad theory on how to specifically choose the index set when certain smoothness assumptions are met. For example, if we assume the unknown function to be an element of the Sobolev Space of mixed smoothness, then the Fourier coefficients of this function decreases with a certain rate. Constructing the subspaces  $W_n$  based on a Fourier basis we consequently observe a certain decay rate of contributions and are able to state the specific ‘optimal’ index set for such functions  $f$ . In general, unfortunately, we do not have sufficient knowledge about the smoothness of the function we want to approximate. In this cases, we have to rely on adaptive algorithms. Thus, there are two approaches to find an ‘optimal’ index set. On the one hand, we assume the unknown function to be an element of a certain function class. Therefore, we are able to define an index set which solves an optimization problem for the general function class. On the other hand, we have no knowledge about the smoothness of the function. In this case, we rely on function evaluations. Here, we adaptively choose the indices for which the approximation error for the given data is minimized. Thus, adaptive methods are particularly tailored to the specific function, while theoretic results apply for a whole function class.

In the following section, we will formulate the problem of finding the optimal index set as an optimization problem.

### 3.2 Choice of an Index Set

We are searching for an optimal index set  $\mathcal{I}^{(opt)} \subset \mathbb{N}_0^d$ , such that the corresponding approximation operator  $\mathcal{O}_{\mathcal{I}^{(opt)}} : V \rightarrow V_{\mathcal{I}^{(opt)}}$  leads to the optimal approximation under a given work count. The aim is to profit from a given work count as much as possible. This problem can be written as the following optimization problem. Let be  $w$  the cost we are willing to pay for the approximation, then

*Optimization Problem*

$$\max_{\substack{f \in V^d \\ |f|=1}} \|f - \mathcal{O}_{\mathcal{I}^{(opt)}} f\| = \min_{\substack{\mathcal{I} \subset \mathbb{N}_0^d \\ |V_{\mathcal{I}}|=w}} \max_{\substack{f \in V^d \\ |f|=1}} \|f - \mathcal{O}_{\mathcal{I}} f\|, \quad (3.2.1)$$

for, yet unspecified, norm  $\|\cdot\|$  and semi norm  $|\cdot|$ . Note that the choice of an optimal index set following (3.2.1) highly depends on the choice of these norm and semi norm and may differ for diverse choices. It is also possible to put the problem in the opposite direction. For a given error  $\epsilon$  one wants to achieve we would like to find the index set with the minimum cost.

To formulate the optimization problem, we have to characterize a notion of cost and benefit coming along with each index set  $\mathcal{I} \subset \mathbb{N}_0^d$ .

*Cost Function*

#### Defintion 3.5: Cost Function

Let be everything defined as in the previous chapter. Define the cost function  $C : \mathbb{N}_0^d \rightarrow \mathbb{N}$  as the function mapping each index set  $\mathcal{I} \subset \mathbb{N}_0^d$  onto its related cost,

$$C(\mathcal{I}) := \dim(V_{\mathcal{I}}).$$

The cost function maps an index set  $\mathcal{I} \subset \mathbb{N}_0^d$  onto the degrees of freedom involved, when approximating a function  $f \in V^d$  by  $\mathcal{O}_{\mathcal{I}} f$ . Doing so, the best approximation is searched in

$$\text{Im}(\mathcal{O}_{\mathcal{I}}) \stackrel{(3.1.6)}{=} \bigoplus_{i \in \mathcal{I}} \text{Im}(\Delta_i) = \bigoplus_{i \in \mathcal{I}} W_i = V_{\mathcal{I}},$$

for that we need to adjust one parameter per basis function of  $V_{\mathcal{I}}$ . So it makes sense to define the local cost of this index as the dimension, i.e. the number of basis functions, of  $V_{\mathcal{I}}$ .

*Benefit Function*

#### Defintion 3.6: Benefit Function

Define the *benefit function*  $B : \mathbb{N}_0^d \rightarrow \mathbb{R}$  as the function mapping each index set  $\mathcal{I} \subset \mathbb{N}_0^d$  onto its related benefit when approximating a function  $f \in V$ ,

$$B(\mathcal{I}) := \|\mathcal{O}_{\mathcal{I}} f\|,$$

for some suitable norm on  $V_{\mathcal{I}}$ .

Now we are able to assign to each index set  $\mathcal{I} \subset \mathbb{N}_0^n$ , its benefit and the cost associated with its calculation. For a given workload  $w$ , the optimization problem (3.2.1) can be rewritten as

$$\max_{\mathcal{I} \subset \mathbb{N}_0^n} B(\mathcal{I}) \quad \text{with} \quad C(\mathcal{I}) = w. \quad (3.2.2)$$

For linear applications it is possible to decompose the cost and benefit function in local parts measuring the cost and benefit of each single index of the index set separately. With a combinatorial argument, see for example [16], the global optimization problem (3.2.1) or (3.2.2), respectively, can then be reduced to the discussion of the local cost–benefit ratios  $b(\mathbf{1})/c(\mathbf{1})$  of the underlying subspaces  $W_{\mathbf{1}}$ .

### Outlook

Later on, in chapter 6, we will construct a machine learning interatomic potential based on an adaptive sparse grid approach. In this case the approximation operator will be based on a penalized least squares regression. As we can not specify the smoothness class of the potential energy surface, we rely on further information in order to construct a well defined approximation operator. Thus, given a data set  $\mathcal{D}_n = \{(x_i, y_i) : i = 1, \dots, n, \}$  the generalized approximation operator will take the form

$$\mathcal{O}_{\mathcal{I}} : V \longrightarrow V_{\mathcal{I}},$$

mapping a function  $f \in V$  onto its approximation

$$\mathcal{O}_{\mathcal{I}} f := \operatorname{argmin}_{f \in V_{\mathcal{I}}} \frac{1}{n} \sum_{i=1}^n \|f(x_i) - y_i\|_2^2 + \|f\|_{\Gamma},$$

for some penalizing norm  $\|\cdot\|_{\Gamma}$  on  $V_{\mathcal{I}}$ . In our application, the function space  $V_{\mathcal{I}}$  will take a rather complex form. Not only physical assumptions will be encoded in it, but it will also reflect the decomposition of the physical system in lower dimensional subsystems. Thus, the function space will have a highly non-linear behavior regarding its index set. In our case, we won't be able to break the global benefit function into local contributions, since they connect in a non-linear way.

## 3.3 Related Work

### Curse of Dimensions

The standard way of representing multidimensional functions are tensor or full grids. Thus a conventional discretization on a uniform grid in  $d$  dimensions and  $O(2^n)$  points in each direction involves  $O(2^{dn})$  degrees of freedom. This phenomenon is associated with the term curse of dimensions, which is first named in 1961 [8]. Since then, circumventing the curse of dimensions has been a much studied field and usually requires additional information about the function to be approximated. If we can assume certain smoothness properties of the function, we are able to significantly decrease the complexity of



the problem.

The idea of a sparse tensor product construction first came up in 1963 with the work of [62], who mainly focused on the use of Fourier transforms, i.e. on function spaces  $V_n$ , which are spanned by Fourier basis functions  $w_k(x) = e^{ikx}$ . The name 'Sparse grid' first came up in [38] who also focused on Fourier transforms and is associated with a specific dyadic choice of function spaces  $V_n$ .

Let be  $f$  an element of the Sobolev space with bounded mixed smoothness  $\mathcal{H}_{\text{mix}}^t(\Omega)$ , ie it is assumed that the  $t$ -th mixed derivatives of the function  $f$  are bounded, and  $\Omega$  represents a product space. This means that the oscillations of the function can be isolated well in the direction of individual coordinates. Under this smoothness assumption an approximation using a sparse grid function space should be preferred to the full grid function space. If one measures the benefit-cost ratio in the  $\mathcal{L}_2$  or  $\mathcal{L}_\infty$  norm, this leads to the *regular sparse grid spaces* involving  $O(2^n d^{n-1})$  degrees of freedom, based on the index set,

$$\mathcal{I}_n^{\text{SG}} = \{\mathbf{k} : |\mathbf{k}|_1 \leq n + d - 1\}.$$

Measuring the benefit-cost ratio in the energy norm, leads to an even sparser grid involving  $O(2^n)$  degrees of freedom, the so-called *energy-norm based sparse grids* [16] constructed by the index set,

$$\mathcal{I}_n^{\text{ESG}} = \{\mathbf{k} : |\mathbf{k}|_1 - C_1(\mathbf{k}) \leq n + d - 1 - C_2(n, d)\},$$

for suitable chosen  $C_1$  and  $C_2$ . The notion of sparse grid arose for the first time in 1991 by Zenger, treating the trigonometric interpolation of a multidimensional periodic function in  $\mathcal{H}_{\text{mix}}^T(\mathbb{T}^n)$  living on the  $d$  dimensional torus.

Considering a Sobolev space of generalized mixed smoothness  $\mathcal{H}_{\text{mix}}^{t,r}(\Omega)$  leads to the *Sobolev Space with generalized mixed Smoothness*

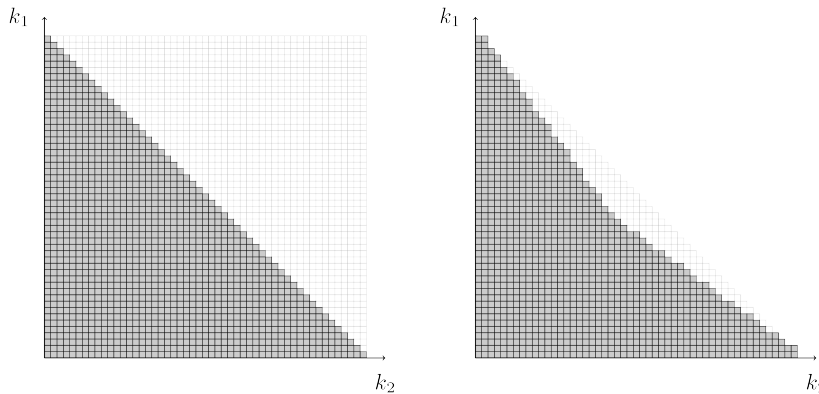


Figure 3.4: A visualization of the mentioned index sets in two dimensions. On the left hand the index set of a regular Sparse Grid  $\mathcal{I}_{50}^{\text{SG}}$  compared to the index set of the Full Grid,  $\mathcal{I}_{50}^{\text{FG}}$ , in the background. On the right hand side the comparison of the index set of an energy-based Sparse Grid  $\mathcal{I}_{50}^{\text{ESG}}$  and regular Sparse Grid of the same level in the background.

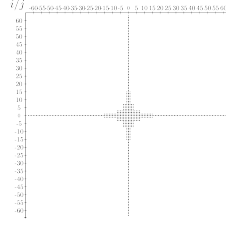


Figure 3.5: The index set of the *dyadic chosen* basis functions, when considering a regular Sparse Grid  $\mathcal{I}_{10}^{SG}$ , forms an hyperbolic cross. The associated function space  $V_{\mathcal{I}_{10}^{SG}}$  is by construction spanned by all basis functions  $\phi_{(i,j)}$  for which  $i \in \{1 - 2^{k_1-1}, \dots, 2^{k_1-1}\}$  and  $j \in \{1 - 2^{k_2-1}, \dots, 2^{k_2-1}\}$  for some  $(k_1, k_2) \in \mathcal{I}_{10}^{SG}$ .

to the generalized Sparse Grids [43] based on the index set

$$\mathcal{I}_n^T = \left\{ \mathbf{k} : \prod_{i=1}^d (1 + k_d) \cdot (1 + |\mathbf{k}|_\infty) \leq n^{(1-T)} \right\}.$$

Here, the parameter  $T \in [-\infty, 1)$  controls the mixture of isotropic and mixed smoothness. The choice of this index sets is a generalization of the upper cases, for  $T = 0$  we are back in the case of the conventional hyperbolic cross or regular sparse grids, for  $T = -\infty$  it describes the full grid.  $T \rightarrow 1$  converges to a latin hypercube and the case  $0 < T < 1$  resembles energy standard based sparse grids. For a visualization of the mentioned index sets, see Figure 3.4.

#### Hyperbolic Cross Approximation

Usually, the support points are chosen dyadically, so each interval is halved in each refinement step. This procedure not only leads to a nesting of the support points, but also to a nesting of any considered function spaces. In addition, the dimension of the underlying function spaces  $V_n$  is doubled in each refinement step, ie  $|V_{n+1}| = 2 \cdot |V_n|$ . This special construction is used in the associated proven complexity and convergence rates. By their specific dyadic construction, the indices of basis functions of  $V_{\mathcal{I}_n^{SG}}$  form the shape of a hyperbolic cross, for a visualization see Figure 3.5. This is why Sparse Grid methods are also known under the name *hyperbolic cross approximation*.

## Chapter 4

# Active Learning

Up to this point, we have assumed the training data  $\mathcal{D}_n$  to be fixed and increased the benefits of an approximation solely by adjusting the underlying search space, with the aim of adapting the algebraic form of the function. Usually, machine learning methods improve their accuracy through increasing the number of fitting parameters, making the allowed algebraic form of the potential function more flexible. However, the estimation does not only depend on the algebraic form, but also on the training data which is used to fit them. During the fitting process, the method is usually unconcerned with the information content of a data point, but learns from what it gets. Since such methods do not question what they are learning from, they are called *passive learning* approaches. Although passive learning approaches are a powerful tool, they are in general necessarily interpolative, which means they fail to give reasonable results in areas outside their training domain. One idea to address inaccuracy in extrapolatory cases would be a proper choice of the training domain, in a way that minimizes those areas of uncertainty and ensures interpolation over relevant regions. This makes the choice of an optimal training set a problem of transferability. Thus, if the method were able to detect extrapolative configurations and add those to the training set, we would be able to improve the approximative power not only by allowing a more flexible algebraic form but also by allowing a more flexible choice of the training data. In such cases, the method actively influences the selection of the training set and is thus called an *active learning* or *query learning* approach. For an overview of the field of active learning, we refer to [58].

*Supervise  
the  
Training Set*

In the following section, we will present a pool-based active learning approach by choosing the training set such that the expected future error is minimized. To this end, we will aim to maximize the Fisher information contained in the data. The Fisher information is the curvature of the log likelihood function and will soon be introduced more formally. We will show, that this approach coincides with the *D-optimality* term of the *optimal experimental design* (OED) in the statistical literature [26]. Afterwards, in subsection 4.4 we will apply the D-optimality criterion to the penalized least squares regression.

*Structure*

## 4.1 From Passive Learning to Active Learning

*Passive  
Learning*

Given a measure space  $X \subset \mathbb{R}^d$  and an Hilbert space  $(Y, \langle \cdot, \cdot \rangle_Y)$ , we are interested in approximating a map

$$f : X \longrightarrow Y, \quad (4.1.1)$$

based on a given a set of  $n$  evaluations  $\mathcal{D}_n := \{(x_i, y_i) : i = 1, \dots, n\}$ . In chapter 2 we introduced the approximation  $\hat{f}$  of (4.1.1) as the solution of a penalized least squares regression problem. Given a randomly chosen set of training data and a prescribed finite dimensional function space  $V_k$ , we defined  $\hat{f}$  as the unique function minimizing the empirical error functional. Here, the parameter  $k$  is a notion of the approximation fineness of the search space  $V_k$ . For example,  $V_k$  could contain nodal functions defined on an underlying grid, which gets finer as  $k$  increases. Choosing the function  $\hat{f}$  is equivalent to the problem of choosing the coefficients  $\hat{w}$  of the basis in the search space. As we explained in 2.2, there are two ways to interpret the solution  $\hat{f}$  of a penalized least squares regression problem. On the one hand it is the function, which minimizes the empirical regression error functional over the search space  $V_k$ . On the other hand it is the function which maximizes the mean of the posterior distribution, which coincides in the Gaussian case with the mode of the posterior. Until here, we solely improved accuracy through increasing the number of fitting parameters by expanding the search space, i.e. increasing  $k$ . Enlarging the search space makes the allowed algebraic form of the estimate more flexible. Nevertheless, an approximation does not only depend on the algebraic form, but also on the training data which is used to fit it. If the data is limited to a small area, then we will not be able to adjust the function outside of this range, even if we allow arbitrarily complex functions. During the fitting process, the method is usually not able to influence the choice of training data in any way.

*Active  
Learning*

We are now tackling the following question: how can we pick the training data in a suitable way, such that it is as easy as possible to choose the optimal approximation? Or, how can we pick the training data to faster converge to the optimal approximation, using less data points? If the likelihood function  $p(y|\mathcal{D}_n, \hat{w})$  is sharply peaked with respect to changes in  $\hat{w}$ , it is easy to indicate the 'correct' value of  $\hat{w}$  from the data, or equivalently, that the data  $\mathcal{D}_n$  provides a lot of information about the parameter  $\hat{w}$ . If  $p(y|\mathcal{D}_n, \hat{w})$  is flat and spread-out, then it would take many, many samples like  $\mathcal{D}_n$  to estimate the actual 'true' value of  $\hat{w}$ . This suggests studying some kind of variance with respect to  $\hat{w}$  or equivalently information of  $\hat{w}$  incorporated in the training data.

*Pool-Based  
Sampling*

The key idea behind active learning is that a machine learning algorithm can achieve greater accuracy with fewer training labels if it is allowed to choose the data from which it learns. Let  $T$  be the total set of all data. So far, we would have chosen a training data set randomly  $\mathcal{D}_n \subset T$ , then label and learn from it. Unlabeled data may be abundant or easily obtained, but labels are

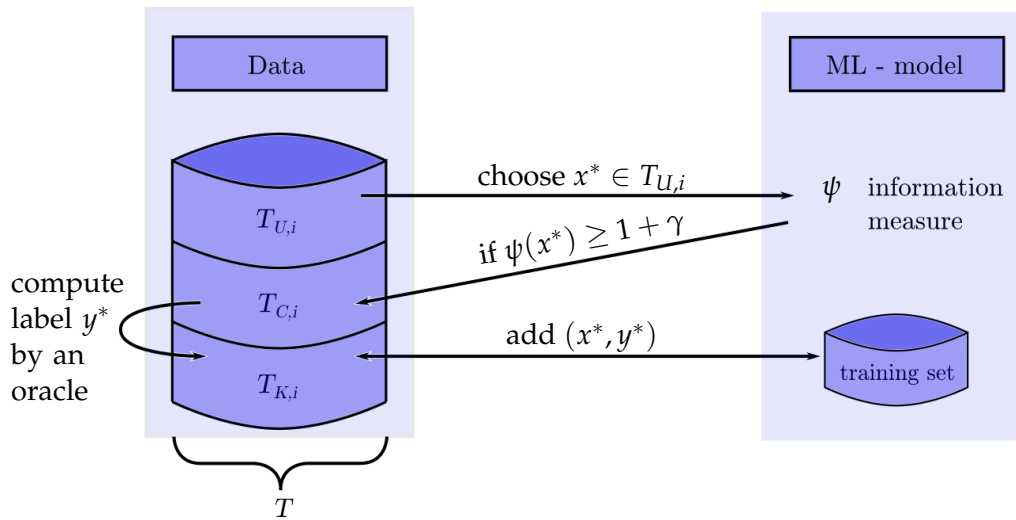
difficult, time-consuming, or otherwise expensive to obtain. An active learner may pose *queries*, usually in the form of unlabeled data instances to request its label, which is usually computationally expensive or time-consuming. In each iteration  $i$ ,  $T$  is broken up into three disjoint subsets  $T = T_{K,i} \cup T_{U,i} \cup T_{C,i}$ . Here, we denote by

- $T_{K,i}$  the subset which contains the data, with known labels,
- $T_{U,i}$  the subset which contains the data with unknown label and
- $T_{C,i}$  containing the data which is chosen to be labeled.

*Dataset  
Splitting*

at iterate  $i$ . Thus, in a *pool-based sampling* approach, we are assuming the existence of a large pool of unlabeled data from which we want to select a certain number of data points containing a maximum of information. Those selected instances would be labeled and used to learn from.

In each step, the data point in  $T_{U,i}$  which contains the maximal information is added to  $T_{C,i}$  and its label is requested. Once the label is computed, which is usually computationally expensive, the data point is added to  $T_{K,i}$ .



All active learning scenarios involve evaluating the informativeness of unlabeled instances with respect to an information measure  $\psi$ . Based on this measure, instances are added to the training set if the incorporated information exceeds a prescribed threshold  $1 + \gamma$  for some  $\gamma > 0$ . The training set of a machine learning model is then chosen as a subset of the labeled data  $\mathcal{D} \subseteq T_{K,i}$ . There are certain query strategies which operate on a training set  $\mathcal{D}_m$  of fixed size  $m$ , i.e. in order to retain the training size an incorporation of the data point  $(x^*, y^*)$  would require removing another data point from the training set. This assumption can simplify the computations by keeping an involved matrix quadratic, as will be observed in subsection 4.4.3. A variety of such so-called query strategies have been developed and an overview can be found

*Information  
Measure*

in [58].

*Variance  
Reduction*

In the following subsection, we will describe one well-known query strategy, variance reduction or, in statisticians literature, the D-optimality approach. Therefore, we first address the mathematical intuition behind that and will point out relationships with the Fisher information.

## 4.2 Variance Reduction

*Expected  
Error  
Reduction*

The variance reduction approach is basically a simplification of the reduction of the expected future error. There, the idea is to estimate the expected future error of a model trained using  $\mathcal{D} \cup (x^*, y^*)$  on the remaining instances  $T \setminus \{\mathcal{D} \cup (x^*, y^*)\}$ , and query the instance  $(x^*, y^*)$  with minimal expected future error. There are a variety of approaches to compute the instance which minimizes the expected future error, for example by minimizing the expected 0/1-loss, minimizing the expected log-loss or maximizing the expected information gain of the query to name a few. But, unfortunately, they all have one thing in common: in most cases a direct minimization of the expected future error is the most computationally expensive query framework, as they require estimating the expected future error over  $T \setminus \{\mathcal{D} \cup (x^*, y^*)\}$  for each query. This means a new model must be incrementally re-trained for each possible query labeling, which in turn iterates over the entire pool. Additionally, often it is not possible to generate a solution in a closed form at all.

*Bias-Variance  
Tradeoff*

Since a direct expected error reduction is computational demanding, we resort on an implicit reduction by reducing the variance. It turns out that whichever function  $\hat{f}$  we select as an approximation, we can decompose its expected error on an unseen sample  $x$  as follows:

$$\begin{aligned} \mathbb{E}[(y - \hat{f}(x))^2 | x, \hat{w}] &= \mathbb{E}[(y - \mathbb{E}[y|x])^2] \\ &\quad + (\mathbb{E}[y - \hat{f}(x) | x, \hat{w}])^2 \\ &\quad + \mathbb{E}[(\hat{f}(x) - \mathbb{E}[\hat{f}(x) | x, \hat{w}])^2 | x, \hat{w}] \\ &= \text{Var}(y|x) + \text{Bias}(\hat{f}(x) | x, \hat{w})^2 + \text{Var}(\hat{f}(x) | x, \hat{w}). \end{aligned} \tag{4.2.1}$$

Here, we used a zero addition with  $\pm \mathbb{E}[y|x]$  and  $\pm \mathbb{E}[\hat{f}(x) | x, \hat{w}]$  and short calculations showing that the mixed terms vanish. In (4.2.1) the first summand on the right hand side denotes the variation of the “true” label and thus encodes the noise incorporated in the data. The noise term does not depend on the choice of  $x$  in our case and thus can be neglected. The second and the third term are the bias and the variance. Often we have a trade-off between these two terms and we have to choose which one to optimize. But the expected error benefits from improvements in both. We will optimize the third term as we will see that this is possible without much computational overhead and even without fitting the model.

A variance reduction query strategy is called in statistical literature the problem of an *optimal experimental design* (OED) [26]. Especially advanced and mathematically sophisticated is the variance reduction for linear models, in particular linear regression models. The basic idea is to choose the training data in such a way, that the incorporated Fisher information is maximal. In the following, we will generally introduce the Fisher information and state some of its properties. Afterwards we will concentrate on the case of a penalized least squares regression model. In this case, the one-to-one correspondence of variance reduction and the maximization of Fisher information will become clear and we will obtain a closed form solution for the variance reduction query strategy.

### 4.3 The Meaning of the Fisher Information

In this section, we will assume that the  $x = \{x_i\}_{1 \leq i \leq n}$  are not random as we actively choose them ourself. That means that one should keep in mind that all expectations only integrate over  $y \in Y^n$ . We denote by  $p(y; \hat{w}, x)$  a parametrized family of distributions of  $y$  on  $Y^n$ . Let  $w$  be the true parameter governing the distribution of  $y$ . Note that by assuming that such a 'true'  $w$  exists we are in the realm of frequentists statistics. If not explicitly stated otherwise we always integrate with respect to  $p(y; x, w)$ , i.e. the true distribution of  $y$ . We denote by  $\hat{w}$  an estimate of  $w$ .

Notation

Choosing the training set with the maximal incorporated Fisher information is the means of choice in the statistical theories of optimal experimental design (OED). The Fisher Information is a way of measuring the amount of information that an observable random variable carries about an unknown parameter  $\hat{w}$ . Since we are dealing with a multi-dimensional parameter vector in the application, we will always describe the multi-dimensional version below.

Fisher Information

#### Definition 4.1: Fisher Information

Consider a  $n$ -dimensional parameter vector,  $\hat{w} = (\hat{w}_1, \dots, \hat{w}_m) \in \mathbb{R}^m$ . Let be  $p(y; x, \hat{w})$  the likelihood distribution for a given value  $x$  and parameter  $\hat{w}$ . Then the Fisher information  $\mathcal{I}(\hat{w}) \in \mathbb{R}^{m \times m}$  takes the form of a symmetric matrix and is given by

$$\mathcal{I}_{i,j}(x, \hat{w}) = \mathbb{E} \left[ \left( \frac{\partial}{\partial \hat{w}_i} \log p(y; x, \hat{w}) \right) \cdot \left( \frac{\partial}{\partial \hat{w}_j} \log p(y; x, \hat{w}) \right) \right]. \quad (4.3.1)$$

Thus it can be seen as the curvature of the log-likelihood function. Near the maximum likelihood estimate, low Fisher information therefore indicates that the maximum appears "blunt", that is, the maximum is shallow and there are many nearby values with a similar log-likelihood. Conversely, high Fisher information indicates that the maximum is sharp.

*Score Function*

The Fisher information is the variance of the so-called *score function*. The score function measures the sensitivity of the likelihood probability distribution  $p(y; \hat{w}, x)$  with respect to its parameter  $\hat{w}$  and is defined as the normalized partial derivative with respect to the parameter, i.e.

$$\begin{aligned} S(y; x, \hat{w}) &:= \frac{1}{p(y; x, \hat{w})} \cdot \frac{\partial}{\partial \hat{w}} p(y; x, \hat{w}) \\ &= \frac{\partial}{\partial \hat{w}} \log(p(y; x, \hat{w})). \end{aligned} \quad (4.3.2)$$

Thus, the score function takes the form of the partial derivative of the log-likelihood function. A short calculation shows that

$$\mathbb{E}[S(y; x, \hat{w})] = 0, \quad (4.3.3)$$

where we used an interchange of integral and partial derivative justified by the second regularity condition. The conditional variance is given by

$$\begin{aligned} \text{Var}(S(y; x, \hat{w})) &= \mathbb{E}[S(y; x, \hat{w})^2] - (\mathbb{E}[S(y; x, \hat{w})])^2 \\ &= \mathbb{E}[S(y; x, \hat{w})^2] \\ &= \mathcal{I}(x, \hat{w}), \end{aligned} \quad (4.3.4)$$

which takes the form of the Fisher Information.

*Cramér-Rao Inequality*

An *estimator*  $t : Y^n \rightarrow \mathbb{R}^m$  is a map  $t(y; x) = \hat{w}$ . We view it as a guess of  $w$  after we have seen some data  $y$ . Define

$$\tilde{\zeta}(\tilde{w}; x) := \mathbb{E}_{\tilde{w}}[t(y; x)].$$

Here,  $\tilde{\zeta}$  depends on  $\tilde{w}$  and  $x$  since we take the expectation with respect to the measure  $p(y; x, \tilde{w})$ . Generally we would like that  $t$  estimates the true parameter  $w$  as well as possible. One important property of an estimator is if it is *biased* or *unbiased*. We call  $t$  an unbiased estimator, if

$$\tilde{\zeta}(\tilde{w}; x) = \tilde{w}. \quad (4.3.5)$$

This would mean that the average of all guesses of  $t$  is the true  $w$ . But since  $y$  is a random variable, so is  $t(y; x)$  and it will vary around its mean. To know how much it will vary around the true parameter, the quantification of the variance is of great interest.



**Theorem 4.1: Cramér-Rao inequality, unbiased case**

Let  $w \in \mathbb{R}^m$  be the  $m$ -dimensional parameter vector governing the distribution of  $y \in Y^n$ . Let  $t : Y \rightarrow \mathbb{R}^m$  be an unbiased estimator of  $w$ , i.e. for all  $\tilde{w}$  we have  $\zeta(\tilde{w}; x) = \tilde{w}$ . Then the covariance matrix satisfies

$$\text{Var}(t(y; x)) \geq \mathcal{I}(w; x)^{-1}$$

where  $\mathcal{I}(w; x) \in \mathbb{R}^{m \times m}$  denotes the Fisher information matrix evaluated at the true parameter  $w$ . The matrix inequality  $A \geq B$  means that  $A - B$  is positive semidefinite.

This tells us that given some samples  $x$  and a true model governing the creation of the  $y$ , we have a lower bound on the variance of any unbiased estimator of  $w$ . An upper bound is not possible in this generality as it heavily depends on the form of the estimator  $t$ . But this gives a criterion for an optimal estimator that we can try to attain. In the next section, we will specify to the case of linear regression. As we have seen, unpenalized least squares coincides with the maximum likelihood estimator. These are unbiased estimators. But using a Tikhonov regularization biases  $\hat{w}$  towards a small  $\|\Gamma \hat{w}\|_2$ . Also, in the Bayesian setting, we saw that using regularization corresponds to choosing a prior. These prior assumptions make  $\hat{w}$  a biased estimator. This means we also need the Cramér-Rao inequality for biased estimators:

**Theorem 4.2: Cramér-Rao inequality, biased case**

Let  $w \in \mathbb{R}^m$  be the  $m$ -dimensional parameter vector governing the distribution of  $y \in Y^n$ ,  $t : Y \rightarrow \mathbb{R}^m$ . Then the covariance matrix satisfies

$$\text{Var}(t(y; x)) \geq \frac{\partial \zeta(w; x)}{\partial w} \mathcal{I}(w; x)^{-1} \frac{\partial \zeta(w; x)^T}{\partial w} \quad (4.3.6)$$

where  $\mathcal{I}(w; x) \in \mathbb{R}^{m \times m}$  denotes the Fisher information matrix evaluated at the true parameter  $w$ .  $\frac{\partial \zeta(w; x)}{\partial w}$  denotes the Jacobian matrix of  $\zeta(w; x)$  as map  $\mathbb{R}^m \rightarrow \mathbb{R}^m$  evaluated at  $w$ .

In the unbiased case,  $\zeta$  is the identity when viewing it only as a map from  $w$ . This means that the Jacobian will also be the identity and the statement is the same as of the theorem before. In the following we will denote the right hand side of 4.3.6 by *Cramér-Rao bound*.

In this section, we summarized important properties of the Fisher information in the general case. In the following, we will turn to the special case of a penalized least squares regression. In this special case, we will obtain that penalized linear regression actually achieves the Cramér-Rao bound. Thus a maximization of the Fisher information directly minimizes the variance. We will be able to find a closed form for the variance reduction query in the linear regression case.

## 4.4 Application to the Linear Regression Model

### 4.4.1 Cramér-Rao for Linear Regression

In a linear regression model, we assume that  $y$  is distributed according to

$$y = Aw + \varepsilon \quad \text{where } \varepsilon \sim \mathcal{N}(0, \sigma^2 I) \quad (4.4.1)$$

for a matrix

$$A = \begin{pmatrix} \phi_1(x_1) & \dots & \phi_m(x_1) \\ \vdots & \ddots & \vdots \\ \phi_1(x_n) & \dots & \phi_m(x_n) \end{pmatrix} \in \mathbb{R}^{n \times m}.$$

with a feature map  $\Phi = (\phi_1, \dots, \phi_m)$ , vector of weights  $w \in \mathbb{R}^m$  and  $n$  data points  $x_i$ . The estimate  $\hat{w}$  is given by the solution to the penalized least squares regression problem.

*Fisher  
Information*

Let us have a closer look at the specific form of the Fisher information matrix in the case of linear regression. Therefore, note that the likelihood function is given by

$$p(y; A, \tilde{w}) \sim \mathcal{N}(A\tilde{w}, \sigma^2 I). \quad (4.4.2)$$

Taking the logarithm of the density of (4.4.2) results in the log-likelihood function, given by

$$\begin{aligned} l(y; \tilde{w}, A) &= \log p(y; A, \tilde{w}) \\ &= -\frac{1}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} |y - A\tilde{w}|^2. \end{aligned} \quad (4.4.3)$$

Since the log-likelihood function is twice differentiable, we are able to compute its Hessian matrix

$$\begin{aligned} H(y; \tilde{w}, A) &= \frac{\partial^2}{\partial w^2} l(y; \tilde{w}, A) \\ &= -\frac{1}{2\sigma^2} \frac{\partial^2}{\partial w^2} (y^T y - 2y^T A\tilde{w} + \tilde{w}^T A^T A\tilde{w}) \\ &= -\frac{1}{2\sigma^2} \frac{\partial}{\partial w} (-2y^T A + 2\tilde{w}^T A^T A) \\ &= -\frac{1}{2\sigma^2} 2A^T A \\ &= -\frac{A^T A}{\sigma^2}. \end{aligned} \quad (4.4.4)$$

Since the Hessian is independent of  $y$ , the Fisher information is given by

$$\mathcal{I}(\tilde{w}; x) = -\mathbb{E}_w[H(y; \tilde{w}, A)] = \frac{A^T A}{\sigma^2} \in \mathbb{R}^{n \times n}. \quad (4.4.5)$$

*Prediction Vari-  
ance*

Before we proceed and explain how we want to minimize such a matrix quan-

tity let us see how our actual variance compares to the Cramér Rao-bound. Recalling that by the normal equations 2.1.16 it holds

$$\hat{w}(y) = (A^T A + \Gamma^T \Gamma)^{-1} A^T y$$

Now we calculate  $\zeta$ :

$$\begin{aligned} \zeta(\tilde{w}) &= \mathbb{E}_{\tilde{w}}[\hat{w}(y)] \\ &= \mathbb{E}_{\epsilon}[\hat{w}(A\tilde{w} + \epsilon)] \\ &= (A^T A + \Gamma^T \Gamma)^{-1} A^T (A\tilde{w}) + (A^T A + \Gamma^T \Gamma)^{-1} A^T \mathbb{E}_{\epsilon}[\epsilon] \\ &= (A^T A + \Gamma^T \Gamma)^{-1} A^T A\tilde{w} \end{aligned} \quad (4.4.6)$$

In the second step we used (4.4.2). In the third step we used linearity of the integral twice. In the last step we use that  $\epsilon$  has a zero mean. Here again one sees that if  $\Gamma = 0$  we would obtain an unbiased estimator as everything would cancel out. In the case of  $\Gamma \neq 0$  the estimator is biased as the average of its predictions is not equal to  $\tilde{w}$  when the data is sampled from  $p(y; A, \tilde{w})$ . To calculate the lower bound in the theorem we need to take the derivative of  $\zeta$ . This is easy as  $\zeta$  is linear:

$$\frac{\partial \zeta(\tilde{w})}{\partial \tilde{w}} = (A^T A + \Gamma^T \Gamma)^{-1} (A^T A)$$

By (4.4.5) we know that  $\mathcal{I}(\tilde{w}; x) = \frac{A^T A}{\sigma^2}$ . We conclude that the Cramér-Rao bound is

$$\frac{\partial \zeta(w; x)}{\partial w} \mathcal{I}(w; x)^{-1} \frac{\partial \zeta(w; x)}{\partial w}^T = \sigma^2 G G^T$$

with  $G = (A^T A + \Gamma^T \Gamma)^{-1} A^T$ . We want to compare this to the actual variance of the estimator:

$$\begin{aligned} \text{Var}_w[\hat{w}(y)] &= \text{Var}_{\epsilon}[\hat{w}(Aw + \epsilon)] \\ &= \text{Var}_{\epsilon}[\hat{w}(Aw) + \hat{w}(\epsilon)] \\ &= \text{Var}_{\epsilon}[\hat{w}(\epsilon)] \\ &= (A^T A + \Gamma^T \Gamma)^{-1} A^T \mathbb{E}_{\epsilon}[\epsilon \epsilon^T] A (A^T A + \Gamma^T \Gamma)^{-1} \\ &= \sigma^2 G G^T \end{aligned} \quad (4.4.7)$$

where in the last step we used that  $\epsilon \sim \mathcal{N}(0, \sigma^2 I)$ . We actually attain the lower bound which means that we in some sense have an optimal estimator. Minimizing the lower bound is equivalent to minimizing the actual variance of our predictor. Since  $\hat{y} = A\hat{w}$  and the prediction is a linear map of  $\hat{w}$  the covariance of our actual predictions  $\hat{y}$  is directly related to the covariance of  $\hat{w}$ :

$$\text{Var}_w[\hat{y}] = A \text{Var}_w[\hat{w}] A^T$$

From this we conclude that in our case minimizing the lower bound of the Cramér-Rao inequality is equivalent to minimizing the variance of the predictor and minimizing the variance in our actual predictions. This means we

can actually directly minimize the variance in (4.2.1). For the case that  $\Gamma = 0$  we have the least squares or maximum likelihood estimator. The fact that these are ideal estimators and attain the Cramér-Rao bound is known under the name *Gauss-Markov theorem*. We see that it even holds for a general regularized least square regression problem. In the following we will minimize the determinant of a matrix which goes under the name of *D-optimality*. Minimizing the determinant of  $G$  is quite costly as it involves calculating matrix products and inverses. We will thus treat the unpenalized case in which  $G = A^{-1}$  and we will assume  $G$  to be a square matrix. This will lead to a very cost-efficient way to minimizing the determinant of  $A^{-1}$  and should also influence the determinant of  $G$  in a similar way. In the numerical experiments we will compare mini

#### 4.4.2 Minimizing the Cramér-Rao bound

Three Optimal  
Design Types

To minimize the Cramér-Rao bound on the variance in the data, we will try to choose  $A$  s.t. the Fisher information is maximized. When there is only one parameter in the model, the Fisher Information takes the form of a one-dimensional continuous function and the maximization is straight forward. In the multi-dimensional case, however, we need to further characterize how to maximize the information. Having a model dependent on  $N_k$  parameters, the Fisher information takes the form of a  $n \times n$  covariance matrix. In the OED literature, there are three common types of optimal design in such cases:

- *A-optimality* minimizes the trace of the inverse information matrix and thus minimizes the averaged variance of  $\hat{w}$ ,
- *D-optimality*, which minimizes the determinant of the inverse matrix and thus maximizes the volume of the confidence ellipsoid of  $\hat{w}$ ,
- *E-optimality*, which minimizes the maximal eigenvalue of the inverse matrix and thus minimizes the maximal possible value for the variance of the one-dimensional components of  $\hat{w}$ .

D-Optimality

In the following, we will concentrate on the concept of D-optimality. So far, we explained the link between covariance and Fisher information matrix, but did not explore how one should choose the training data in order to obtain maximal Fisher information. To this end, we will obtain a closed form solution for a D-optimality ansatz in the linear regression case. For a large data set  $T$ , we aim to choose the optimal training data  $\hat{D} \subseteq T$  in the way that minimizes the determinant of the inverse fisher information. Thus

$$\begin{aligned}
 \hat{D} &:= \operatorname{argmin}_{D \subseteq T} \det(\mathcal{I}(w)^{-1}) \\
 &= \operatorname{argmax}_{D \subseteq T} \det(\mathcal{I}(w)) \\
 &= \operatorname{argmax}_{D \subseteq T} \det\left(\frac{A^T A}{\sigma^2}\right) \\
 &= \operatorname{argmax}_{D \subseteq T} \det(A^T A)
 \end{aligned} \tag{4.4.8}$$

Hence, we are interested in choosing the training set which maximizes the absolute value of the determinant of the matrix  $AA^T$ . In the following, we will define a query strategy, which gradually chooses the optimal data point in each step. To define the query strategy, we follow the maxvol algorithm.

The maxvol algorithm is based on an application of the Cramer's rule. Cramer's rule observes how the determinant of a quadratic matrix changes, when replacing one row. This directly applies to our problem, since the addition of one data point comes in by changing singular rows of the matrix  $A$ .

*Cramer's Rule*

#### Theorem 4.3: Cramer's Rule

Consider a system of  $n$  linear equations and  $n$  unknowns, represented in matrix multiplication by

$$Ax = b, \quad (4.4.9)$$

where  $A \in \mathbb{R}^{n \times n}$  and  $\det(A) \neq 0$ , known values  $b \in \mathbb{R}^n$  and unknown coefficients  $x \in \mathbb{R}^n$ . Then (4.4.9) has a unique solution and the coefficients are given by

$$x_i = \frac{\det(A_i)}{\det(A)}, i = 1, \dots, n \quad (4.4.10)$$

where  $A_i \in \mathbb{R}^{n \times n}$  is the matrix formed by replacing the  $i$ th column with  $b$ , i.e.

$$A_i = \begin{pmatrix} a_{1,1} & \dots & a_{1,i-1} & b_1 & a_{1,i+1} & \dots & a_{1,n} \\ \vdots & & \vdots & \vdots & \vdots & & \vdots \\ a_{n,1} & \dots & a_{n,i-1} & b_n & a_{n,i+1} & \dots & a_{n,n} \end{pmatrix} \quad (4.4.11)$$

#### 4.4.3 Query Strategy with fixed Training Size

In the following, we assume the number of data points in the training set to be fixed. But we do not fix the number in an arbitrary way, but we make sure that we include exactly as many data points, such that the matrix  $A$  is quadratic. Thus, we are interested in choosing  $m$  data points in order to obtain  $A \in \mathbb{R}^{m \times m}$ . We are also interested in searching the training set  $\mathcal{D}_m \subset T$  which is D-optimal. This way, the form of the optimal data set (4.4.8) can be further simplified

*Fixed Training Size*

$$\begin{aligned} \hat{\mathcal{D}}_m &= \operatorname{argmax}_{\mathcal{D} \subseteq T, |\mathcal{D}|=m} \det(A^T A) \\ &= \operatorname{argmax}_{\mathcal{D} \subseteq T, |\mathcal{D}|=m} \det(A)^2 \\ &= \operatorname{argmax}_{\mathcal{D} \subseteq T, |\mathcal{D}|=m} |\det(A)|. \end{aligned} \quad (4.4.12)$$

Maximizing the determinant of  $AA^T$  reduces to a maximization of the determinant of  $A$ . This opens up the possibility to define a query strategy directly

based on Cramer's rule in Theorem 4.3 in order to avoid very expensive determinant calculations.

*Information Measure*

Following the maxvol algorithm we form the row vector

$$C := (\phi_1(x^*), \dots, \phi_m(x^*)) \cdot A^{-1}, \quad (4.4.13)$$

for a data point  $x^* \in T$ . If we replace the  $k$ th row of the matrix  $A$  by the vector  $\Phi(x^*) = (\phi_1(x^*), \dots, \phi_m(x^*))$  then, using Cramer's rule,  $|\det(A)|$  changes with a factor  $|C_k|$ ,

$$|\det(A_k)| = |C_k| \cdot |\det(A)|. \quad (4.4.14)$$

Where  $A_k$  is  $A$  with the  $k$ -th row replaced by  $\Phi(x^*)$ . Thus, if  $|C_k| > 1$  holds the absolute determinant increases when replacing the  $k$ -th row by  $\Phi(x^*)$ . Based on that, we define

$$\psi(x^*) := \max_{1 \leq k \leq m} |C_k|, \quad (4.4.15)$$

which describes the maximal possible increase of the absolute determinant, when replacing a row of  $A$  with  $\Phi(x^*)$ . Or equivalently, it is a measure for the maximal information we gain, when incorporating  $x^*$  in the training set.

*Query Strategy*

The resulting query strategy is as follows. Let  $\gamma > 0$  be a prescribed value, which encodes the allowed extrapolation grade of our model. We start with a randomly chosen training set containing  $m$  data points. Based on the underlying least squares regression problem, we set up the matrix  $A$  according to the training set. Iteratively, we run through all the data points  $x^*$  which are not yet included in the training set and compute their information measures  $\psi(x^*)$ . We add each instance  $x^*$  to the training set if it improves the determinant in a way such that

$$\psi(x^*) > 1 + \gamma. \quad (4.4.16)$$

When adding a data point to the training set another data point has to be removed in order to preserve the training size. Therefore, we remove  $x^{\hat{k}}$ , where  $\hat{k} = \operatorname{argmax}_k |C_k|$ . Otherwise, the training set set remains unchanged. This is repeated, until the absolute determinant of the matrix  $A$  can not be improved any further.

**Data:**  $\gamma > 0$

**Result:** Matrix  $A$

Assemble quadratic initial matrix  $A$  samples ;

**while**  $\exists x^*$  with  $\psi(x^*) > 1 + \gamma$  **do**

    | let  $\hat{k} = \operatorname{argmax}_k |C_k|$  with  $C_k$  as in 4.4.14 ;

    | replace  $\hat{k}$ th row of  $A$  by  $\Phi(x^*)$  ;

**end**

**Algorithm 1:** Active Learning

$\Phi(x^*)$  can be seen as a grade of dissimilarity to the other data points. Hence in [49], is called the *extrapolation grade* and  $\gamma$  is called the maximal allowed extrapolation grade.

## Chapter 5

# The Born-Oppenheimer Potential Energy Surface

The Born Oppenheimer-Potential Energy Surface (BO-PES) is a concept based on the simplifying approach of Born and Oppenheimer who decomposed the Schrödinger's equation into separate parts describing the slow large nuclei and the fast small electrons. A Potential Energy Surface aims to describe the energy of a quantum mechanical system solely by knowing the positions of its nuclei. Since an exact approximation of the PES by exact QM methods is computationally demanding, one is especially interested in further simplifying the setting and limiting exact calculations to a minimum.

This chapter is structured as follows. In section 5.1 we will introduce the concept of the Born-Oppenheimer Potential Energy Surface in more detail. In section 5.2 we will state the well known atomic decomposition ansatz. To open up the possibility to describe infinite crystal structures and reduce the dimension, we will decompose the energy of the whole system into energy contributions of smaller subsystems. In section 5.3 we state physical and computational assumptions on a suitable PES. A further decomposition of the smaller subsystems is described by the many body decomposition in section 5.4. In section 5.5 we broach the issue on how to decipher a quantum mechanical system containing nuclei and electrons in numbers which can serve as an input for an algorithm.

For a comprehensive introduction to quantum mechanics, we refer to [71] and [57].

### 5.1 The Potential Energy Surface

The Born-Oppenheimer potential energy surface is a concept of quantum mechanics (QM). The field of quantum mechanics goes back to the 20th century and deals with the description of the smallest particles, atoms and even smaller ones. In 1926 Schroedinger laid the foundation for modern quantum mechanical development with the publication of the *Schrödinger's equation*

*Schrödinger's Equation*

tion (SE), for which he received the Nobel Prize in Physics in 1933 [55]. Where Newtonian mechanics dealt with classical mechanics, Schrödinger's equation was now able to describe changes over time of a quantum mechanical system consisting for example of nuclei and electrons. Given a QM system, the time-independent Schrödinger equation comes in the form of an eigenvalue problem,

$$\mathcal{H}\phi = \mathcal{E}\phi, \quad (5.1.1)$$

where  $\mathcal{H}$  is a suitable Hamiltonian, a wave function  $\phi$  which uniquely describes one state of the QM system, and  $\mathcal{E}$  describing its corresponding energy. Let us consider an atomic configuration consisting of  $M$  nuclei, given by its atomic numbers and spatial coordinates

$$(\mathbf{Z}, \mathbf{R}) = (Z_i, R_i)_{i=1}^M \in (\mathbb{Z} \times \mathbb{R}^3)^M$$

and  $N$  electrons given by their coordinates  $\mathbf{r} = (r_i)_{i=1}^N \in (\mathbb{R}^3)^N$ . Here the atomic number, or proton number,  $Z_A$  of a nucleus  $A$  is the number of protons found in the nucleus and describes its charge. In this situation, the time-independent Schrödinger equation depends on  $3 \cdot (M + N) + M$  dimensions and the associated Hamilton operator reads

$$\begin{aligned} \mathcal{H}(\mathbf{Z}, \mathbf{R}, \mathbf{r}) = & - \sum_{i=1}^N \frac{1}{2} \nabla_i^2 - \sum_{A=1}^M \frac{1}{2M_A} \nabla_A^2 \\ & - \sum_{i=1}^N \sum_{A=1}^M \frac{Z_A}{d_{iA}} \\ & + \sum_{i=1}^N \sum_{j>i}^N \frac{1}{d_{ij}} + \sum_{A=1}^M \sum_{B>A}^M \frac{Z_A Z_B}{d_{AB}}. \end{aligned} \quad (5.1.2)$$

In the above equation,  $M_A$  is the ratio of the mass of nucleus  $A$  to the mass of an electron and  $d_{iA} := \|r_i - R_A\|_2$  denotes the distance of electron  $i$  and nucleus  $A$ ,  $d_{ij}$  and  $d_{AB}$  describe the distance between two electrons and nuclei respectively. The Laplacian operators  $\nabla_i^2$  and  $\nabla_A^2$  involve differentiation with respect to the coordinates of the  $i$ -th electron and the  $A$ -th nucleus. The first summand in equation (5.1.2) is the operator for the kinetic energy of the electrons, the second term is the operator for the kinetic energy of the nuclei. The kinetic energy of an object is the energy that it possesses due to its motion. The third term represents the Coulomb attraction between electrons and nuclei. Electrons and nuclei are charged particles. The electrostatic potential (Coulomb potential) at any position in space  $r_i$  is the energy required to bring a single positive charge from infinity to that point. Assuming only one single nucleus  $A$  to be present, the Coulomb potential is given by  $Z_A/d_{iA}$  in atomic units. An electron moving in this potential possesses the potential energy  $-Z_A/d_{iA}$ . The fourth and fifth terms represent the repulsion between electrons and between nuclei, respectively.

Unfortunately the time-independent Schrödinger equation is only analytically solvable for, besides systems containing just a few atoms, hydrogen or hydrogen-like atoms. Taking just a single particle, the Schrödinger equation is already



a complex second order differential equation in three dimensions and even in the classical mechanics there is no general solution for the three body problem. A first simplification has been made by the Born Oppenheimer approximation.

Still in beginnings of the early period of quantum mechanics, Born and Oppenheimer came up with the idea to decouple the behaviour of electrons and nuclei to simplify the approach of Schrödinger. Due to the enormous difference in mass and movement of electrons and nuclei, the nuclei are regarded to be motionless compared to the fast electrons. Fixing a configuration of slow nuclear positions, one still has to solve the Schrödinger equation for the fast electrons. This equation is called the *electronic Schrödinger's equation*,

*Born-  
Oppenheimer  
Approximation*

$$\mathcal{H}_e \phi_e = \mathcal{E}_e \phi_e. \quad (5.1.3)$$

Here the *electronic Hamiltonian*  $\mathcal{H}_e$  describes the motion of  $N$  electrons in the field of  $M$  point charges

$$\begin{aligned} \mathcal{H}_e(\mathbf{Z}, \mathbf{R}, \mathbf{r}) &= \mathcal{H}_{ee}(\mathbf{Z}, \mathbf{R}, \mathbf{r}) + \sum_{A=1}^M \sum_{B>A}^M \frac{Z_A Z_B}{d_{AB}}, \\ \mathcal{H}_{ee}(\mathbf{Z}, \mathbf{R}, \mathbf{r}) &= - \sum_{i=1}^N \frac{1}{2} \nabla_i^2 + \sum_{i=1}^N \sum_{j>i}^N \frac{1}{d_{ij}} - \sum_{i=1}^N \sum_{A=1}^M \frac{Z_A}{d_{iA}}. \end{aligned} \quad (5.1.4)$$

This approximation allows us to focus on the electronic energy first and adding the repulsion term of nuclei later.

The time-independent Schrödinger equation and its electronic counterpart comes in the form of an eigenvalue problem. Since the Hamiltonian is an Hermitian operator, the eigenvalues are real and the corresponding eigenfunctions can be chosen orthonormal. The *ground state* of a quantum mechanical system is its lowest energy state and hence, regarding the spectrum of the Hamiltonian operator, the eigenvalue at the bottom of the spectrum. On the contrary,

*Ground  
Energy*

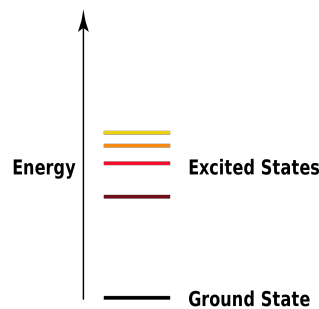


Figure 5.1: Energy levels.

an excited state is any state with energy greater than the ground state energy. An excited state is generally unstable and decays into a less excited state eventually, where the time span can range from a split of a second to many years. Hence, a closed physical system typically seeks the state of minimal energy.

Thus, we are especially interested in the ground state for a fixed atomic configuration, which forms the Born Oppenheimer potential surface.

*Born-  
Oppenheimer  
Potential  
Energy  
Surface  
(BO-PES)*

We define the *Born Oppenheimer Potential Energy Surface* (BO-PES) as the minimal eigenvalue of the eigenvalue problem (5.1.3) fixing one specific atomic configuration  $(\mathbf{Z}, \mathbf{R})$ ,

$$V^{BO}(\mathbf{Z}, \mathbf{R}) := \inf_{\|\phi\|=1} \left\{ \int \phi^*(\mathbf{r}) \mathcal{H}_e(\mathbf{Z}, \mathbf{R}, \mathbf{r}) \phi(\mathbf{r}) d\mathbf{r} \right\} \quad (5.1.5)$$

A solution for the ground state exists by Zhislin's theorem and its corresponding eigenspace is finite dimensional if the charged system is neutral or positively charged, i.e.  $\sum_{A=1}^M Z_A > N - 1$  [28, 72]. Since the Hamiltonian is a self-adjoint operator and the wave functions  $\phi$  are elements of an Hilbert space, we can therefore conclude that there are finitely many eigenfunctions corresponding to the minimal eigenvalue. In the following work, we will concentrate on the approximation of (5.1.5) given a fixed atomic configuration  $(\mathbf{Z}, \mathbf{R})$ . Note, that we have to handle a very high dimensional function

$$V^{BO} : (\mathbb{R}^3 \times \mathbb{Z})^M \rightarrow \mathbb{R},$$

which moreover depends on the number of nuclei  $M$ , thus taking an arbitrary physical situation containing nuclei and electrons, we are ending up with different dimensions of the PES. Note furthermore, that we can not assume differentiability of the PES due to its specific construction as the minimal eigenvalue of an eigenvalue problem. It could happen, that the least and the second smallest eigenvalue approach each other and eventually switch positions. This situation is called an *eigenvalue crossing* and the reason why we can't assume differentiability for  $V^{BO}$ .

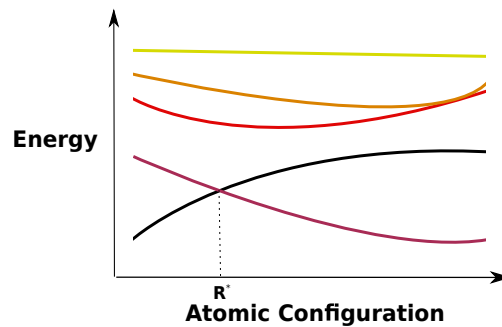


Figure 5.2: Eigenvalue Crossing. No differentiability of  $V^{BO}$  at  $\mathbf{R}^*$ .  $\mathbf{R}^*$  is called a *bifurcation point*.

Moreover, having a look at the construction of the electronic Hamiltonian (5.1.4), we note that an eigenvalue of  $\mathcal{H}_e$  is also an eigenvalue of  $\mathcal{H}_{ee}$ . Unfortunately  $\nabla \mathcal{H}_{ee}$  may be unbounded due to singularities of the Coulomb potential.

*Minima and  
Saddle Points*

Intuitively, we think of a PES as a function mapping spacial coordinates and

charges of nuclei onto its corresponding energy. In general, the nuclei are assumed to be stationary. Thus, we neglect that an atom can occupy different levels of vibrational energy due to the surrounding temperature. Especially, one is interested in local minima and saddle points of a PES. Is a atomic configuration located in a minimum, or physically speaking in a metastable state, the smallest random perturbation of its coordinates would lead to an higher energy. Is a configuration located at a saddle point of the PES, a random perturbation of the coordinates in each direction but one would lead to a smaller energy except in one direction. This state is called a transition state and each perturbation leading to a smaller energy leads to a different metastable state. Thus, one can observe the metastable states of a PES and its paths between them forming a trajectory over the potential energy hyper surface.

In summary, we keep in mind, that the problem is extremely high dimensional and irregular due to its dimensional dependence on the specific physical situation, and we can not assume differentiability. Moreover it is only analytically solvable for very few simple situations, which is why we have to resort on numerical approximation. For a comprehensive introduction to quantum mechanics, we refer to [71] and [57].

## 5.2 The Atomic Decomposition Ansatz

Motivated by the concept of nearsightedness of electronic matter introduced by Kohn [44], a usually taken assumption is that if two nuclei are far apart from each other, the interaction potential between them can be neglected. This ansatz allows us to decompose the energy of a system into energies of fractional subsystems, which are in a sense closed under interaction. For a threshold  $r_{\text{cut}} > 0$ , one only assumes an interaction between nuclei located at distance  $r_{\text{cut}}$  or closer. The choice of the maximal interaction distance crucially depends on the precise nuclei considered. Two hydrogen atoms are for example typically at a distance of  $0.74\text{\AA}$ , whereas two fluorine atoms are usually  $1.42\text{\AA}$  apart [1]. It is therefore not surprising that a change in the cut-off radius brings huge differences not only in the dimension of the approximation problem, but also in the approximation power. Lets fix a configuration containing  $M$  nuclei described by  $\mathbf{X} = (X_1, \dots, X_M)$ , where

$$X_i := (Z_i, R_i),$$

with atom number  $Z_i \in \mathbb{N}$  and spatial coordinates  $R_i \in \mathbb{R}^3$ . Denote by  $\mathbf{N}_i$ , the *local atomic environment* or the *local neighborhood* of the  $i$ -th nuclei given by

$$\mathbf{N}_i := (X_j)_{\substack{j=1, \dots, M \\ \|R_i - R_j\|_2 \leq r_{\text{cut}}}}.$$

For a visualization see Figure 5.3. The local atomic environment contains all nuclei of the system, which affects the  $i$ -th nuclei in any way by being closer than  $r_{\text{cut}}$ . The local atomic environment is obviously not necessary of the same dimension for every nuclei of an arbitrary system. There could be a nucleus

*Local  
Atomic  
Environment*

which is so far apart from the others, that there is no interaction with another nucleus. The atomic environment of this nucleus would contain only itself. The other extreme is a nucleus affecting every other nucleus in the the system, such that the atomic environment is again the whole system.

The  
Atomic  
Decomposition  
Ansatz

The aim of this thesis is to find a suitable *potential function*

$$V : (\mathbb{N} \times \mathbb{R}^3)^M \longrightarrow \mathbb{R},$$

given the effort, we are willing or able to spend. The potential function  $V$  spans a corresponding potential energy surface (PES) approximating the very high-dimensional Born Oppenheimer Potential Energy Surface (BO-PES) spanned by  $V^{BO}$ . From now on, we make the simplifying assumption that we can represent the global potential function  $V$  as the sum over *local atomic potential functions*  $V_{\text{atomic}}$ , which describes only a part of the system. We assume that the potential function can be written as a sum of atomic functions depending only on local atomic environments of single nuclei.

#### Definition 5.1: Atomic Decomposition

Let be  $\mathbf{X} = (X_1, \dots, X_M)$  an atomic configuration containing  $M$  nuclei and  $r_{\text{cut}} > 0$  the maximal interaction distance. Then, we call the decomposition

$$V(\mathbf{X}) = V(X_1, \dots, X_M) = \sum_{i=1}^M V_{\text{atomic}}(\mathbf{N}_i) \quad (5.2.1)$$

the *atomic decomposition*. Here,  $\mathbf{N}_i$  is given by

$$\mathbf{N}_i = (X_j)_{\substack{j=1, \dots, M \\ \|R_i - R_j\|_2 \leq r_{\text{cut}}}}$$

and is called the *local atomic environment* or neighborhood of the  $i$ -th nuclei.

Here, the *atomic potential function*  $V_{\text{atomic}}$  is of varying dimensionality, due to the varying dimensionality of the input. Although physically motivated, this

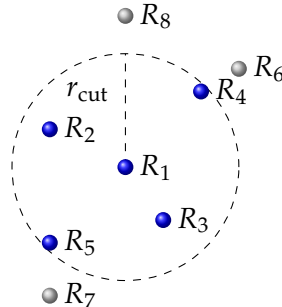


Figure 5.3: The local atomic environment of the first nuclei with respect to the cut-off radius includes all atoms which are closer than  $r_{\text{cut}}$ ,  $\mathbf{N}_1 = (X_1, X_2, X_3, X_4, X_5)$

approach leads to a numerical feasibility by a drastic reduction of the functional dimension, from the information of the whole system to the information of local atomic environments. This locality assumption (5.2.1) is true in most systems with short-range interactions (as opposed to, e.g., Coulomb interaction in charged or polarized systems). For rigorous proofs of this statement for simple QM models we refer to recent works [18, 48].

### 5.3 Physical and Computational Assumptions on the Potential Function

We saw previously that we can assume  $V^{BO}$  (5.1.5) to be continuous but not differentiable, due to the definition as the minimal eigenvalue of an eigenvalue problem. Nevertheless, we can limit the potential functions in question by taking meaningful physical requirements into account. Therefore, we will state physical assumptions on a suitable atomic potential function

$$V_{\text{atomic}} : (\mathbb{N} \times \mathbb{R}^3)^n \rightarrow \mathbb{R},$$

for an arbitrary dimension  $n \leq M$ . As an input we think of a neighborhood of one nucleus containing  $n - 1$  atoms which are closer than a threshold  $r_{\text{cut}}$  and the centering nucleus itself. By the atomic decomposition ansatz (5.2.1), those assumptions will also apply to the resulting potential function  $V$ . Take  $n$  arbitrary nuclei of the system  $X_1, \dots, X_n \in \mathbf{X}$ . First, we should assume that a QM system has the same energy after simply renaming the nuclei, thus the potential function should be *permutational invariant*. I.e., for any  $\sigma \in \mathcal{S}_n$ ,

$$V_{\text{atomic}}(X_1, \dots, X_n) = V_{\text{atomic}}(X_{\sigma(1)}, \dots, X_{\sigma(n)}). \quad (5.3.1)$$

*Permutational  
Invariance*

Also, the energy should be invariant under rotation, reflection or translation of the whole system in space. I.e., for any  $Q \in O(3)$ , where  $O(3)$  denotes the orthogonal group in  $\mathbb{R}^3$ ,

$$V_{\text{atomic}}(X_1, \dots, X_n) = V_{\text{atomic}}(QX_1, \dots, QX_n). \quad (5.3.2)$$

*Transformational  
Invariance*

Here, the operation is applied to the position in space  $QX_i := (Z_i, QR_i)$  for  $i \in \{1, \dots, n\}$ . To deal with the varying dimensionality of the atomic potential function  $V_{\text{atomic}}$ , we have to integrate the nearsightedness in the atomic potential function. Introducing a nucleus  $X_{n+1}$  which is farther apart than  $r_{\text{cut}}$  from all the other nuclei should not have any effect to the energy. If  $\|R_{n+1} - R_i\|_2 \geq r_{\text{cut}}$  for all  $i \in \{1, \dots, n\}$ , then *continuity for nuclei entering and leaving* the interaction radius is ensured by assuming

$$V_{\text{atomic}}(X_1, \dots, X_n) = V_{\text{atomic}}(X_1, \dots, X_n, X_{n+1}). \quad (5.3.3)$$

*Continuity  
w.r.t. Cut-Off  
Radius*

### 5.4 The Many-Body Expansion

Fragment-based methods represent a promising path toward reducing the computational scaling with respect to the number of nuclei  $M$ . In such methods,

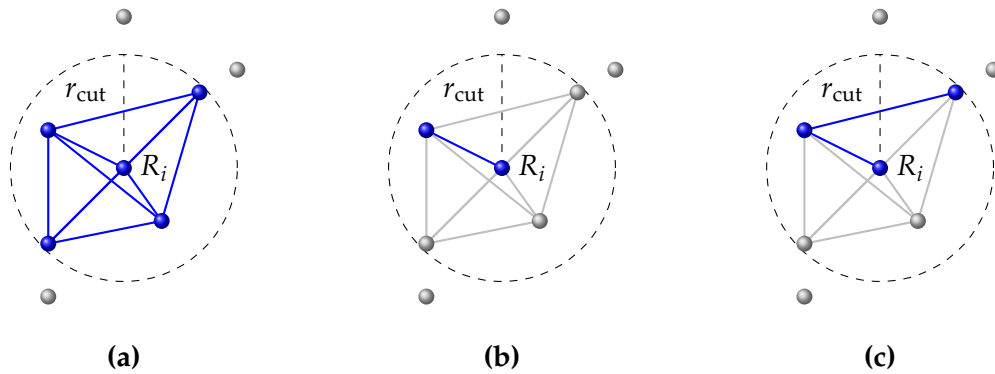


Figure 5.4: Fractional parts of a local atomic environment  $\mathbf{N}_i$ , where (a) describes the whole environment, (b) one two-body and (c) one three-body of it.

the energy of a local atomic environment,  $V_{\text{atomic}}(\mathbf{N}_i)$ , is further decomposed in lower-dimensional terms. The so-called *dimension wise decomposition* or *many-body expansion* is comparable to the well-known analysis of variance (ANOVA) decomposition.

#### Example

Let us Visualize the decomposition by a concrete example. Consider an atomic environment containing four atoms ( $n = 4$ ). Then, the atomic energy decomposes in energy contributions of all possible environmental fractions, which are all sets of one to four nuclei.

$$\begin{aligned}
 V_{\text{atomic}}(\text{---}) &= w_0(\text{---}) \\
 &+ w_1(\text{---}) + w_1(\text{---}) + w_1(\text{---}) + w_1(\text{---}) \\
 &+ w_2(\text{---}) + w_2(\text{---}) + w_2(\text{---}) \\
 &+ w_3(\text{---}) + w_3(\text{---}) + w_3(\text{---}) \\
 &+ w_4(\text{---}).
 \end{aligned} \tag{5.4.1}$$

A set of  $k$  nuclei will be in the following referred to as a  $k$ -body, which explains the designation ‘many-body expansion’. In the above expansion, the one, two, three and four-body terms are separated in rows. The first summand describes the energy of a completely empty configuration, containing no atom at all. In the following, we will neglect the so-called zero-point energy (first row) since it won’t affect the model in any way. Similarly, there is no influence between atoms in the one atom case (second row), these summands represent the energy due an external force field such as gravitational or electrostatic, which the system is immersed into. The first term which describes interactions of particles are the two-body terms (third row).

In general, the input dimension of  $V_{\text{atomic}}$  can differ, but can not be larger than the total number  $M$  of nuclei contained in the whole system  $\mathbf{X}$ . Therefore, a universal dimension wise decomposition like in (5.4.1) would involve all possible fractional functions  $w_0, \dots, w_M$ . In the previous example the evaluation of the functions  $w_5, \dots, w_M$  would just be zero, since there simply are no bodies of order greater than four in (5.4.1). For the general case, we obtain

$$V_{\text{atomic}}(\mathbf{N}_i) = \sum_{k=1}^M \sum_{\substack{U \subseteq \mathbf{N}_i \\ |U|=k}} w_k(U). \quad (5.4.2)$$

Here, the first sum runs over all possible orders of bodies  $k$  and the second sum runs over all  $k$ -bodies in the given local atomic environment. Combining the locality assumption we introduced in subsection 5.2 with the decomposition in lower-dimensional terms (5.4.2), we get

$$\begin{aligned} V(\mathbf{X}) &= \sum_{i=1}^M V_{\text{atomic}}(\mathbf{N}_i) \\ &\stackrel{(5.4.2)}{=} \sum_{i=1}^M \left( \sum_{k=1}^M \sum_{\substack{U \subseteq \mathbf{N}_i \\ |U|=k}} w_k(U) \right). \end{aligned} \quad (5.4.3)$$

Rearranging the summands, we obtain the equivalent form

$$\begin{aligned} V(\mathbf{X}) &= \sum_{i_1}^M w_1(\mathbf{X}_{\{i_1\}}) + \sum_{i_1 < i_2}^M w_2(\mathbf{X}_{\{i_1, i_2\}}) + \dots + w_M(\mathbf{X}) \\ &= \sum_{\substack{u \subseteq \{1, \dots, M\} \\ 0 < |u|}} w_{|u|}(\mathbf{X}_u), \end{aligned} \quad (5.4.4)$$

where  $\mathbf{X}_u = (X_i)_{i \in u}$ . Here, we directly sum up over all possible pairs, triples and so on of the whole system instead of summing first over each single nucleus and split up its atomic environment. Note that those expressions are indeed equivalent. In (5.4.3) we initially considered only local atom environments, where each nucleus lies within an interaction radius around one central atom, thus pairs of nuclei with a larger distance do not occur. In (5.4.4) we included all possible pairs and sets, which raises the question whether those expressions are equal. In fact, we assume the functions  $w_k$  to satisfy the same physical assumptions which we made for an atomic potential function. Consequently, the assumption of continuity for nuclei entering and leaving the cut-off radius, concludes to a vanishing of all additional terms we were concerned of.

**Definition 5.2: Many-Body Expansion**

Let be  $\mathbf{X} = (X_1, \dots, X_M)$  an atomic configuration containing  $M$  nuclei and  $r_{\text{cut}} > 0$  the maximal interaction distance. Then we call the decomposition

$$V(\mathbf{X}) = \sum_{\substack{u \subseteq \{1, \dots, M\} \\ 0 < |u|}} w_k(\mathbf{X}_u)$$

a *many-body expansion*, under the assumption that the functions  $w_u$  are permutational invariant, transformational invariant and invariant under nuclei entering and leaving the cut-off radius with respect to  $r_{\text{cut}}$ . To simplify notation, we will in the following use the shortcut

$$V(\mathbf{X}) = G(w_1, \dots, w_M)(\mathbf{X}).$$

Group  
Structure

Note, that the many-body expansion suggests that the fractional functions  $w_k$  properly describe all  $k$ -bodies. In application, those functions will again decompose in groups for which the corresponding sets  $\mathbf{X}_u$  are physically equivalent. For example, if  $\mathbf{X}_u$  can be obtained by a rotation in space of  $\mathbf{X}_{u'}$ ,  $w_u$  and  $w_{u'}$  are assumed to be equal. It is also interesting, that those sets  $u \subseteq \{1, \dots, M\}$ , which are physically equivalent are forming a group. Assuming the equivalence under the physical assumptions presented in section 5.3, associativity follows since if  $\mathbf{X}_u$  is equivalent to  $\mathbf{X}_{u'}$  and  $\mathbf{X}_{u'}$  is equivalent to  $\mathbf{X}_{u''}$ ,  $\mathbf{X}_u$  is equivalent to  $\mathbf{X}_{u''}$  as well. Also, inverse elements are included and the identity exists. Based on the application in the next chapter, we will explain this group structure in more detail.

K-Body  
Potential

Neglecting all bodies of order greater than some threshold  $k$  forms a general *k-body potential* or *interatomic potential of order k*,

$$V_k(\mathbf{X}) := \sum_{\substack{u \subseteq \{1, \dots, M\} \\ 0 < |u| \leq k}} w_k(\mathbf{X}_u).$$

Equivalently,

$$V_k(\mathbf{X}) = G(w_1, \dots, w_k)(\mathbf{X}).$$

As seen by this expansion, the pair potential,  $V_2$ , is the simplest possible model to evaluate atomic interactions. Though those potentials, such as the Lennard-Jones 6-12 or Morse potential can be fast and easily calculated for arbitrarily large atomic configurations, they are only rather good approximations for the simplest closed shell systems. For more complex situations, such as strongly covalent systems like semiconductors they appear completely inapplicable [64]. In practice, the zero-point energy term and the external force terms are usually ignored, while considering the sum up to pair-potentials or maximally three-body interaction [15, 24, 35, 65]. A Many-body approach for  $k \geq 3$  is rarely seen. An implementation of higher body terms is an intractable task and leads to a great increase of degrees of freedom. However, the general three-body



potentials, have been criticized for their inability to describe the energetics of all possible bonding geometries [10, 11, 64].

One may ask what we gained by this decomposition and the worth depends on the physical situation. Of course one aims to reduce the computational cost by assuming a decrease of energy contributions for increasing order and treating low dimensional terms with more effort, while treating higher dimensional terms with less. If we are able to neglect high order terms without losing too much information, we can break the curse of dimensions. If we are able to truncate the sum at all is also dependent on the situation. Nevertheless, a decay of the energy contributions for raising order is seen for most cases in which relatively weaker interaction energies are considered, and for most organic molecules [34].

*Truncating the Expansion*

## 5.5 Descriptors of Local Atomic Environments

How a physical situation of atoms and electrons can be deciphered in numbers is a much discussed topic. By choosing a *descriptor*, we choose how to describe a physical situation with numbers that can ultimately act as input to an algorithm. Until here, we explicitly assumed the main describing factors of a physical situation given by the involved atomic number and position in space. We made this choice, because the Schrodinger equation solely depends on these values. However, other parameters can be chosen. Regardless of this, there is a wealth of descriptors based on coordinates in space and atomic numbers, for a comprehensive overview see [66]. The choice of a well chosen descriptor has some advantages. On the one hand we are able to mostly overcome the physical assumptions on  $V_{\text{atomic}}$  by transforming the input variables into a format which has already some invariance properties. On the other hand, by filtering out excess information, we are able to reduce the input dimension.

*Choice of the Input*

Further investigations are based on the assumption of atomic decomposition and many body expansion, i.e.

*Descriptor Based On Our Assumptions*

$$V(\mathbf{X}) = \sum_{\substack{u \subseteq \{1, \dots, M\} \\ 0 < |u|}} w_{|u|}(\mathbf{X}_u),$$

for suitable functions  $w_u$ . Thus, we turn to descriptors, which describe those sets  $\mathbf{X}_u$ . The descriptor takes the form of a linear coordinate transformation, which maps the input parameters onto a meaningful description of an atomic environment

$$D : (\mathbb{R}^3 \times \mathbb{N})^{|u|} \rightarrow \mathbb{R}^{d(|u|)}$$

which takes a set of  $|u|$  nuclei and maps it to a description of dimension  $d(|u|)$ , which is usually a reduction. For a chosen descriptor the problem we end up

with

$$V(\mathbf{X}) = \sum_{\substack{u \subseteq \{1, \dots, M\} \\ 0 < |u|}} \hat{w}_u(\mathbf{X}_u),$$

for  $\hat{w} = w \circ D$ . To simplify notation, we will in the following still write  $w$  instead of  $\hat{w}$ .

**Example 5.1.** 1) If each atom in one configuration has the same atomic number, i.e.  $Z_i = Z_j$  for all  $i, j \in \{1, \dots, M\}$  the atomic numbers lose their meaning as a distinctive factor and a possible descriptor could filter out this information. Lets fix the  $i$ -th atom and consider  $n$  atoms with indices  $(t_1, \dots, t_n)$  lying in its cut off neighborhood given by  $r_{\text{cut}} > 0$ . Define the descriptor

$$\begin{aligned} D(\mathbf{N}_i) &= D(Z_{t_1}, \dots, Z_{t_n}, R_{t_1}, \dots, R_{t_n}) \\ &:= (|R_i - R_{t_k}|)_{k=1, \dots, n}, \end{aligned} \quad (5.5.1)$$

which maps the information onto the pairwise distances. The dimension reduces from  $(1 + 3)^n$  to  $d(n) = n$ .

2) Assume  $Z_i \neq Z_j$  for some  $i, j \in \{1, \dots, M\}$ , then one possible descriptor is given by the Coulomb potentials of each pair of atoms, i.e.

$$\begin{aligned} D(\mathbf{N}_i) &= D(Z_{t_1}, \dots, Z_{t_n}, R_{t_1}, \dots, R_{t_n}) \\ &:= \left( \frac{Z_i Z_{t_k}}{|R_i - R_{t_k}|} \right)_{k=1, \dots, n}, \end{aligned} \quad (5.5.2)$$

again reducing the dimension to  $d(n) = n$ .

### Revision

Lets recap what we've done so far. In this subsection we introduced the Born-Oppenheimer potential energy surface (BO-PES) as the minimal eigenvalue of an eigenvalue problem given by the electronic Schrödinger equation. Given a fixed system of nuclei, a PES describes the corresponding energy. We explained the complexity of the problem due to its high-dimensionality and non-differentiability. Afterwards, we made two simplifying assumptions. The first simplification we made was the atomic decomposition ansatz,

$$V(X) = \sum_{\text{nuclei} \in X} V_{\text{atomic}}(B_{r_{\text{cut}}}(\text{nucleus})).$$

This assumes that the energy of the whole system can be broken up into an energy of subsystems which are closed under interaction with respect to a maximal interaction radius  $r_{\text{cut}}$ . Additionally, we decomposed the local atomic environments dimension wise into parts containing only one nucleus, pairs of nuclei, triples of nuclei and so on,

$$V_{\text{atomic}}(B_{r_{\text{cut}}}(\text{nucleus})) = \sum_{\text{pairs}} w_2(\text{pair}) + \sum_{\text{triples}} w_3(\text{triple}) + \dots$$

We stated physical meaningful assumptions on the potential function, such as rotational invariance, permutation invariance and a maximal interaction

radius with respect to  $r_{\text{cut}}$ . We briefly mentioned the usage of a suitable descriptor of an atomic environment which can be used to filter out redundant information and reduce the dimension of the problem.



## Chapter 6

# Approximating the PES

Unfortunately, it is not generally possible to solve the Schrödinger equation and consequently the PES analytically; thus we rely on numerical approximations and simulations. In the following chapter, we will make the approximation of the Born-Oppenheimer PES a subject of discussion using the least squares regression. All following considerations will be based on the many-body expansion. Likewise, we will need to capture a dimension-wise decomposition in the search set. Aside from the local energy contributions, the molecular dynamics applications also need to compute the forces acting on each particle. Thus, we will also incorporate the negative gradient of the potential with respect to the coordinates of the particles. *Many Body Expansion*

For a configuration  $\mathbf{X} = (X_1, \dots, X_M)$  containing  $M$  atoms, the many body decomposition takes the form, *Simplification*

$$V(\mathbf{X}) = \sum_{\substack{u \subseteq \{1, \dots, M\} \\ 0 < |u|}} w_u(\mathbf{X}_u).$$

As we mentioned earlier many of those functions  $w_u$  will coincide with respect to a group structure dependent on the specific physical situation. Nevertheless, for the sake of notational simplicity of this chapter, we will assume all functions  $w_u$  and  $w_{u'}$  to be equal if and only if  $|u| = |u'|$ . Thus, we will assume the simplified form

$$\begin{aligned} V(\mathbf{X}) &= \sum_{\substack{u \subseteq \{1, \dots, M\} \\ 0 < |u|}} w_{|u|}(\mathbf{X}_u) \\ &=: \langle w_1, \dots, w_M \rangle(\mathbf{X}). \end{aligned} \tag{6.0.1}$$

This assumption is solely meant to simplify the following notation and is in general inaccurate, since it would assign two configuration fragments the same energy, even though they are not physical equivalent. For example in our implementation we have two different functions  $w_4^1$  and  $w_4^2$ . One is evaluated on 4-bodies that form a star-like constellation where one middle particle has three neighbours. The other one is used for 4-bodies that are more like a snake where the start and end of the snake only have one neighbour and the two

other particles both have two neighbours. Particles are considered neighbours when their distance is smaller than  $r_{\text{cut}} > 0$ . This is necessary as depending on which kind of constellation the particles build we have to allow for different types of symmetry: For a star like constellation we can rotate the three outer particles around the middle one and still have a physically equivalent constellation. For the snake-like constellation there is only one symmetry which is reversing the order of the snake and start counting from the other side. With more than 4 bodies the number of different constellations and especially the number of symmetries allowed per constellation will of course grow rapidly. Nevertheless, the approximation theory presented below can be easily transferred to the general case which we will do in the following chapters.

## 6.1 Constructing a Regression MLIP $\hat{V}_{D_n, V_p}$

### *The Problem*

In the following section, we'll introduce a machine learning potential based on the simplified many-body expansion. Thus, we are searching for an approximation of the form

$$\hat{V}_K = \langle \hat{w}_1, \dots, \hat{w}_K \rangle \approx \langle w_1, \dots, w_M \rangle = V, \quad (6.1.1)$$

for some  $K \leq M$ . Here,  $\hat{V}_K$  forms a  $K$ -body potential.

### *Structure*

This section is structured as follows. First, in subsection 6.1.1, we choose a suitable descriptor to in advance reduce the dimension of the problem and filter out access information, i.e. we decide how a chemical construction of a  $k$ -body enters the function  $w_k$ . Due to the dimension wise decomposition of the potential function, we will in subsection 6.1.2 also introduce a search space which decomposes dimension wise. Thus we will define a search space  $W_{k, \leq p}$  associated to the fractional potential  $w_k$  of order  $k$  and an approximation accuracy  $p$ . In subsection 6.1.3 we introduce a least squares regression problem on the prescribed search space. With that, the least squares regression will take the form:

$$\hat{c} := \underset{c}{\operatorname{argmin}} \left\| \begin{pmatrix} A_1 & \dots & A_M \end{pmatrix} \cdot c - y \right\|_2. \quad (6.1.2)$$

Here, the sub matrices  $A_1$  will correspond to the one-nuclei part, the sub matrix  $A_2$  to the pair-nuclei part and so on. From now on, we will denote a multi index or multidimensional elements by bold letters.

### 6.1.1 Choosing a Descriptor

#### *Pairwise Distances*

The fractional energy function  $w_k$  requires sufficient information of a set of  $k$  particles in order to describe its energy. This information is based on its spatial coordinates and atomic numbers, thus on  $k \cdot (3 + 1)$  parameters. The reduction of this dimension is a matter of choosing a suitable descriptor, i.e. a map taking

those information and going to a lower dimensional space, without losing too much information. We will describe a  $k$ -body by all pairwise distances

$$d_{ij} := \|R_i - R_j\|_2$$

between its  $i$ th and  $j$ th nuclei. We do not lose information about the positioning since the potential is required to be invariant under rotation and translation of the positions anyway. In case all nuclei in the data also have the same atomic number we also do not lose any information at all by using the above descriptor. An association of  $k$  nuclei is thus uniquely defined by  $n_k$  distances, where

$$n_k := \begin{cases} 1 & , \text{if } k = 1 \\ \sum_{i=1}^{k-1} i = \frac{k \cdot (k-1)}{2} & , \text{otherwise.} \end{cases} \quad (6.1.3)$$

This forms a unique description under the assumption of identical atom numbers. The descriptor of a  $k$ -body using pairwise distances is given by

$$D : (\mathbb{R}^3, \mathbb{N})^k \longrightarrow \mathbb{R}^{n_k}, \quad \mathbf{X}_u \longmapsto (d_{i,j})_{i,j \in u} \quad (6.1.4)$$

for a  $k$ -body  $\mathbf{X}_u$  described by the index set  $u \subseteq \{1, \dots, M\}$  with  $|u| = k$ . A further common way to describe a  $k$ -body uses a mixture of pairwise distances and adjacent angles. Nevertheless, both descriptions are equivalent and can be easily transformed into each other. If the atom numbers of the nuclei differ, one may take a Coulomb potential, a scaled distance, instead of just the pairwise distances.

### 6.1.2 Choosing the Search Set

In the previous subsection we've chosen a descriptor. Now we are interested in finding a suitable feature map. With that, one maps the description of an atomic configuration into a high dimensional feature space and searches for the best linear representation of the target function there. This makes the optimization problem we solve at the end a linear one but gives as a solution that is non-linear as a map from distances to energy. Taking the previously defined descriptor into account, the problem of interest of (6.1.1) transforms to

$$\hat{V}_K = \langle \tilde{w}_1 \circ D, \dots, \tilde{w}_K \circ D \rangle \approx \langle w_1, \dots, w_M \rangle = V, \quad (6.1.5)$$

where  $\hat{w}_k = \tilde{w}_k \circ D$ . Once fixed a descriptor, we aim to find the optimal functions  $\tilde{w}_k$ . Using the pairwise distance descriptor introduced in (6.1.4) the potential function takes the form

$$\hat{V}_K(\mathbf{X}) = \sum_{i_1}^M \tilde{w}_1(1) + \sum_{i_1 < i_2}^M \tilde{w}_2(d_{i_1 i_2}) + \dots + \sum_{i_1 < \dots < i_K}^M \tilde{w}_K((d_{m,n})_{m,n \in \{i_1, \dots, i_K\}}). \quad (6.1.6)$$

Note that the function  $\tilde{w}_k$  is  $n_k$ -dimensional. Due to the dimension wise decomposition of the potential function, we will also need the construction of a

*Many-Body  
Expansion*

search space which decomposes dimension wise.

Lets fix an arbitrary approximation accuracy  $\mathbf{p} = (p_1, \dots, p_K) \in \mathbb{N}_0^K$ . Here, we think of  $p_k$  as the approximation accuracy of  $\tilde{w}_k$ . In the next section, we will introduce an adaptive procedure, how to chose this notion in an optimal way. But for now, we assume the choice to be arbitrary. The form of  $\mathbf{p}$  suggests that just like the function itself, the search space also decomposes dimension wise. In fact, the search space will take the form

*Dimension-wise decomposition of the Search Set*

$$V_{\mathbf{p}} = \bigoplus_{q=1}^{p_1} W_{1,q} \oplus \dots \oplus \bigoplus_{q=1}^{p_K} W_{K,q}, \quad (6.1.7)$$

With that, we are searching the optimal estimate  $\tilde{w}_k \in \bigoplus_{q=1}^{p_k} W_{k,q}$  for each  $k = 1, \dots, K$ . To this end, we will modify orthogonal polynomials such that they meet the required physical assumptions of section 5.3. This way, we encode all assumptions we would like the potential function to satisfy already in the function space and do not have to worry about the suitability of the estimate later on. The  $\bigoplus_{q=1}^{p_k} W_{k,q}$  consist of non-linear polynomials of the descriptor but by searching the best linear combination of these our optimization problem will become linear in the end.

*Orthogonal Polynomials*

The foundation of the basis functions is a sequence of orthogonal polynomials. We start with a definition of one-dimensional orthogonal polynomials and construct a multi-dimensional generalization later on.

#### Definition 6.1: Sequence of $\mu$ -orthogonal Polynomials

Let  $\mu$  be a Borel measure on  $I \subseteq \mathbb{R}$  and denote by  $L^2(X, d\mu)$  the associated Hilbert space of all square integrable functions with respect to  $\mu$  with the inner product  $\langle f, g \rangle = \int_I f(x)g(x)d\mu(x)$ . Then the sequence  $(P_n)_{n \in \mathbb{N}_0}$  is called *sequence of  $\mu$ -orthogonal polynomials*, if for all  $m, n \in \mathbb{N}_0$

$$\deg P_n = n \quad (6.1.8)$$

and

$$\langle P_n, P_m \rangle_{\mu} = 0 \text{ for } m \neq n. \quad (6.1.9)$$

If  $\langle P_n, P_n \rangle_{\mu} = 1$ , the sequence is called  *$\mu$ -orthonormal*.

In more detail, we will consider Legendre, Chebyshev and Laguerre polynomials

**Example 6.1.** 1) **Legendre Polynomials:** Let be  $X = [-1, 1]$  and the Borel measure  $\mu_w$  characterized by the easiest weight function

$$w \equiv 1.$$

The so corresponding orthogonal polynomials  $(P_n)_{n \in \mathbb{N}_0}$  are orthogonal to the standard inner product of square integrable functions. Legendre



polynomials can be constructed with a Three-Term recurrence,

$$(n+1)P_{n+1}(x) = (2n+1)xP_n(x) - nP_{n-1}(x)$$

and starting values  $P_0 = 1, P_1 = x$ .

- 2) **Chebyshev Polynomials:** Let again be  $X = [-1, 1]$  and  $\mu_w$  the Borel measure characterized by

$$w(x) = \frac{1}{\sqrt{(1-x)^2}}$$

Then the  $\mu_w$ -orthogonal Chebyshev polynomials  $(T_n)_{n \in \mathbb{N}_0}$  can be constructed by the Three-Term recursion

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x)$$

and starting values  $T_0(x) = 1, T_1(x) = x$ .

- 3) **Laguerre Polynomials:** For  $X = [0, \infty)$  the polynomials  $(L_n)_{n \in \mathbb{N}_0}$  are orthogonal with respect to the Borel measure  $\mu_w$  constructed by the weight function

$$w(x) = e^{-x}.$$

The Three-Term recursion reads

$$(n+1)L_{n+1}(x) = ((2n+1) - x)L_n(x) - nL_{n-1}(x)$$

with starting values  $L_0(x) = 1, L_1(x) = -x + 1$ .

To construct  $n_k$ -dimensional orthogonal polynomials, we use a tensor product of the one dimensional functions,

$$P_{\mathbf{q}}(\mathbf{x}) := \prod_{i=1}^{n_k} P_{q_i}(x_i),$$

for  $\mathbf{q} = (q_1, \dots, q_{n_k}) \in \mathbb{N}_0^{n_k}$  and  $\mathbf{x} = (x_1, \dots, x_{n_k}) \in \mathbb{R}^{n_k}$ . The so constructed multi dimensional polynomials are again orthogonal with respect to  $\mu^{n_k} = \mu \otimes \dots \otimes \mu$ . Moreover, if the one dimensional sequence of polynomials is orthonormal, the multi dimensional polynomials are again orthonormal. The degree of this multi dimensional polynomial is given by

$$\deg(P_{\mathbf{q}}) = |(q_1, \dots, q_{n_k})|_1 = \sum_{i=1}^{n_k} q_i.$$

Those polynomials span the space of square integrable functions over  $\mathbb{R}^{n_k}$ , which lies dense in the space of all continuous functions. But in fact we do not want to take all those functions into account, but only those who meet the physical assumptions of a potential function.

A general sequence of orthogonal polynomials a priori does not meet the con-

*Modifying the  
Orthogonal  
Polynomials*

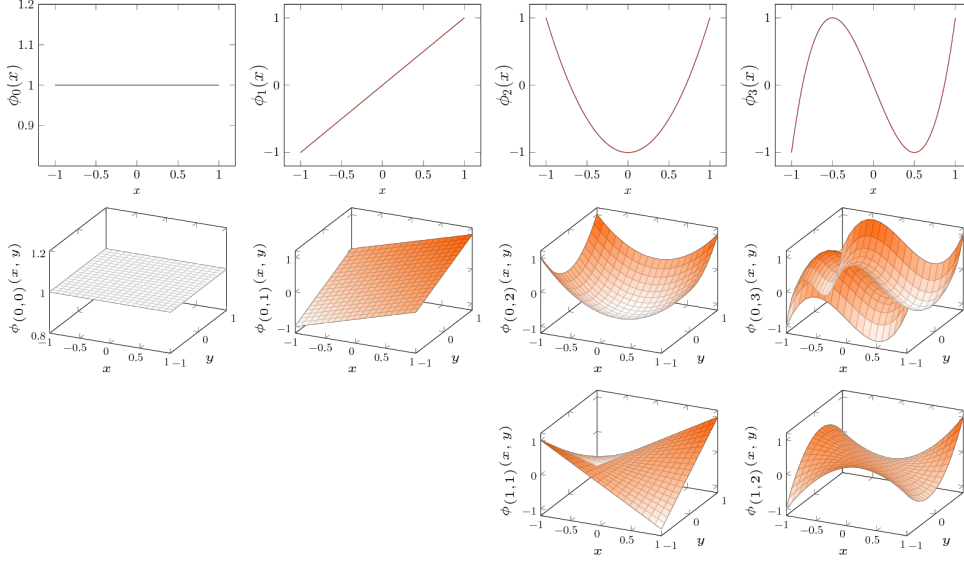


Figure 6.1: One- and two-dimensional Chebyshev polynomials constructed by a tensor product up to degree three.

sidered requirements. As already noted we will ignore particles that have a distance greater than  $r_{\text{cut}}$ . Equivalently, the search spaces should only consist of functions with domain in  $r_{\text{cut}}$ . Since we also want the potential to be continuous it has to go to 0 smoothly when approaching  $r_{\text{cut}}$ . Thus we require this property from our basis of  $w_k$ :

$$\lim_{|\mathbf{x}|_{\infty} \rightarrow +r_{\text{cut}}} w_k(\mathbf{x}) = 0. \quad (6.1.10)$$

In practice, one forces the functions to have this property by multiplying a suitable *cut off function* that approaches 0 smoothly. In our implementation we use for  $x \in \mathbb{R}$

$$f_{\text{cut}}(x) = \begin{cases} 1 & \text{for } r \in (0, r_{\text{cut}} - d] \\ \frac{1}{2} \cos\left(\pi \frac{x - r_{\text{cut}} + d}{d} + 1\right) & \text{for } r \in (r_{\text{cut}} - d, r_{\text{cut}}], \\ 0 & \text{for } r \in [r_{\text{cut}}, \infty) \end{cases}, \quad (6.1.11)$$

where  $d \in [0, r_{\text{cut}}]$  determines the width of the cut off region. Using that, we are forcing the multi-dimensional polynomial to have the property (6.1.10) by

$$\prod_{i=1}^{n_k} P_{q_i}(x_i) \cdot f_{\text{cut}}(x_i).$$

Another restriction we have on our potential is reflection and rotation invariance. As already noted earlier, the allowed reflections and rotations that map a configuration to an equivalent one form a group  $S$ . A simple way to make a function  $\phi$  invariant under the group action of a group  $S$  is just to take

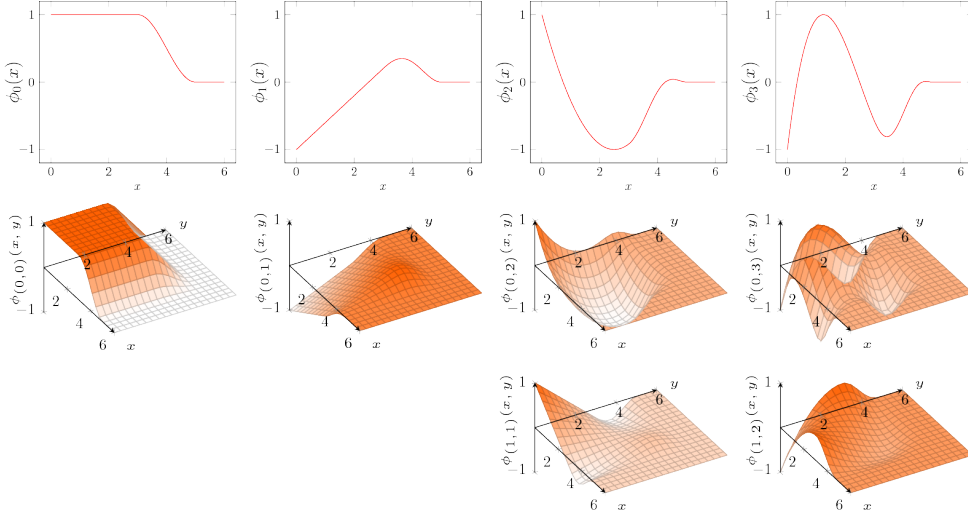


Figure 6.2: One- and two-dimensional basis functions based on Chebyshev polynomials up to degree three. The Chebyshev polynomials are multiplied with the cut-off function (6.1.11) with  $r_{\text{cut}} = 5$  and  $r_l = 2$ .

$\tilde{\phi} = \sum_{g \in S} \phi(g(x))$ . It is easy to see that due to the group property if  $y = g(x)$  for some  $g \in S$ , then  $\tilde{\phi}(y) = \tilde{\phi}(x)$ . In our case all allowed symmetries correspond exactly to permutations of the inputs of the function. So we can treat  $S$  as a subgroup of the symmetric group and get:

$$\phi_{\mathbf{q}}(\mathbf{x}) := \frac{1}{|S|} \sum_{\sigma \in S} \prod_{i=1}^{n_k} P_{q_i}(x_{\sigma(i)}) \cdot f_{\text{cut}}(x_{\sigma(i)}) = \frac{1}{|S|} \sum_{\sigma \in S} \prod_{i=1}^{n_k} P_{q_{\sigma(i)}}(x_i) \cdot f_{\text{cut}}(x_i) \quad (6.1.12)$$

for all  $\mathbf{q} \in \mathbb{N}_0^{n_k}$ . Note that we did not further specify the specific form of  $S$ . In subsection 7.1 we will analyze its structure in detail.

We define

$$\mathcal{G}_{k,p} := \{\mathbf{q} \in \mathbb{N}_0^{n_k} \text{ such that } |\mathbf{q}|_1 = p\}.$$

*Search  
Space*

which are the indexes of all basis functions approximating  $k$ -bodies s.t. their degrees add up to  $p$ . Note that this Using this we can define our search spaces with precision  $p$ :

$$W_{k,p} := \text{span}\{\phi_{\mathbf{q}} : \mathbf{q} \in \mathcal{G}_{k,p}\}, \quad (6.1.13)$$

and we write down the basis into a vector:

$$\Phi_{k,p} := (\phi_{\mathbf{q}})_{\mathbf{q} \in \mathcal{G}_{k,p}}. \quad (6.1.14)$$

Additionally define,

$$W_{k,\leq p} := \bigoplus_{q=1}^p W_{k,q}.$$

and

$$\mathcal{G}_{k,\leq p} := \{\mathbf{q} \in \mathbb{N}_0^{n_k} \text{ such that } |\mathbf{q}|_1 \leq p\}.$$

Here,  $W_{k,\leq p}$  incorporates all  $n_k$ -dimensional basis function as constructed in (6.1.12), based on orthogonal polynomials up to degree  $p$  and serves as a search space of the fractional function  $\tilde{w}_k$ . Note that this index set corresponds to the a classical sparse grid construction. For a multi index  $\mathbf{p} = (p_1, \dots, p_K) \in \mathbb{N}_0^K$ , we define the general search space by

$$\begin{aligned} V_{\mathbf{p}} &:= W_{1,\leq p_1} \oplus \dots \oplus W_{K,\leq p_K} \\ &= \bigoplus_{k=1}^K W_{k,\leq p_k}. \end{aligned} \quad (6.1.15)$$

i.e. we are back to (6.1.7) with the desired format. One should note that by the choice of the finite dimensional search space we also already are doing some regularization. For example due to all  $W_{k,q}$  consisting only of  $C^\infty$  functions our approximation will also be a  $C^\infty$  function.

In the following section, we will define the optimal approximation with the help of a given search set as a least squares regression solution. For a given approximation accuracy  $\mathbf{p} = (p_1, \dots, p_K)$  and training data we choose the optimal approximation  $\hat{V}_K = \langle \hat{w}_1, \dots, \hat{w}_K \rangle \in V_{\mathbf{p}}$  of the regression function  $V = \langle w_1, \dots, w_M \rangle$ .

### 6.1.3 K-Body Regression Potential

*MLIP of  
Order K*

In the following, we construct a MLIP based on nuclei interactions up to order  $K$ , i.e. up to  $K$  nuclei at once. Based on the training set  $\mathcal{D}_n$ , we aim to choose a suitable estimation  $\langle \hat{w}_1, \dots, \hat{w}_K \rangle$  such that

$$\hat{V}_{\mathcal{D}_n, V_{\mathbf{p}}} = \langle \hat{w}_{1,p_1}, \dots, \hat{w}_{K,p_K} \rangle \approx V, \quad (6.1.16)$$

where  $\hat{w} = \tilde{w} \circ D$ .

*The Data Set*

Let be given a database of atomic configurations together with their reference energies and forces,

$$\mathcal{D}_n = \{(\mathbf{X}_i, \mathbf{f}_i, e_i) : i = 1, \dots, n\}.$$

One data point represents the information of one whole atomic system containing  $M_i$  nuclei,  $\mathbf{X}_i = (X_1, \dots, X_{M_i})$ . The forces and energies are computed with a costly but relatively accurate quantum mechanical ESM method,

$$\begin{aligned} V^{(ESM)}(\mathbf{X}_i) &= e_i, \\ -\nabla V^{(ESM)}(\mathbf{X}_i) &= \mathbf{f}_i. \end{aligned} \quad (6.1.17)$$

Here,  $e_i$  denotes the ground energy of the system  $\mathbf{X}_i$  and  $\mathbf{f}_i \in \mathbb{R}^{M_i \times 3}$  its forces. By incorporating the forces of a system in the learning process, we gain a wealth of additional information, because we do not only have information about the function evaluation but also about the gradient. Nevertheless, to

use this further knowledge, we need to calculate the derivatives of the very high dimensional estimation.

Having defined the search space, we are able to formulate the empirical regression problem. To do so, we will first focus on the energy based information of the training data and incorporate the forces later on. The energy-based empirical error functional with respect to the data set  $\mathcal{D}_n$  is given by *Energy-Based  
Empirical  
Regression  
Problem*

$$\mathcal{E}_{\mathcal{D}_n}^E(g) := \frac{1}{n} \sum_{i=1}^n \|g(\mathbf{X}_i) - e_i\|_2^2. \quad (6.1.18)$$

Using that, the energy-based regression problem on the finite dimensional function space  $V_p$  is then given by

$$\begin{aligned} \hat{V}_{\mathcal{D}_n, V_p}^E &= \langle \hat{w}_{1, p_1}, \dots, \hat{w}_{K, p_K} \rangle \\ &:= \operatorname{argmin}_{\langle g_1, \dots, g_K \rangle \in V_p} \mathcal{E}_{\mathcal{D}_n}^E(\langle g_1, \dots, g_K \rangle) \end{aligned} \quad (6.1.19)$$

As in the previous section, this problem can be equivalently expressed as an optimization problem of the coefficients in the search space,

$$\hat{c}_{\mathcal{D}_n, V_p}^E := \operatorname{argmin}_c \frac{1}{n} \|n A_K^E \cdot c - e\|_2^2, \quad (6.1.20)$$

where  $e := (e_1, \dots, e_n)$  denotes the vector with energies of the atomistic configurations in the training set  $\mathcal{D}_n$ . Since the search space has an extraordinary form in this application, so has the matrix  $A_K^E$ . Therefore, we will go into more detail about the construction of this matrix.

To understand the form of the matrix  $A_K^E$  lets have a closer look at the functions contained in the search space  $V_p$ . For one configuration  $\mathbf{X} = (X_1, \dots, X_M)$  containing  $M$  nuclei, we recall the definition of the many-body expansion of a  $K$ -body potential, *Construction of  
the Matrix  $A_K^E$*

$$\begin{aligned} \langle g_1, \dots, g_K \rangle(\mathbf{X}) &:= \sum_{|u|=1} g_1(\mathbf{X}_u) + \dots + \sum_{|u|=K} g_K(\mathbf{X}_u) \\ &= \sum_{k=1}^K \sum_{|u|=k} g_k(\mathbf{X}_u). \end{aligned} \quad (6.1.21)$$

Here, the sums run over all possible index sets  $u \subseteq \{1, \dots, M\}$  and as before, we denote by  $\mathbf{X}_u := \{X_i\}_{i \in u}$  the subset of nuclei with an index contained in the index set  $u$ . First, we are summing over all single nuclei in the atomic system, then over all pairs of nuclei and so on. Since  $g_k \in W_{k, \leq p_k}$  by construction of the search space, we can write this function as linear combination of the corresponding basis functions, i.e.

$$g_k = c_{k, \leq p_k} \cdot \mathbf{E}_{k, \leq p_k}(\mathbf{X})$$

where

$$\mathbf{E}_{k,\leq p_k}(\mathbf{X}) := \left( \sum_{\substack{|u|=k \\ \mathbf{1} \in \mathcal{G}_{k,\leq p_k}}} \phi_{\mathbf{1}}(\mathbf{X}_u) \right) \quad (6.1.22)$$

is the vector of basis functions of  $W_{k,\leq p_k}$  summed over all subsets of size  $k$ . The  $j$ th entry of the vector  $\mathbf{E}_{k,p_k}(\mathbf{X})$  simply denotes a summation of the  $j$ th basis function of  $W_{k,\leq p_k}$  evaluated at all possible  $k$ -bodies. Now putting it all together

$$\langle g_1, \dots, g_K \rangle(\mathbf{X}) = \sum_{k=1}^K \mathbf{c}_{k,\leq p_k} \cdot \mathbf{E}_{k,\leq p_k}(\mathbf{X}),$$

We titled this array with an large  $\mathbf{E}$  to remind us, that we until here only incorporated the energy information. Define the matrix

$$A_K^E := \frac{1}{n} \begin{pmatrix} \mathbf{E}_{1,p_1}(\mathbf{X}_1) & \dots & \mathbf{E}_{K,p_K}(\mathbf{X}_1) \\ \vdots & \ddots & \vdots \\ \mathbf{E}_{1,p_1}(\mathbf{X}_n) & \dots & \mathbf{E}_{K,p_K}(\mathbf{X}_n) \end{pmatrix}. \quad (6.1.23)$$

Here, each  $\mathbf{E}_{k,p_k}$  stands for all basis functions used to approximate  $\hat{w}_k$  with prescribed accuracy  $p_k$ . In the previous section, the matrix  $A$  took the form  $A_{i,j} = \phi_i(x_j)/n$  for basis functions  $\phi_i$  of the search space and data points  $x_j$ . Here, instead of describing one single basis function,  $\mathbf{E}_{k,p_k}$  describes all the basis functions needed to approximate  $w_k$ . This unusual form is due to the dimension wise decomposition of the search space in (6.1.7).

### *Incorporating The Forces*

The energy of the system is not the only information we have, we also assume the knowledge about the incorporated forces. To this end, we need the introduction of the derivatives of the basis functions. Therefore, we have a look at the force-based empirical error functional based on the data set  $\mathcal{D}_n$ , which is given by

$$\mathcal{E}_{\mathcal{D}_n}^F(g) := \frac{1}{n} \sum_{i=1}^n \|\nabla g(\mathbf{X}_i) - \mathbf{f}_i\|_2^2. \quad (6.1.24)$$

Based on that, the force-based empirical regression problem using the search space  $V_{\mathbf{p}}$  can be written as

$$\hat{V}_{\mathcal{D}_n, V_{\mathbf{p}}}^F := \operatorname{argmin}_{\langle g_1, \dots, g_K \rangle \in V_{\mathbf{p}}} \mathcal{E}_{\mathcal{D}_n}^F(\langle g_1, \dots, g_K \rangle). \quad (6.1.25)$$

Again, we can write equivalently

$$\hat{c}_{\mathcal{D}_n, V_{\mathbf{p}}}^F := \operatorname{argmin}_c \frac{1}{n} \|\nabla n A_K^E \cdot c - f\|_2^2, \quad (6.1.26)$$

for the matrix  $A_K^E$  as defined in (6.1.23) and  $f := (\mathbf{f}_1, \dots, \mathbf{f}_n)$ . Moreover, an empirical regression problem using the whole dataset at once can be obtained by defining the empirical error functional

$$\mathcal{E}_{\mathcal{D}_n} := \mathcal{E}_{\mathcal{D}_n}^E + \mathcal{E}_{\mathcal{D}_n}^F. \quad (6.1.27)$$

The empirical regression problem becomes

$$\hat{c}_{\mathcal{D}_n, V_{\mathbf{p}}} := \operatorname{argmin}_c \frac{1}{n} \|A_K \cdot c - y\|_2^2, \quad (6.1.28)$$

where  $A_K := (A_K^E | -\nabla A_K^E)$  and  $y := (e_1, \dots, e_n, \mathbf{f}_1, \dots, \mathbf{f}_n)$ .

One has to keep in mind that every of these  $\mathbf{E}$  consists of a rather big sum of evaluations of  $\phi$  on each subset of a given cardinality of  $\mathbf{X}$ . Finding the subsets  $\{X_i\}_{i \in U}$  such that the neighbourhood relations are all smaller than  $r_{\text{cut}}$  is a task that grows fast with the number of  $k$ -bodies one wants to treat. The  $\phi$  themselves then again are sums over all allowed symmetries for a given  $k$ -body. The number of these symmetries also grows. These are the reasons that the cost to assemble the matrix  $A_K$  grow fast with the number of  $k$ -bodies one wants to treat.

Lets recall what we did in this section. We assumed the potential energy surface (PES) introduced in chapter 5 to fulfill the atomic decomposition ansatz and the many body expansion ansatz. Assuming this format, we defined a descriptor, which describes each set of  $k$  nuclei by its pairwise distances. To formulate a regression problem, we defined a search space, which just like the function we want to approximate itself decomposes dimension wise. Therefore, we fixed a maximal order of bodies  $K$  and an approximation accuracy vector  $\mathbf{p} = (p_1, \dots, p_K)$  encoding with how much effort we wan each body to treat. Then we formulated the least squares regression problem. We explained the explicit form of the matrix involved, which again decomposes dimension wise. Also, we pointed out, that we do not only have data points in the training set regarding function evaluations due to the energy, but also over the derivation due to the forces. *Review*

In the following subsection, we will specify how one can optimally choose the maximal involved body order  $K$  and the approximation accuracy  $\mathbf{p} = (p_1, \dots, p_K)$ , which we until here assumed to be arbitrary. Therefore, we will use an application of a sparse tensor product construction. Since we can't specify the smoothness characteristic of the PES, we rely on an adaptive choice of the index set. Nevertheless, we will see in the results that a decay of the contributions can be recorded. *Outlook*

## 6.2 Choosing an Optimal Search Set $V_{\text{opt}}$

In the previous section, we constructed a machine learning interatomic potential given a maximal body order  $K$  and accuracy  $\mathbf{p} = (p_1, \dots, p_K)$  by *The Problem*

$$\hat{V}_{\mathcal{D}_n, V_{\mathbf{p}}} := \operatorname{argmin}_{\langle g_1, \dots, g_K \rangle \in V_{\mathbf{p}}} \mathcal{E}_{\mathcal{D}_n}(\langle g_1, \dots, g_K \rangle) \approx V.$$

In this section, we present an adaptive method to choose the search set in an efficient way under a prescribed workload.

Therefore, define the generalized search space based on an index set  $\mathcal{I} \subseteq \{1, \dots, M\} \otimes \mathbb{N}_0^M$  by

$$V_{\mathcal{I}} := \bigoplus_{(k,q) \in \mathcal{I}} W_{k,q}.$$

Starting at the smallest index set possible, we will gradually include indices which maximizes a benefit-cost-ratio. We start with the coarsest approximation possible, which is approximating the PES only with one-nuclei parts ( $K = 1$ ) using constants ( $p_1 = 0$ ), i.e.  $\mathcal{I} = \{(1,0)\}$ . We strive the index set to be admissible. The starting index set  $\mathcal{I} = \{(1,0)\}$  fulfills this condition and we will only add indices that preserve this property. Therefore, we define the set of indices which are allowed to be added next to an index set  $\mathcal{I}$  by

$$\mathcal{N}(\mathcal{I}) := \{(k,q) \in \{1, \dots, M\} \otimes \mathbb{N}_0^M : \mathcal{I} \cup (k,q) \text{ is admissible}\}.$$

*Cost-function*

To decide which index to incorporate next, we need to measure its worth. To this end, we define the cost of an index  $(k,q) \in \{1, \dots, M\} \otimes \mathbb{N}_0^M$  by

$$c(k,q) := |W_{k,q}|.$$

Thus, we define the cost of an index set as the incorporated degrees of freedoms involved, when adding it to the index set. Since the cost-function depends on one single index, we name it with a small  $c$  to mark its locality. We define the benefit-function by

*Benefit-function*

$$B(\mathcal{I}, (k,q)) := \mathcal{E}_{\mathcal{D}_n}(\hat{V}_{\mathcal{D}_n, V_{\mathcal{I}}}) - \mathcal{E}_{\mathcal{D}_n}(\hat{V}_{\mathcal{D}_n, V_{\mathcal{I} \cup (k,q)}})$$

for an index set  $\mathcal{I} \subset \{1, \dots, M\} \otimes \mathbb{N}_0^M$  and an index  $(k,q)$ . The benefit thus is not solely dependent on one index, but describes its worth given an already chosen index set. We marked its non-local character by a capital  $B$ . Thus, the benefit of an index given a prescribed index set is the change of the empirical error functional based on the used least squares regression. Note, that this definition does make the problem a non-linear one. This means, the refinement in one dimension is not independent of the refinements in other dimensions of the index set. This makes the choice of the next index set non-unique, if there is more than one index set with exactly the same benefit-cost ratio the algorithm would take just one arbitrary, which may lead to a different behavior afterwards. Nevertheless, in application this case is very unlikely.

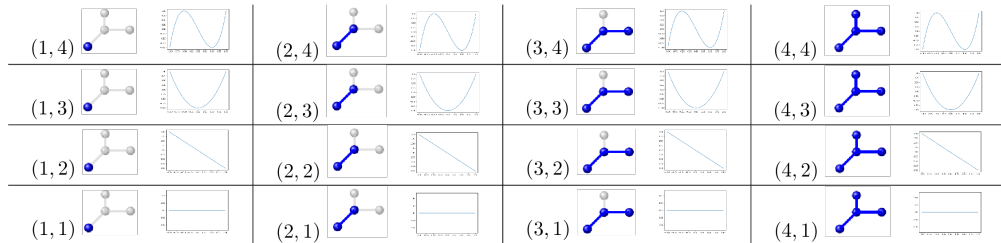


Figure 6.3: Visualization of the used index set. The tuple  $(k,q)$  encodes the approximation of  $k$ -bodies with the help of basis functions based on polynomials of degree  $q$ .



- 1: **Input:**  $C_{\max}$
- 2: **Output:**  $\mathcal{I}_{\text{opt}}$
- 3: **Initialize:**  $\mathcal{I} := \emptyset, cost = 0, \mathcal{N}(\mathcal{I}) = \{(0, 1)\}$
- 4: **while**  $cost < C_{\max}$  **do**
- 5: Compute the values  $\eta(\mathcal{I}, (q, k)) := B(\mathcal{I}, (q, k)) / c(q, k)$  for all  $(q, k) \in \mathcal{N}(\mathcal{I})$  and set  $cost := cost + c(q, k)$
- 6: Select the next index  $(q^*, k^*) := \operatorname{argmax}_{(q, k) \in \mathcal{N}(\mathcal{I})} \eta(\mathcal{I}, (q, k))$  and move it to  $\mathcal{I}$
- 7: Generate new allowed index set  $\mathcal{N}(\mathcal{I})$
- end**

**Algorithm 2:** Adaptive algorithm to construct quasi-optimal index set  $\mathcal{I}_{\text{opt}}$  under a given maximal cost  $C_{\max}$ .



# Chapter 7

## Assessment and Validation

In this chapter, we will present the approximation procedure in detail. In section 7.1, we will explain the group structure under which the basis functions are invariant in more detail. In section 7.2 we will compare the different basis sets. In section 7.3 we make several remarks to the implementation phase and in section 7.4 we present the data sets we evaluate our methods on.

### 7.1 Group Structure for used Physical Applications

We remember, the general formulation of the many body expansion under an application of descriptor  $D$  reads

$$V(\mathbf{X}) = \sum_{\substack{u \subseteq \{1, \dots, M\} \\ 0 < |u|}} w_{|u|} \circ D(\mathbf{X}_u).$$

As we mentioned previously, many of the functions  $w_u$  of the many body expansion can be assumed to coincide if the corresponding sets  $\mathbf{X}_u$  are physically equivalent.

#### 7.1.1 Solids with Periodic Boundary

Lets first consider the case where each atomic number in the data set is equal as in the W-14 data set. Therefore the atomic number loses its describing character and we can filter out this information by a suitable descriptor. Let be  $D$  in this case the pairwise distance descriptor introduced in subsection 6.1.1.  $D$  maps each fraction of an atomic environment  $\mathbf{X}_u = (X_i)_{i \in u}$  onto its pairwise distances. Since we describe each  $k$ -body uniquely by an array containing all pairwise distances, a rotation or translation in space can be expressed by a corresponding permutation of this distance array, see Figure 7.1. This way, the group structure of subsets  $u \subseteq \{1, \dots, M\}$  for  $|u| = k$  equivalently induces a group structure on the symmetric group of permutations  $S_{n_k}$ . Here  $n_k$  is the number of all pairwise distances needed to uniquely describe one  $k$ -body. Nevertheless, not every set of  $k$  atoms induces the same group structure, but the choice also depends on a *grade of connection* incorporated in the body. In

Figure 7.1 , we list all permutational groups for all possible bodies up to order five.

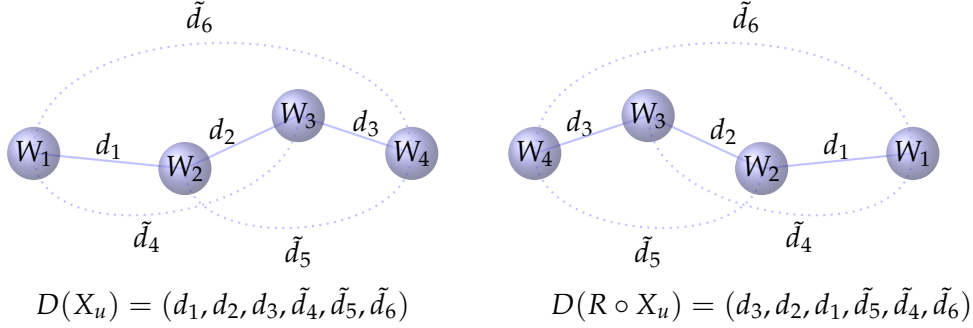


Figure 7.1: The one-to-one correspondence of an physical transformation and the permutation of the description. A 4-body containing solely tungsten (W) atoms is reflected through the middle axis. Equivalently, the pairwise distances are permuted with the permutation  $\sigma = (3, 2, 1, 5, 4, 6)$ . Here,  $d_i$  are describing distances spanning the body  $|d_i| < r_{\text{cut}}$  and  $\tilde{d}_i$  are introduced to additionally describing the body uniquely, they might be larger than  $r_{\text{cut}}$ .

For the case of identical atom numbers, we not only write the potential function as the sum of functions  $w_{|u|}$  which assigns each  $|u|$ -body to its energy, but decompose those functions again to distinguish between different interaction grades. For example, we decompose the function  $w_4$  in the sum of  $w_{41}$  and  $w_{42}$ . Here, we used the notation of 7.1 describing with 41 a 4-body which has a snake-like form and with 42 a star-like 4-body. For example, the potential function incorporating up to bodies of order five can be written by

$$V_5 = \langle w_1, w_2, w_3, w_{41}, w_{42}, w_{51}, w_{52}, w_{53} \rangle.$$

This decomposition is also reflected in the search space which is used to approximate  $V_5$  with the help of a penalized least squares regression. Given a prescribed approximation accuracy  $p_k \in \mathbb{N}$  for the function  $w_k$ , the search space used to approximate  $w_k$  reads,

$$W_{k, \leq p_k} = \text{span}\{\phi_1 : \mathbf{1} \in \mathcal{G}_{k, \leq p_k}\}$$

for basis functions which are invariant under the group action defined by  $S_k$  (see 7.1, 'permutation invariance'), i.e.

$$\phi_1(\sigma \circ D(\mathbf{X}_u)) = \phi_1(D(\mathbf{X}_u)), \quad \text{for all } \sigma \in S_k.$$

This is achieved by simply summing over all permutations and normalizing,

$$\phi_1 := \frac{1}{|S_k|} \sum_{\sigma \in S_k} \phi_1 \circ \sigma.$$

Note, that we already defined the basis functions in exactly this way. Nevertheless, we did not mention how the group action  $S_k$  looks like but only made reference that it should have a suitable form to reflect the physical assumptions.


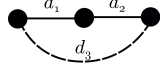
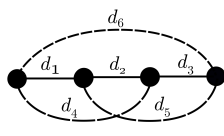
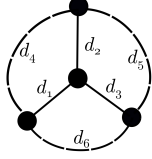
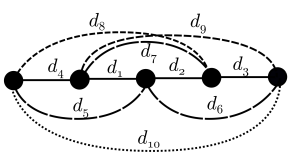
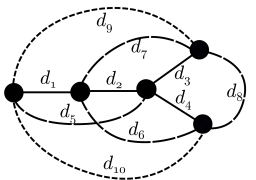
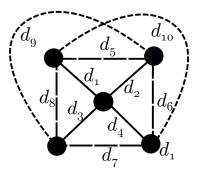
Body	Notation	Format	Permutation Invariance
2 Body	2		$S_2 = (1)$
3 Body	3		$S_3 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & 1 & 3 \end{pmatrix}$
4 Body	41		$S_{41} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 3 & 2 & 1 & 5 & 4 & 6 \end{pmatrix}$
	42		$S_{42} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 \\ 1 & 3 & 2 & 5 & 4 & 6 \\ 2 & 1 & 3 & 4 & 6 & 5 \\ 2 & 3 & 1 & 6 & 4 & 5 \\ 3 & 1 & 2 & 5 & 6 & 4 \\ 3 & 2 & 1 & 6 & 5 & 4 \end{pmatrix}$
5 Body	51		$S_{51} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 2 & 1 & 4 & 3 & 6 & 5 & 7 & 9 & 8 & 10 \end{pmatrix}$
	52		$S_{52} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 1 & 2 & 4 & 3 & 5 & 7 & 6 & 8 & 10 & 9 \end{pmatrix}$
	53		$S_{53} = \begin{pmatrix} 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 & 9 & 10 \\ 1 & 4 & 3 & 2 & 7 & 6 & 5 & 10 & 9 & 8 \\ 2 & 1 & 4 & 3 & 5 & 9 & 8 & 7 & 6 & 10 \\ 2 & 3 & 4 & 1 & 7 & 9 & 10 & 5 & 6 & 8 \\ 3 & 2 & 1 & 4 & 8 & 6 & 10 & 5 & 9 & 7 \\ 3 & 4 & 1 & 2 & 10 & 6 & 8 & 7 & 9 & 5 \\ 4 & 1 & 2 & 3 & 8 & 9 & 5 & 10 & 6 & 7 \\ 4 & 3 & 2 & 1 & 10 & 9 & 7 & 8 & 6 & 5 \end{pmatrix}$

Table 7.1: All possible bodies of order two to five with a varying grade of connection and invariant permutations of the distances. The permutations  $S_k$  form a group of the corresponding symmetric group  $S_{n_k}$ . The four body with the least interaction grad is marked by 41, the larger the grade gets the larger gets the second number. Solid lines are marking direct neighbors, the more dashed a line is, the more particles are located in between. Moreover, the enumeration of the distances are stored from close neighbors to far ones and are in a way sorted, such that they can be uniquely connected to a certain rotation or reflection in space.

### 7.1.2 Molecules

Different than before, one configuration  $\mathbf{X} = (X_1, \dots, X_M)$  does not only contain nuclei of the same atom number, but varying ones, i.e. there exists at least one pair  $i, j \in \{1, \dots, M\}$  such that  $Z_i \neq Z_j$ . Since the interaction energy between two nuclei highly depends on the atom number of both, simply taking the pairwise distance between all nuclei is not a suitable description anymore and we need to specify a suitable way describing a molecule. Of course, there is a wealth of ways to do so. For example, molecules can be represented as Coulomb matrices [40, 47, 52], scattering transforms [42], bags of bonds [39], smooth overlap of atomic positions [4, 5] or generalized symmetry functions [6, 7]. We present two approaches to include information about different atomic numbers in our previous model. On the one hand, we could adjust the description of an local atomic environment. Other than describing an interaction between environment solely by its incorporated pairwise distances  $d_{i,j} = \|\mathbf{R}_i - \mathbf{R}_j\|_2$ , we need the additional information about the atom numbers,  $(Z_i, Z_j, d_{i,j})$ . Thus, the form of the model in this case would be given by

$$V(\mathbf{X}) = \sum_{i_1=1}^M w_1(Z_{i_1}) + \sum_{i_1 < i_2=1}^M w_2(Z_{i_1}, Z_{i_2}, d_{i_1, i_2}) + \dots \\ + w_M \left( (Z_{i_1}, Z_{i_j}, d_{i_1, i_2})_{i_1, i_2 \in \{1, \dots, M\}} \right). \quad (7.1.1)$$

On the other hand, one can learn an energy potential for each specific interaction separately,

$$V(\mathbf{X}) = \sum_{i_1=1}^M w_{(Z_i)}(1) + \sum_{i_1, i_2=1}^M w_{(Z_{i_1}, Z_{i_2})}(d_{i_1, i_2}) + \dots \\ + w_{(Z_1, \dots, Z_M)} \left( (d_{i_1, i_2})_{i_1, i_2 \in \{1, \dots, M\}} \right). \quad (7.1.2)$$

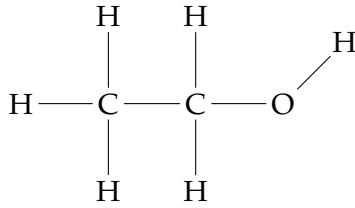
While in the first approach one function is trained describing all bodies of the same order, in the second approach, we aim to describe each type of body separately. To illustrate the second approach, lets have a look at an example.

**Example 7.1** (Ethanol). Lets consider an Ethanol molecule  $C_2H_5OH$ , given by the atom numbers and positions in space of each nuclei,

$$\mathbf{X} = ((Z_C, \mathbf{R}_1), (Z_C, \mathbf{R}_2), \\ (Z_H, \mathbf{R}_3), (Z_H, \mathbf{R}_4), (Z_H, \mathbf{R}_5), (Z_H, \mathbf{R}_6), (Z_H, \mathbf{R}_7), \\ (Z_O, \mathbf{R}_8), (Z_H, \mathbf{R}_9)).$$

In this case, there are six possible pair interactions C-C, C-H, C-O, O-H, O-O and H-H.

Thus, in this case we would train functions  $w_{C,C}, w_{C,H}, w_{C,O}, w_{O,H}, w_{O,O}$  and  $w_{H,H}$  describing each possible pair interaction. We proceed analogously for bodies of higher order.



## 7.2 Comparing the different Basis Functions

During our thesis, we tested four different basis function sets, all based on different orthogonal polynomials. One sequence of orthogonal polynomials we tested but excluded from our observations very early was the set of Hermite polynomials given by

$$H_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2},$$

which are defined on  $\mathbb{R}$  and symmetric around zero. Nevertheless, they seemed to be impractical for our purpose due to their enormous increase of amplitude when increasing the degree. Although, we transformed the polynomials and its degree by multiplying a cut off function, the large absolute function evaluations lead to great round-off errors and badly conditioned regression matrices. A different type of orthogonal polynomials we considered are the Chebyshev polynomials and the Legendre polynomials. Both are special cases of the more general Jacobi polynomials. Those are defined on the interval  $[-1, 1]$ . With a linear transformation, we defined the polynomials on the interval  $[0, r_{\text{cut}}]$  along with a further transformation due a multiplication with the cut-off function. Basis functions based on Chebyshev and Legendre polynomials behave well to a certain degree. Since the polynomials are only defined on the interval  $[0, r_{\text{cut}}]$ , the resulting basis functions have still a degree no more less than the original polynomials. Logically, those basis functions are not suitable for a degree higher than 10, since the basis functions would oscillate too much. One way around would be to define the polynomials not on  $[0, r_{\text{cut}}]$  but on a larger interval, including polynomials with a varying upper interval bound or searching the optimal bound in an expanded regression model. Nevertheless, we only considered Chebyshev and Legendre polynomials on the interval  $[0, r_{\text{cut}}]$  which behave well up to a underlying degree of ten. The fourth set of basis functions are defined on a sequence of Laguerre polynomials on  $[0, \infty]$ . Since Laguerre polynomials are defined on the whole positive real line, the basis function resulting after multiplying with a cut-off function are of much smaller degree than its ancestor polynomial. Additionally, having a look at the resulting basis functions in Figure 7.2, the minima and maxima of the polynomial seems to seamlessly explore the definition area, while the extreme points of the basis functions based on the Chebyshev or Legendre polynomials are more distributed. This makes the set of basis functions based on Laguerre polynomials seem much more stable and easier to use than the other considered basis functions. This assumption will be confirmed in the numerical results.

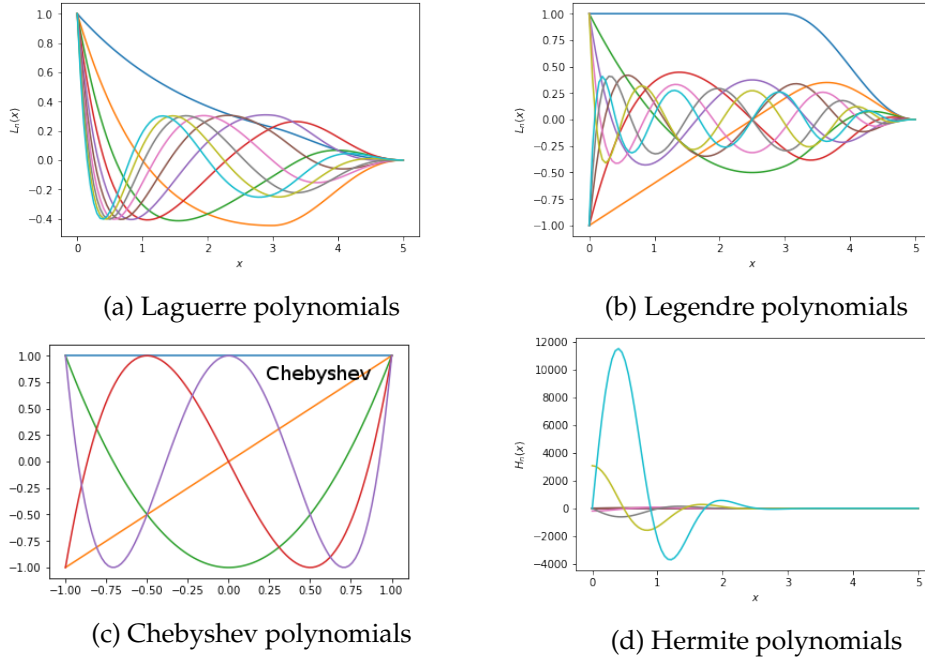


Figure 7.2: The first ten basis functions of basis sets based on Laguerre, Legendre, Chebyshev and Hermite polynomials and transformed by a multiplication with an cut-off function with respect to  $r_{\text{cut}} = 5\text{\AA}$ .

### 7.3 Remarks on the Implementation

*A  
Graph-  
Representation*

Since our methods is based on the atomic decomposition ansatz and the many-body expansion ansatz, the assembling of all pairs, triples and even bodies of an higher order in a system is necessary. Therefore, we use the structure of the underlying problem, which connects two atoms if they are close enough. This opens up the possibility to save the atoms or the molecules as an undirected graph. A graph  $G = (V, E)$  is given by a set of vertices  $V$  and a set of edges  $E$ . For a configuration  $\mathbf{X} = (X_1, \dots, X_M)$  incorporating  $M$  atoms, the corresponding graph is given by

$$G_{\mathbf{X}} = (V_{\mathbf{X}}, E_{\mathbf{X}}),$$

where the vertices are given by the incorporated atoms  $V_{\mathbf{X}} = \{X_1, \dots, X_M\}$  and a set of edges  $E = \{e_1, \dots, e_{|E|}\}$ . Here, two vertices are connected with an edge  $e = \{X_i, X_j\}$  if the atoms are 'close enough'. I.e. for a given maximal interaction distance  $r_{\text{cut}} > 0$ , two vertices are connected if their pairwise distance is smaller than  $r_{\text{cut}}$ . In this setting, the problem of finding all subsets of  $k$ -atoms, which are interacting can be reformulated to the problem of finding all connected components of size  $k$  of the graph  $G_{\mathbf{X}}$ . This way, we incorporate easily even rings and To compute such properties of graphs, so-called depth-first search (DFS) and breadth-first search (BFS) methods are used which typically involve  $O(|V|)$  cost as they step over each vertex exactly once, see [25]. This approach of a graph representation in combination with the many-body



expansion is treated in more detail in [33].

In our methods, we need to solve a least squares regression with a Tikhonov regularization. Therefore, we used the `sklearn.linear_model.Ridge` model to use its efficient solving methods. However, the Ridge model solely solves the linear least squares regression with Tikhonov regularization  $\Gamma = \alpha I$ . For a  $\alpha > 0$ , it minimizes the objective function

$$\|A \cdot c - y\|_2^2 + \alpha \|c\|_2^2, \quad (7.3.1)$$

for a matrix  $A \in \mathbb{R}^{n \times m}$ , weight vector  $c \in \mathbb{R}^m$  and target vector  $y \in \mathbb{R}^n$ . To use it for various Tikhonov matrices  $\Gamma \in \mathbb{R}^{m \times m}$ , we note that

$$\|\tilde{A} \cdot c - \tilde{y}\|_2^2 = \|A \cdot c - y\|_2^2 + \|\Gamma \cdot c\|_2^2$$

for the matrix

$$\tilde{A} := \begin{pmatrix} A \\ \Gamma \end{pmatrix} \in \mathbb{R}^{(n+m) \times m}$$

and the vector

$$\tilde{y} := \begin{pmatrix} y \\ 0 \end{pmatrix} \in \mathbb{R}^{n+m}$$

To optimize the most costly part of the code, which is the assembling of derivatives of the basis functions to incorporate the forces, we tried to use Numba. Numba translates Python functions to optimized machine code at runtime using the industry-standard LLVM compiler library. Nevertheless, in order to compile it with Numba we needed to change various data structures, for example it does not allow the usage of lists containing lists. Using the Numba compilation led to a speed up of 30 percent of the costly function, but marginal slowed down other parts due to the change to allowed data types. We stopped this investigation because the changes were very error prone.

## 7.4 Data Sets

In order to demonstrate the flexibility of the applications, we will test our methods on data sets of different characteristics. Therefore, we will examine periodic atomic environments of solids as well as trajectories of molecules.

### 7.4.1 The Tungsten Data Set

Tungsten is a white shiny, malleable metal of medium hardness, high density and strength. The density is almost the same as that of gold. Tungsten has the highest melting point (3695 K) and the highest boiling point (5828 K) among all metals and its alloys are utilized in numerous technological applications, for example as filament in incandescent lamps. The details of the atomistic processes behind the plastic behavior of tungsten have been investigated for a long time and many interatomic potentials exist in the literature reflecting an

evolution over the past three decades.

The W-14 data set used in [63] incorporates 9693 configurations of diverse atomic environments of elemental tungsten and is provided freely accessible in xyz format [2]. Here, each configuration is given by a periodic unit cell and 1 – 135 atoms in the contained in it. Of the different configurations, 25 percent contains only one particle, the vast amount of 65 percent contains 12 particles, 8 percent incorporating 50 and 2 percent 125 particles. The energies are computed using DFT/PBE and given in eV/Å. Also included are forces and stresses, including bcc primitive cell, 128-atom bcc cell, vacancies, low index surfaces, gamma-surfaces and dislocation cores. *Data Set*

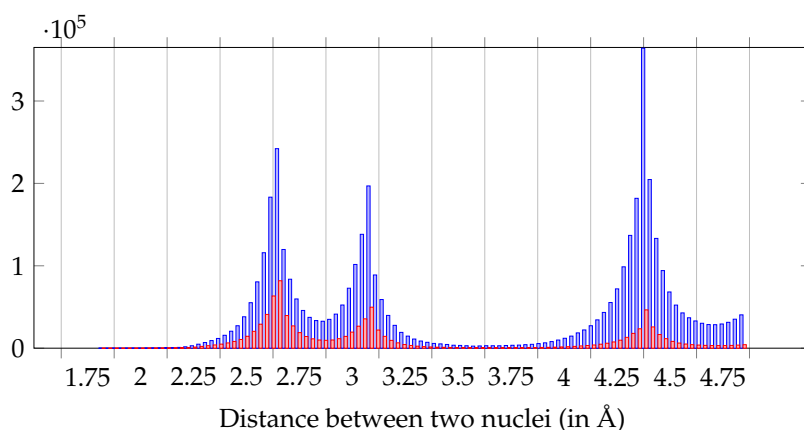


Figure 7.3: The histogram in blue shows the pairwise distances distribution in the data set W-14. The red histogram shows the normalized quantity of pairwise distances. Here, the value is divided by the area of possible occurrence. Favorable distances are at 2.75Å, 3.18Å and 4.47Å.

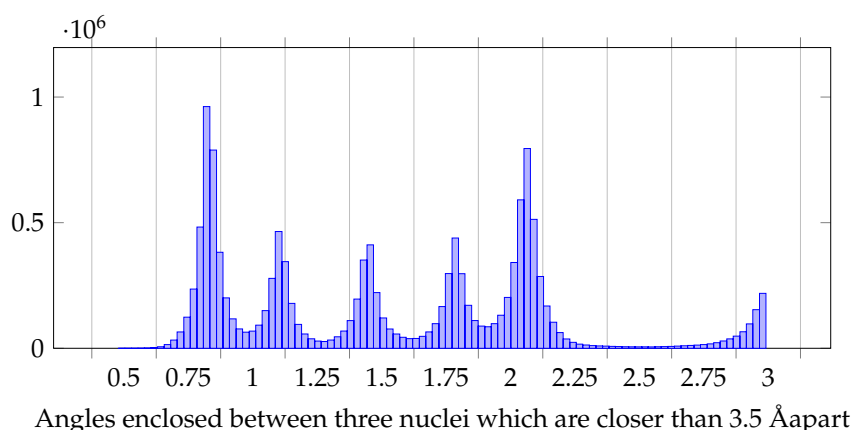


Figure 7.4: The histogram shows the angle distribution in the data set W-14. Here, we measured all occurring angles enclosed between particles which are closer than 3.5Å.

We will compare our results on the tungsten data set W-14 with the performance of similar approaches. Amongst other things we will consult potentials which are also based on the many-body decomposition up to a specific order. The Morse and Tersoff potential are pair potentials, i.e. they incorporate only pair interactions and are therefore not able to detect angle or dihedral angles. A more advanced potential is the Stillinger-Weber Potential (StiWe), which additionally incorporates interactions between three atoms and is thus receptive for angles. Moreover, we will compare our results to the Gaussian Approximation Potential (GAP) and the Moment Tensor Potential (MTP) [60]. Here, the GAP is also constructed based on the atomic decomposition ansatz, but not necessarily assumes a further decomposition into many bodies as we did. As a reminder, the atomic decomposition ansatz is the assumption, that two atoms do not interact if they are farther apart than a threshold  $r_{\text{cut}} > 0$ . With that, the energy of one atomic configuration decomposes in energies of connected components, which are closed under interaction. Those connected components, or local atomic environments, are further decomposed in a ANOVA-like form into its contributions of pairs, triples and on. In the GAP, one assumes that the energy function has a linear representation in some search space characterized by a basis set  $\Phi$ ,  $f = \mathbf{w} \cdot \Phi$ . Additionally, we assume the coefficients, or weights,  $\mathbf{w}$  to be Gaussian distributed. We shortly introduced this approach in section 2.2. Equivalently,  $f$  forms Gaussian process over the function space spanned by  $\Phi$ , since each evaluation of  $f$  has by construction Gaussian distribution. As in section 2.2 the GAP is given by mean of the predictive distribution, which in practice breaks down to the calculation of a matrix vector product. For a further introduction in the GAP, we refer to [5]. As the GAP and all the other potentials, the MTP is also based on the atomic decomposition ansatz. Instead of further dividing those connected components in pairs, triples and so on, Bartók constructed a certain kind of basis functions spanning the search set, which take an nondecomposed local atomic environment as an input. Moreover, he proved that the so constructed search space covers all possible energy potentials possible. Onto this search set, he performed a regression with a penalization with respect to the  $\mathcal{L}_2$  and the  $\mathcal{L}_0$  norm.

*Comparative  
Results*

#### 7.4.2 The Molecular Dynamics (MD) Data Set

The molecular dynamics (MD) MD-17 data sets used in [19, 20, 52] contains a wealth of trajectories of eight small molecules represented in Figure 7.5. A trajectory of a molecule is given by a number of frames, describing its physical movement. The atoms and molecules are allowed to interact for a fixed period of time, giving a view of the dynamic evolution of the system. The number of frames in the data set MD-17 range in size from 150k to nearly 1M conformational geometries. All trajectories are calculated at a rather high temperature of 500 K to achieve exhaustive exploration of the potential-energy surface of such small molecules. The molecules have different sizes, from the smallest one, Ethanol with 9 atoms, to the largest incorporated, Aspirin with 21 atoms, and the molecular PESs exhibit different levels of complexity. The energy range across all data points within a set spans from 20 to 48 kcal/mol.

*Data Set*

Force components range from 266 to 570 kcal/mol/Å. The total energy and force labels for each dataset were computed using the PBE+vdW-TS electronic structure method. All geometries are in Angstrom, energies and forces are given in kcal/mol and kcal/mol/Å respectively. The data is provided in xyz format with one file per conformation and is freely available [2]. The energy and force labels for each geometry are included in the comment line. Positions are given in Angstroms, energies are given in kcal/mol. Due to performance reasons, we will only restrict to a manageable subset of the data. Therefore, we randomly chose 2500 frames per molecule and test our methods on this subset to get a feeling for the applicability.

### Comparative Results

In [56] there are Comparative deep tensor neural networks (DTNN) used receiving the molecular structures Results through a vector of nuclear charges and a pairwise atomic distances expanded in a Gaussian basis. Similar approaches have been applied to the entries of the Coulomb matrix for the prediction of molecular properties in [47]. The authors of [20] developed a gradient-domain machine learning (GDML) approach to construct molecular force fields. Instead of computing the derivative of the PES, the interatomic forces are directly learned by a functional relationship using a generalization of a kernel

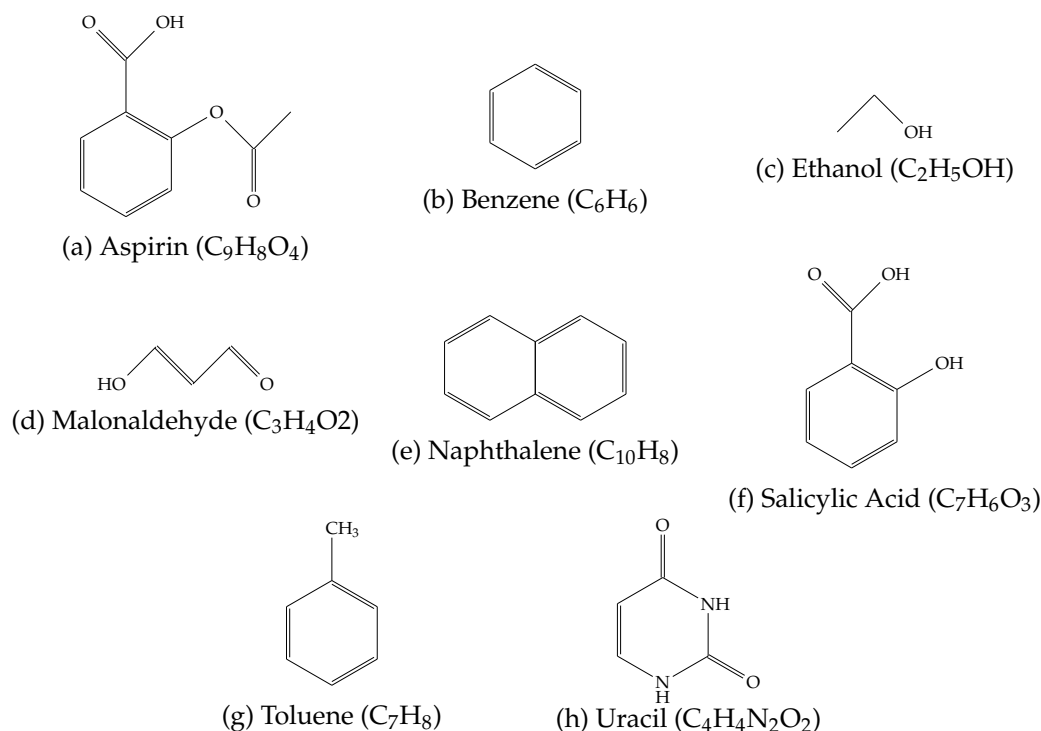


Figure 7.5: Incorporated molecules in the MD-17 dataset in skeletal formula. In a skeletal formula, carbon atoms are represented by the unlabeled vertices (intersections or termini) of line segments. Hydrogen atoms attached to carbon are implied. An unlabeled vertex is understood to represent a carbon attached to the number of hydrogens required to satisfy the octet rule.

ridge regression. To obtain the PES, the force field is integrated. In [20], the GDML method is expanded by additionally incorporating spatial and temporal physical symmetries (sGDML).



## Chapter 8

# Numerical Results

In the following we test our mathematical approaches to different data sets, we presented in section. To show the different applications, we consider solids embedded in a periodic environment as well as MD trajectories of small molecules. During the results on tungsten, we present the errors in the format mean  $\pm$  standard variation. Errors on energy terms will be in eV/Å and on forces in eV.

### 8.1 Results on Tungsten Data

In this section, we present our results on the W-14 data set. Considering the highest order to be two, i.e. only considering pairs of atoms in our problem, we are able to achieve a root mean squared error of 0.22eV on the energy terms and 0.25 eVÅ<sup>-1</sup> on the forces over the whole data set using a 9-fold cross validation and a  $L^2$ -regularization. In comparison, the Morse potential and the Tersoff potential which are both pair potentials as well achieve an error of 1.32eV (0.49 eVÅ<sup>-1</sup>) and 0.53eV (0.37 eVÅ<sup>-1</sup>) respectively. When raising the order of incorporated bodies to three, the error reduces to 0.12eV on the energy and 0.14eVÅ<sup>-1</sup> on the forces. In comparison to another three-body potential, the Stillinger-Weber potential which achieves an error of 1.73eV (0.39eVÅ<sup>-1</sup>). In both terms we are able to reduce the error by more than 50 percent. Note that we are able to especially increase the accuracy on the energy terms. Two other potential is the Moment Tensor potential (MTP) and the Gaussian Approximation potential (GAP). The GAP is able to achieve an error of 0.06eVÅ<sup>-1</sup> and the MTP an error of 0.04eVÅ<sup>-1</sup> on the force terms. But this is no surprise since they do not truncate the incorporated sets of particles as we do, but incorporate all lying inside an cut-off radius. Additionally, all of those comparative potentials are also incorporating stress terms coming in as the second derivative of the potential function, while we only involved energy and force terms.

#### 8.1.1 Varying the Penalization Norm

As we mentioned earlier, the choice of the penalizing norm in a least squares

*The  
Penalization  
Norm*

regression problem is crucial. In the regression formulation this norm defines the search set as a bounded ball in a prescribed function space. In the Bayesian formulation this norm specifies a prior distribution on the weights. In practice, we are able to encode a certain grade of additional knowledge with the help of the penalization norm. In the following, we will investigate the impact of three different penalizations.

### *L<sup>2</sup> Norm*

As the first case, we consider the basic  $L^2$  regularization characterized by a Tikhonov matrix  $\Gamma = \alpha/2 \cdot I$ . This case corresponds to a gaussian prior distribution with covariance matrix  $\alpha^{-1}I$  and a  $L^2$ -bounded ball in the function space, which gets larger as  $\alpha$  decreases. Nevertheless, we will observe an oscillating behavior of the approximation in regions which are underrepresented by the data. Since the W-14 data set does not include distances smaller than  $2\text{\AA}$ , the fitted functional oscillates in the interval  $[0, 2\text{\AA}]$  in order to better capture the behavior in represented regions of larger distances. For an overview over the error distribution with an  $L^2$  regularization, see table 8.19.

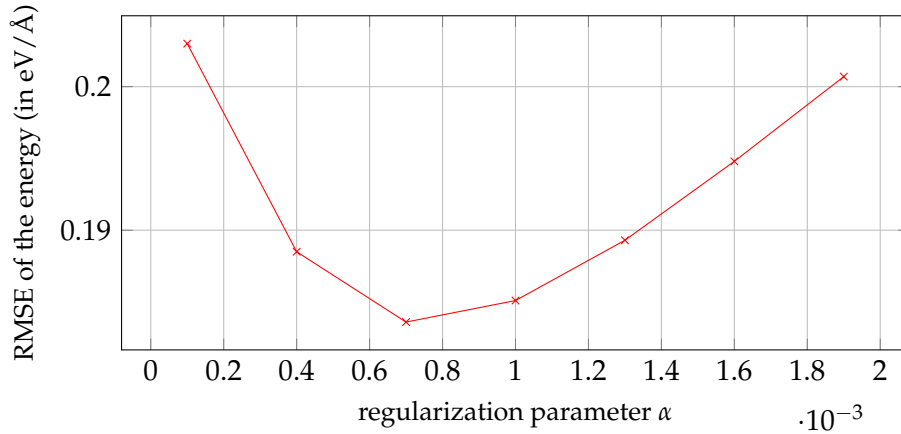


Figure 8.1: A typical error evolution under a varying regularization term. Here, the root mean squared error (RMSE) on the energy is plotted against the regularization parameter  $\alpha$  of the  $L^2$  regularization. The decreasing part describes a phase of overfitting. Here, a lack of regularization leads to a too strong orientation on the training data. Otherwise, the increasing part describes a phase where the regularization is too high to capture the model.

### *Sobolev-like Norm*

To confine those oscillations, we also observed the behavior under an  $H^s$ -like norm regularization. As we introduced in section this norm forms a Tikhonov regularization with  $\Gamma = (M + DMD^T)^{s/2}$  with mass matrix  $M$  and total derivative of the basis functions  $D$ . Even though we do not consider Fourier basis functions in our numerical application, we will define a Sobolev-like norm by considering the Tikhonov matrix  $\Gamma = (\text{diag}((1 + 1^2)^{s/2}, \dots, (1 + p^2)^{s/2}))$ , where  $p$  encodes the degree of the underlying orthogonal polynomials. Thus, we choose a stricter penalization for higher oscillations of the basis functions to limit oscillating behavior in uncertain regions. Penalizing with respect to this



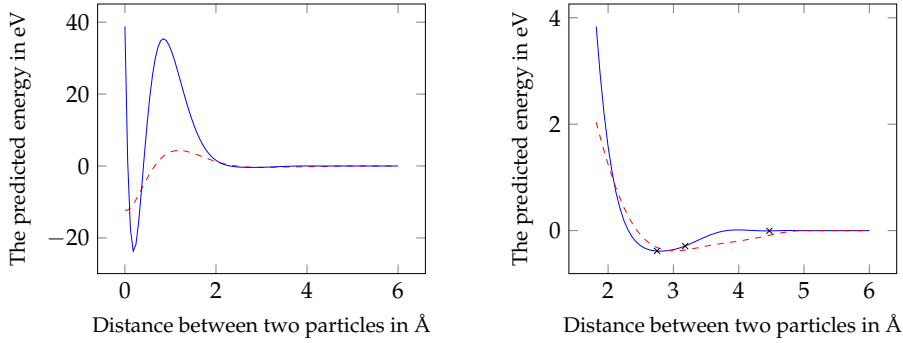


Figure 8.2: The two body part of the learned potential, mapping the distance between two particles onto its associated energy contribution. The potential fitted with a cut-off radius of  $5\text{\AA}$ , based on Laguerre polynomials up to degree 20. The blue solid line shows the outcome using a  $L^2$  regularization with  $\alpha = 0.0007$ . On the left hand side, we plotted the potential function starting from zero  $\text{\AA}$ . Since the data does not include particles which are closer than two  $\text{\AA}$ , we notice an oscillation behaviour of the function in those regions. On the right hand side we plotted the same potential function on the area which is presented in the data. We notice that in fact the potential function captures the characteristics of the data. The three additionally marked points in the right plot are the favourable distances incorporated in the data set and are in fact local minima of the learned function. In comparison, the red dashed function is fitted using a  $H^{1/2}$ -like regularization. While the outcome is way more shallow in the unrepresented data, the favourable distances are not as well captured.

norm is equivalent to assuming a gaussian prior on the weights with covariance matrix  $2\Gamma^{-1/2}$ . Here, the entries decrease along the diagonal. Since the weights are also centered, this describes that the entries are becoming more unlikely to differ from zero as the index increases. In the regression view point, this penalization describes a bounded ellipse in the underlying function space, getting narrower in the direction of higher oscillation. In figure , we see the impact of a  $H^{1/2}$  regularization (dashed) in comparison to a  $L^2$  regularization. Here, one sees the learned energy potential based on Laguerre polynomials, describing pairs of atoms by mapping its pairwise distance onto its corresponding energy contribution. The function is not only flatter in the unrepresented area, but also in the presented one, which is not desired and leads to an higher error.

To also penalize the order of the body, we did a Tikhonov regularization with  $\Gamma = (\text{diag}((1 + 1^2)^{(s+k)/2}, \dots, (1 + p^2)^{(s+k)/2}))$ . Here, additionally to the degree of the underlying polynomials, also the body order is penalized. This will bias the model towards having smaller weights for bodies of higher order. As we see in table 8.19 this makes the prediction worse compared to  $L^2$  regularization, if we are considering only small bodies. On the other hand for five bodies it suddenly outperforms  $L^2$  and the Sobolev-like regularization.

*Introducing a Penalization for High Body Orders*

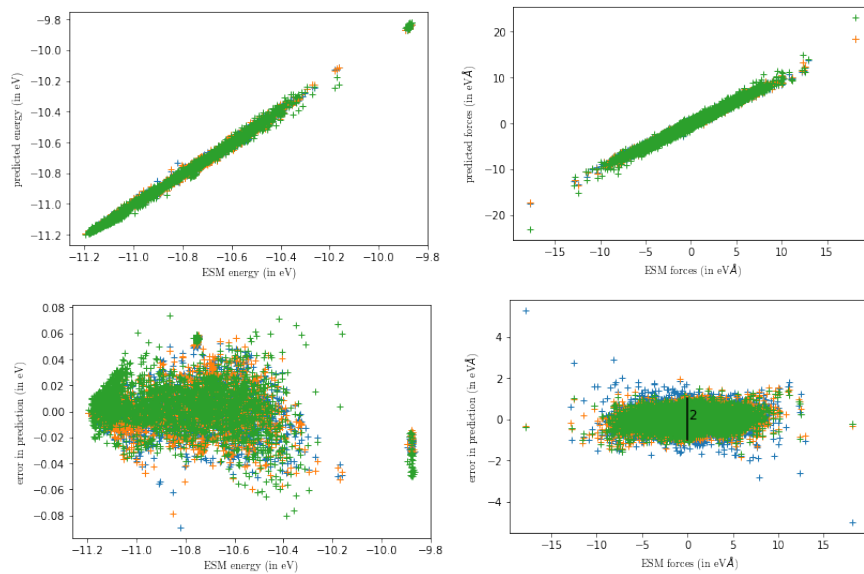


Figure 8.3: Comparison of the predicted energy and forces with the ESM calculated ones. Here, the green marks describes the behaviour of a model trained on Laguerre polynomials, the orange marks a model trained on Chebyshev polynomials and the blue marks a model trained on Legendre Polynomials. All models are trained on Polynomials up to degree ten and with a  $L^2$  regularization with  $\alpha = 0.0007$ . In this stadium, the polynomials behave very similar in their predicting accuracy.

		force (eVÅ <sup>-1</sup> )		
	$\alpha$	RMSE	MSE	WCE
Laguerre	0.0007	0.1445 ± 0.023	0.1053 ± 0.014	1.1733 ± 0.132
	0.1	0.1809 ± 0.031	0.1220 ± 0.025	1.3486 ± 0.210
	1	0.1934 ± 0.035	0.1304 ± 0.028	1.3901 ± 0.259
	10	0.2086 ± 0.038	0.1500 ± 0.031	1.2939 ± 0.554
Legendre	0.007	0.1511 ± 0.260	0.1031 ± 0.021	0.9739 ± 0.154
	0.1	0.1574 ± 0.028	0.1071 ± 0.023	1.1367 ± 0.208
	1	0.1609 ± 0.027	0.1094 ± 0.023	1.1861 ± 0.258
	10	0.1640 ± 0.028	0.1117 ± 0.024	1.1463 ± 0.117
Chebyshev	0.007	0.1559 ± 0.028	0.1069 ± 0.022	1.0280 ± 0.157
	0.1	0.1554 ± 0.028	0.1077 ± 0.022	1.0405 ± 0.121
	1	0.1582 ± 0.028	0.1094 ± 0.023	1.0725 ± 0.099
	10	0.1634 ± 0.028	0.1126 ± 0.023	1.1556 ± 0.120

		energy (eV)		
	$\alpha$	RMSE	MSE	WCE
Laguerre	0.0007	0.1289 ± 0.036	0.0978 ± 0.027	0.5749 ± 0.187
	0.1	0.1711 ± 0.033	0.1269 ± 0.031	0.6357 ± 0.084
	1	0.1919 ± 0.034	0.1369 ± 0.035	0.7863 ± 0.258
	10	0.2370 ± 0.042	0.1678 ± 0.045	0.8643 ± 0.217
Legendre	0.007	0.1511 ± 0.026	0.1031 ± 0.021	0.9739 ± 0.154
	0.1	0.1537 ± 0.036	0.1100 ± 0.030	0.5678 ± 0.129
	1	0.1684 ± 0.025	0.1197 ± 0.022	0.7535 ± 0.154
	10	0.2181 ± 0.052	0.1586 ± 0.049	0.8302 ± 0.186
Chebyshev	0.007	0.2065 ± 0.084	0.1435 ± 0.061	0.6237 ± 0.134
	0.1	0.1785 ± 0.037	0.1309 ± 0.032	0.6229 ± 0.207
	1	0.1792 ± 0.026	0.1306 ± 0.027	0.5624 ± 0.090
	10	0.2088 ± 0.030	0.1441 ± 0.034	0.6732 ± 0.143

Figure 8.4: The error distribution of a least squares regression penalizing with an  $L^2$  norm with Tikhonov matrix  $\Gamma = \alpha I$ . Here, we incorporate bodies up to order three and approximate on the search set based on Laguerre, Legendre and Chebyshev polynomials up to degree twenty. Therefore, we operate a eight-fold cross validation on the entire data set W-14 and train on the energies and forces, a cut-off radius  $r_{\text{cut}} = 5\text{\AA}$  and a set including 1k data points. In the above tabular one sees the error on the force terms in eVÅ and on the bottom tabular the error on the energy terms in eV. RMSE describing the root mean squared error, MSE the mean squared error and WCE the worst case error.

$k_{max}$	$\Gamma$	forces (eVÅ <sup>-1</sup> )		
		MAE	RMSE	WCE
2	$L^2$	0.2503± 0.02	0.3676± 0.02	1.7253± 0.14
2	$H^{d/2}$	0.3015± 0.04	0.4517± 0.05	3.1087± 0.79
2	$H^{(d+k)/2}$	0.3469± 0.01	0.5393± 0.02	4.0410± 0.36
3	$L^2$	0.1620± 0.01	0.2426± 0.03	1.0229± 0.13
3	$H^{d/2}$	0.1871± 0.01	0.2729± 0.02	1.4044± 0.04
3	$H^{(d+k)/2}$	0.2688± 0.08	0.4037± 0.13	2.3439± 1.05
41	$L^2$	0.1494± 0.00	0.2169± 0.00	1.1746± 0.12
41	$H^{d/2}$	0.1715± 0.01	0.2472± 0.01	1.3288± 0.13
41	$H^{(d+k)/2}$	0.1678± 0.02	0.2497± 0.03	1.3955± 0.31
42	$L^2$	0.1570± 0.02	0.2335± 0.03	1.3202± 0.28
42	$H^{d/2}$	0.1530± 0.02	0.2296± 0.03	1.3793± 0.25
42	$H^{(d+k)/2}$	0.1367± 0.01	0.2299± 0.00	1.7909± 0.05
51	$L^2$	0.1724± 0.01	0.2460± 0.01	1.3647± 0.11
51	$H^{d/2}$	0.1319± 0.00	0.1940± 0.00	1.1356± 0.17
51	$H^{(d+k)/2}$	0.1310± 0.00	0.1934± 0.00	1.0948± 0.15

$k_{max}$	$\Gamma$	energy (eV)		
		MAE	RMSE	WCE
2	$L^2$	0.1678± 0.04	0.2235± 0.04	0.4336± 0.13
2	$H^{d/2}$	0.3882± 0.14	0.5605± 0.19	1.2438± 0.40
2	$H^{(d+k)/2}$	0.4364± 0.14	0.6984± 0.30	1.3134± 0.24
3	$L^2$	0.1382± 0.02	0.2279± 0.05	0.5290± 0.13
3	$H^{d/2}$	0.2384± 0.05	0.3397± 0.06	0.7628± 0.18
3	$H^{(d+k)/2}$	0.3695± 0.21	0.6013± 0.46	1.2892± 0.48
41	$L^2$	0.1304± 0.03	0.1990± 0.05	0.5289± 0.16
41	$H^{d/2}$	0.2359± 0.01	0.3496± 0.03	0.8881± 0.12
41	$H^{(d+k)/2}$	0.2779± 0.05	0.3939± 0.07	0.9973± 0.29
42	$L^2$	0.2366± 0.07	0.3334± 0.09	0.8143± 0.32
42	$H^{d/2}$	0.2595± 0.08	0.3772± 0.11	0.8516± 0.29
42	$H^{(d+k)/2}$	0.3274± 0.03	0.5737± 0.08	0.8692± 0.15
51	$L^2$	0.4069± 0.06	0.5719± 0.09	1.6639± 0.92
51	$H^{d/2}$	0.2693± 0.03	0.4150± 0.07	0.9928± 0.24
51	$H^{(d+k)/2}$	0.2535± 0.03	0.3970± 0.06	1.0139± 0.19

Figure 8.5: Comparison of the different Tikhonov regularizations. Here,  $k_{max}$  denotes the maximal incorporated order of bodies.  $L^2$  the a regularization with respect to  $\Gamma = \alpha I$ ,  $H^{d/2}$  the Sobolev-like regularization and  $H^{(d+k)/k}$  the Sobolev-like regularization with an additional penalization of the high body orders.

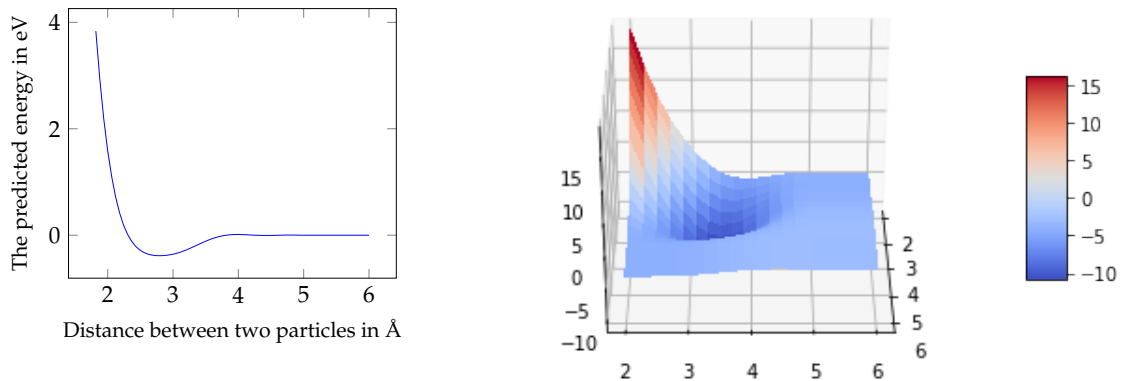
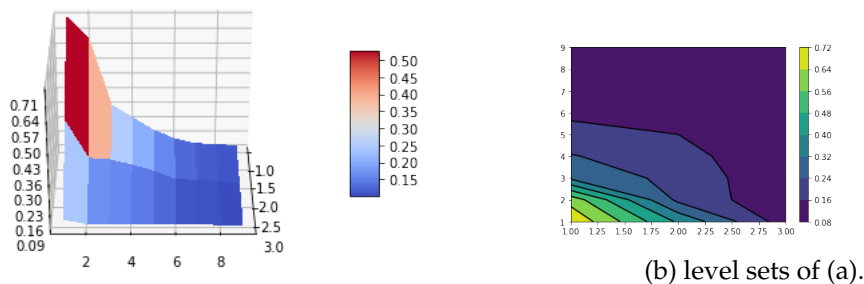
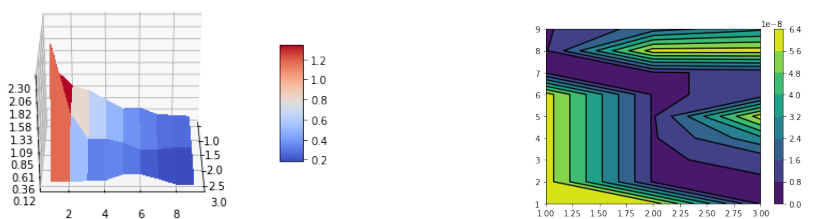


Figure 8.6: The fitted potential function. On the LHS the pair potential mapping each distance between two particles onto an energy in eV. On the RHS a hypersurface of the three-body potential. To visualize the three-dimensional function, the third distance is fixed to the value  $2\text{\AA}$ , the other two distances are marked in the  $x$ - and  $y$ -axis. The corresponding energy is given in the  $z$ -axis. The two-body potential captures the most occurring distance.



(a) Test error evolution of the adaptive sparse grid approach.



(c) corresponding choice of the regularization parameter  $\alpha$  in each step.

(d) level set of (c)

Figure 8.7: Regularization parameter  $\lambda$  optimally chosen in each step, with adaptive SG approach using Laguerre polynomials incorporating bodies up to order four. With an adaptive regularization an error of  $0.12eV/\text{\AA}$  on the forces can be achieved. Choosing the optimal overall regularization results into an error of  $0.18eV/\text{\AA}$ . on the  $x$ -axis one sees the size of the search space, one  $y$ -axis the body order.

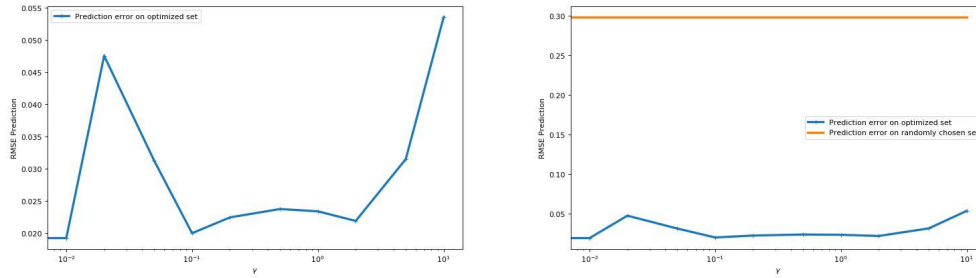


Figure 8.8: Both plots show the RMSE of a least-squares predictor. In the left picture one sees the prediction value for different values of  $\gamma$ . When one averages over all values of  $\gamma$  one gets a mean of 0.029 and a standard deviation of 0.011. As a comparison in the right plot we also plotted the prediction error of a penalized least squares regression trained on a random subset of the data. It has a RMSE of 0.029

### 8.1.2 Incorporating the Regularization Parameter

A sign of too little data is an overfitting behaviour, this can be cushioned to a certain degree by a more flexible penalization. So far, we have chosen a *general* regularization parameter  $\lambda$  in an optimal way. Based on this penalization we then defined a cost function. Nevertheless, the dimensional splitting of the problem in lower dimensional sub-problems also allows a more diverse regularization. One can incorporate the regularization parameter itself in the adaptive method by first discretizing an interval  $\Lambda$  of possible regularization parameters  $\lambda$  and then expand the cost-function by  $\lambda$ , instead of just previously fixing one. The expanded adaptive sparse grid approach allows to choose the optimal regularization parameter in each step, allowing a stricter penalization in some regions than in others. This leads to a more flexible algorithm that is able to cushion overfitting behavior by increasing the regularization in these regions and consequently improving the approximation power without requiring more information, see 8.7.

### 8.1.3 Active Learning

As we derived in section 4.4 we want to choose the training set which minimizes the determinant of  $GG^T$  where

$$G = (A^T A + \Gamma^T \Gamma)^{-1} A^T,$$

for the Tikhonov matrix  $\Gamma$ . Let us first consider the unregularized case. To simplify things, we fix the number of training points such that the matrix  $A$  is quadratic. To do this we only use 136 training examples as this is also the number of features and thus columns of  $A$ . This way, the calculation reduces to the maximization of  $|\det(A)|$  and is very cost-efficient. The active learning algorithm got a pool of 6000 data points to choose the best 136 from. We compare it with 136 randomly chosen ones. As we see in 8.8 the predictor trained on the output of the active learning algorithm does considerably better. On the

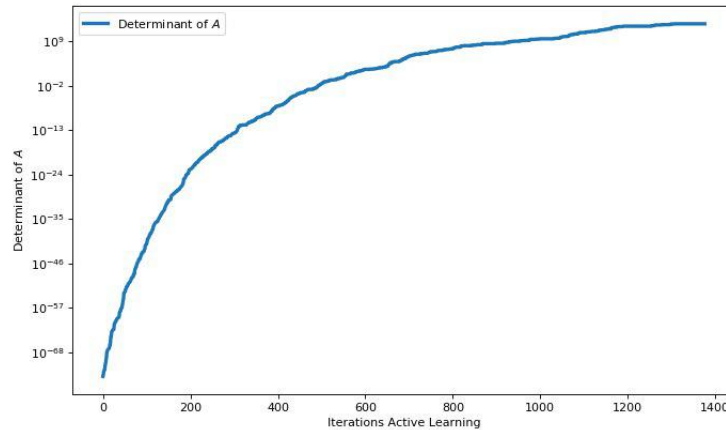


Figure 8.9: The determinant of the square matrix  $A$  while exchanging rows for other data points to increase the determinant. Since this is the case where the algorithm is based on a square matrix  $A$  the training set size is 136.

$x$ -axis there are different values for the extrapolation grade  $\gamma$ . Even for values as big as 10 which imply that we only take new rows into the matrix if they increase determinant eleven-fold. The weak dependence on  $\gamma$  is explained when one sees the algorithm run. It initially increases the determinant rapidly, going from a near-zero determinant to determinants in the trillions and then gets slower. When it comes to the point that  $\gamma$  matters and the algorithm halts the matrix determinant was already vastly improved. This can also be seen in figure 8.9.

A further step is to actually maximize the determinant of Fisher information  $A^T A$  and thus minimize the Cramér-Rao bound. This also lead to minimization of our prediction variance. In figure 8.10 one can see the comparison of a least squares algorithm trained on a random subset of a given size and a least squares algorithm trained on the resulting matrix  $A$  after optimizing the determinant of  $A^T A$  using active learning. Here  $\gamma$  is fixed at 0.01. Table 8.11 gives a more precise view on the data used to generate figure 8.10. Especially for small training sets the optimized training data results in a much better prediction. When we increase the size of the training data the randomly chosen subset catches up. The whole data set consisted of 9693 and choosing a random subset of over 1000 data points probably already covers most of the area on which we will later test. Therefore the predictor will not need to extrapolate and does well. The optimization of the determinant of  $A^T A$  which also seeks to reduce the amount of extrapolation that will be needed later does not deliver such a big gain anymore. This can also be seen in figure 8.12. Here the training set size is 800. The determinant still grows but if one compares the  $y$ -axis of 8.12 to 8.9 there is less growth. There are also longer phases where the algorithm does not find a new data point  $\Phi(x^*)$  with a sufficient extrapolation grade  $\psi(x^*)$  and hence the determinant stays constant. This also underpins the interpretation of that the 800 random data points already cover the space

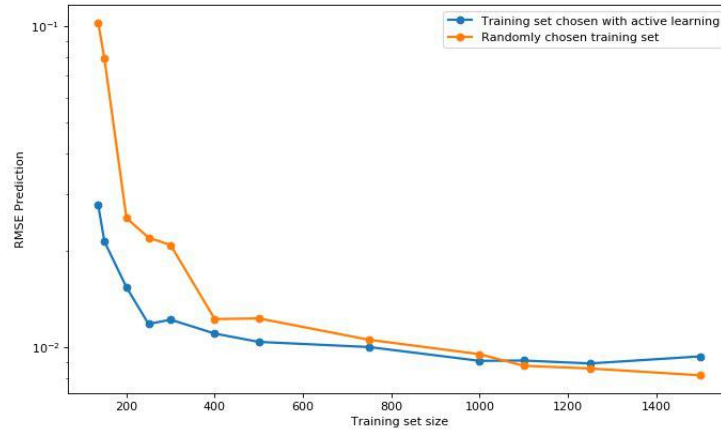


Figure 8.10: One data point corresponds to fixing a training size and either choosing a random subset of that size or maximizing the determinant of  $A^T A$ . Then we varied the training set size by just adding new random points to the training set for the blue line. For the orange line we increased the size of the matrix  $A$  hence giving the active learning algorithm a bigger matrix to optimize.

#TS	optimal TS	random TS
136	0.02770	0.10240
150	0.02136	0.07937
200	0.01537	0.02525
250	0.01182	0.02195
300	0.01219	0.02081
400	0.01103	0.01224
500	0.01039	0.01230
750	0.01002	0.01055
1000	0.00907	0.00952
1100	0.00909	0.00876
1250	0.00891	0.00859
1500	0.00937	0.00818

Figure 8.11: RMSE evolution on the energy ( $\text{eV}\text{\AA}^{-1}$ ) for an optimal chosen training data set versus a randomly chosen one. More explanation about the creation of the data is given in the description of figure 8.8. The first column is the size of the training set.



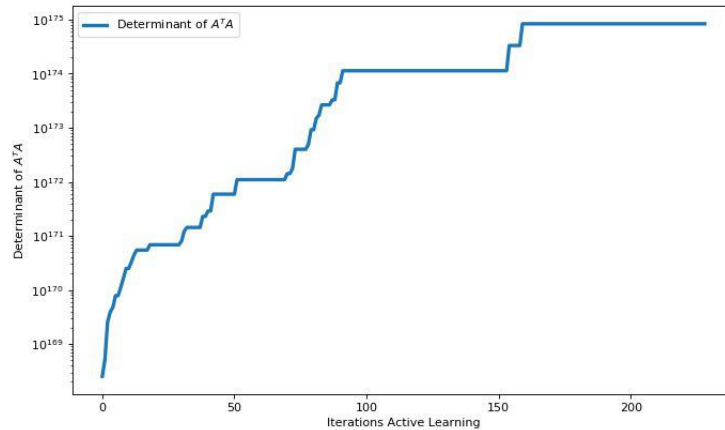


Figure 8.12: The figure shows how the determinant of  $A^T A$  develops when iterating the active learning algorithm. The training set size is set to 800. The growth is slower than the growth in figure 8.10 where the training set size is only 136.

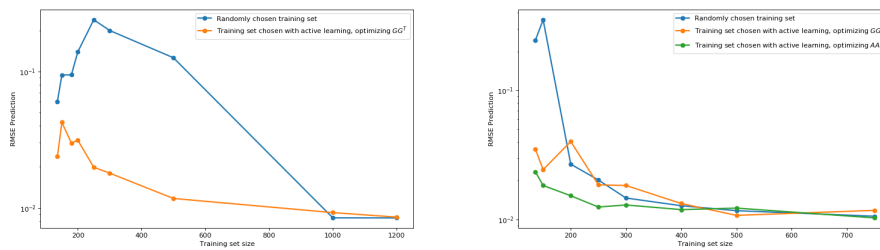


Figure 8.13: In this figure the active learning algorithm minimizes the determinant of  $GG^T$  through maximizing the determinant of  $(GG^T)^{-1}$ . On the left hand side we plot the regression error of a linear regression trained on a random subset as comparison. In the right figure we did another experiment where we also ran an active learning algorithm optimizing  $A^T A$ .

well enough and it is hard to find exploratory data points to improve  $A$ .

In theory the optimal prediction variance should be attained if we minimize the variance of  $GG^T$  directly. One does not have to take any matrix inverse as the determinant commutes with the inverse and we can calculate it directly. The active learning algorithm maximizes the determinant of  $(GG^T)^{-1}$ . The result of this can be seen in figure 8.13. Optimizing  $GG^T$  still outperforms just random sampling but it seems to be worse than optimizing  $A^T A$  in practice. Other experiments we did also indicated that optimizing  $A^T A$  leads to a smaller prediction error than optimizing  $GG^T$ .

#TS	maximized $\det(A^T A)$	minimized $\det(GG^T)$	random subset
136	0.02341	0.03508	0.24456
150	0.01834	0.02435	0.35243
200	0.01527	0.04017	0.02682
250	0.01248	0.01853	0.02022
300	0.01296	0.01834	0.01466
400	0.01192	0.01329	0.01278
500	0.01225	0.01076	0.01170
750	0.01029	0.01177	0.01059

Figure 8.14: Data used to generate the right figure in 8.13. The leftmost column is the size of the training set.

## 8.2 Results on the Molecular Dynamics (MD) Data Set

In the following section, we will apply our methods to a randomly chosen subset of the molecular dynamics MD-17 data set. Although we mentioned existing methods in section 7.4.2, the aim of this section is not to outperform them, but to optimize the accuracy under a given work load in a simplified setting. Instead of considering all atoms of a molecule, we will restrict ourself to simply incorporate subsets up to a specific size. Due to reasons of cost, we will limit the maximal set of atoms by four, which is a drastic simplification when describing molecules up to 21 atoms like aspirin. Nevertheless, we will show that even in this simplified setting, we are able to achieve chemical accuracy of less than 1 kcal/mol on most molecules, except on aspirin, see Figure 8.15 . However, an energy prediction when considering even smaller sets, i.e. sets of pairs or triples, fail to capture the energy properly. Thus, an incorporation of up to four atoms seems to be the smallest regime which leads to an appropriate prediction, see Figure 8.16 and Table 8.19.

In the following, we use a cut off radius of  $2\text{\AA}$ . This choice ensures that all covalent bonds of the molecules are represented. A covalent bond, also called a molecular bond, is a chemical bond that involves sharing of electron pairs between atoms. In the structural formula of a molecule they are represented by an edge connecting two atoms. The distances between two atoms forming an covalent bond represented in the data are all smaller than  $2\text{\AA}$ , but not much smaller.

Thus, when incorporating all pairs of atoms which are not farther apart than  $2\text{\AA}$  they are most likely also forming a covalent bond. Moreover, we will most of the time only considering one specific basis set instead of comparing all polynomials, since they behave quite similar. This is due to the small size of one molecule compared to a periodic environment describing a solid and the consideration of basis functions which oscillate not that much. Additionally, we solely will train on the energies of the data set.

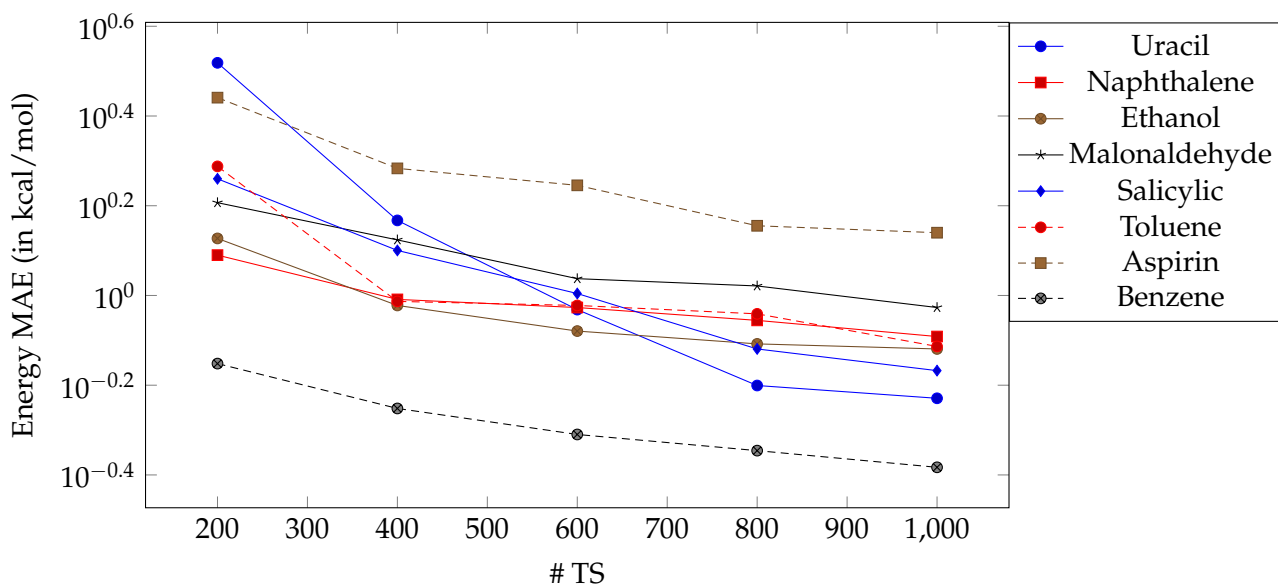


Figure 8.15: The accuracy of the regression model in terms of the mean absolute error (MAE) as a function of training set size. Incorporating only interactions of four atoms, we are able to achieve chemical accuracy of less than 1 kcal/mol for all molecules except aspirin, which is with 21 atoms the largest one, see Table 8.2

### 8.2.1 A Sparse Tensor Product Approach

We observe, that we are able to achieve a similar accuracy incorporating less degrees of freedom when considering a more sparse search space. Depending on the complexity of the molecule, we are in some cases able to notably increase the approximation accuracy while using a fixed amount of degrees of freedom. In the following sections we focus on four molecules: Benzene, Uracil, Salicylic and Aspirin. Note, that even if we are approximating each molecule with the same search set, the degrees of freedom vary, since it additionally depends on the molecule itself. As we mentioned in section 7.1, in the case of differing atom numbers we fit a separate function for each possible coalition of atoms. This depends on the number of different atoms incorporated in one molecule. For example, in the case of Benzene  $C_6H_6$  there are only three different types of atomic pairs (C-C, H-H and C-H), while in Uracil  $C_4H_4N_2O_2$  there are already ten. Thus, although we are approximating the energy of each and every molecule by the same search set, the different functional forms leads to different degrees of freedom.

Molecule	Formula	Size Training Set				
		200	400	600	800	1000
Uracil	$C_4H_4N_2O_2$	3.3	1.47	0.93	0.63	0.59
Naphthalene	$C_{10}H_8$	1.23	0.98	0.94	0.88	0.81
Ethanol	$C_2H_6O$	1.34	0.95	0.833	0.78	0.76
Malonaldehyde	$C_3H_4O_2$	1.61	1.33	1.09	1.05	0.94
Salicylic	$C_7H_6O_3$	1.82	1.26	1.01	0.76	0.68
Toluene	$C_7H_8$	1.94	0.97	0.95	0.91	0.77
Benzene	$C_6H_6$	0.70	0.56	0.49	0.45	0.41
Aspirin	$C_9H_8O_4$	2.76	1.92	1.76	1.43	1.38

Table 8.1: The accuracy of the regression model in terms of the mean absolute error (MAE) dependent of the training set size. Incorporating only interactions of four atoms, we are able to achieve chemical accuracy of less than 1 kcal/mol for all molecules except aspirin, which is with 21 atoms the largest one.

Aspirin	Deg. of Freedom	47	233	716	1784	3860	7628	9397
	FG	31.72	30.33	27.73	21.26	6.59	3.49	3.44
	SG	29.55	19.58	11.80	8.45	4.98	3.64	3.40
Benzene	Deg. of Freedom	21	97	294	735	1604	3200	5927
	FG	27.10	22.92	14.67	4.07	0.76	0.45	0.35
	SG	24.67	19.74	4.45	0.90	0.53	0.42	0.35
Salicylic	Deg. of Freedom	45	229	721	1831	4021	6437	12229
	FG	26.70	25.46	21.82	13.60	3.96	2.13	1.20
	SG	26.36	23.80	20.17	4.13	2.49	1.74	1.08
Uracil	Deg. of Freedom	61	309	693	1361	2487	4277	5487
	FG	25.98	20.97	15.66	8.71	3.23	1.09	0.82
	SG	20.25	7.18	3.42	1.86	0.91	0.72	0.65
Toluene	Deg. of Freedom	15	120	369	927	2025	4038	6134
	FG	24.62	23.52	19.51	13.22	2.96	1.66	1.15
	SG	23.38	12.74	4.28	2.53	2.18	1.00	
Ethanol	Deg. of Freedom	35	105	354	909	1573	3286	7716
	FG	16.89	16.05	14.02	10.87	7.61	2.79	1.19
	SG	15.75	6.46	4.22	2.72	1.97	1.38	1.08

Table 8.2: Comparison of the performance on full and sparse grid in terms of the mean absolute error (MAE) dependent of the allowed maximal number of degrees of freedom.

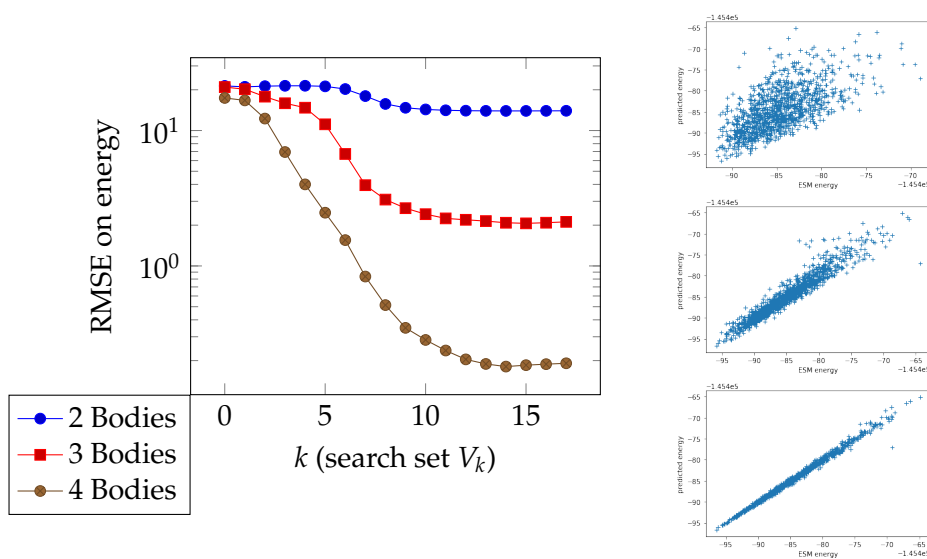


Figure 8.16: The approximation accuracy when incorporating all sets of maximally two, three or four atoms which are closer than  $r_{\text{cut}} = 2\text{\AA}$ . On the left hand side, the root mean squared error (RMSE) on the energy of Benzene is written as function of  $k$ , which describes the search set  $V_k$  (based on Legendre polynomials). On the right hand side, the corresponding predictive accuracy when incorporating body orders up to two, three or four respectively.

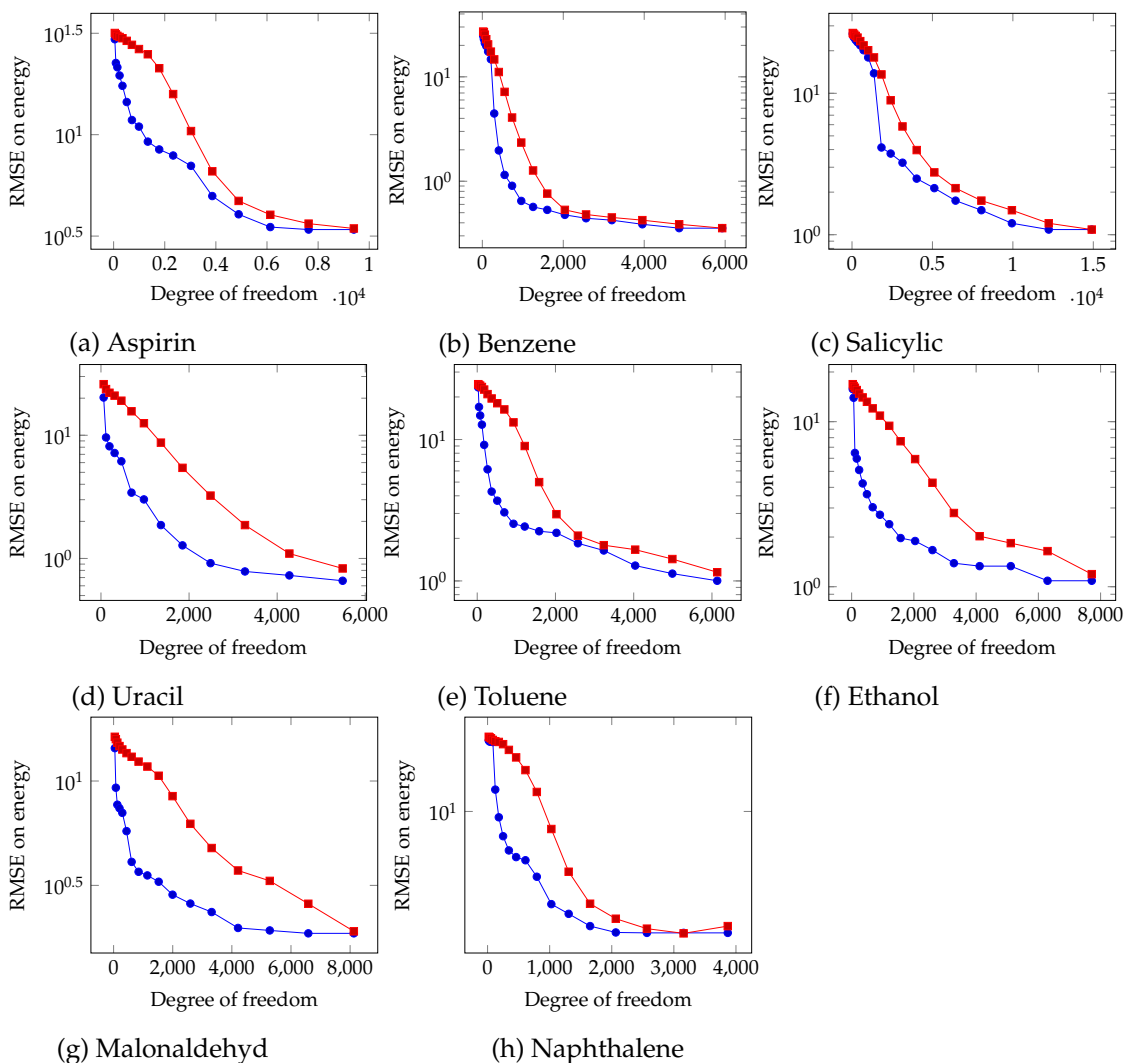


Figure 8.17: Comparison of the performance on full grid (red squares) and sparse grid (blue dots) in terms of the mean absolute error (MAE) depending on the allowed maximal number of degrees of freedom. See figure 8.18 for an overview over the corresponding index set. See table 8.2 for details on the MAE error.

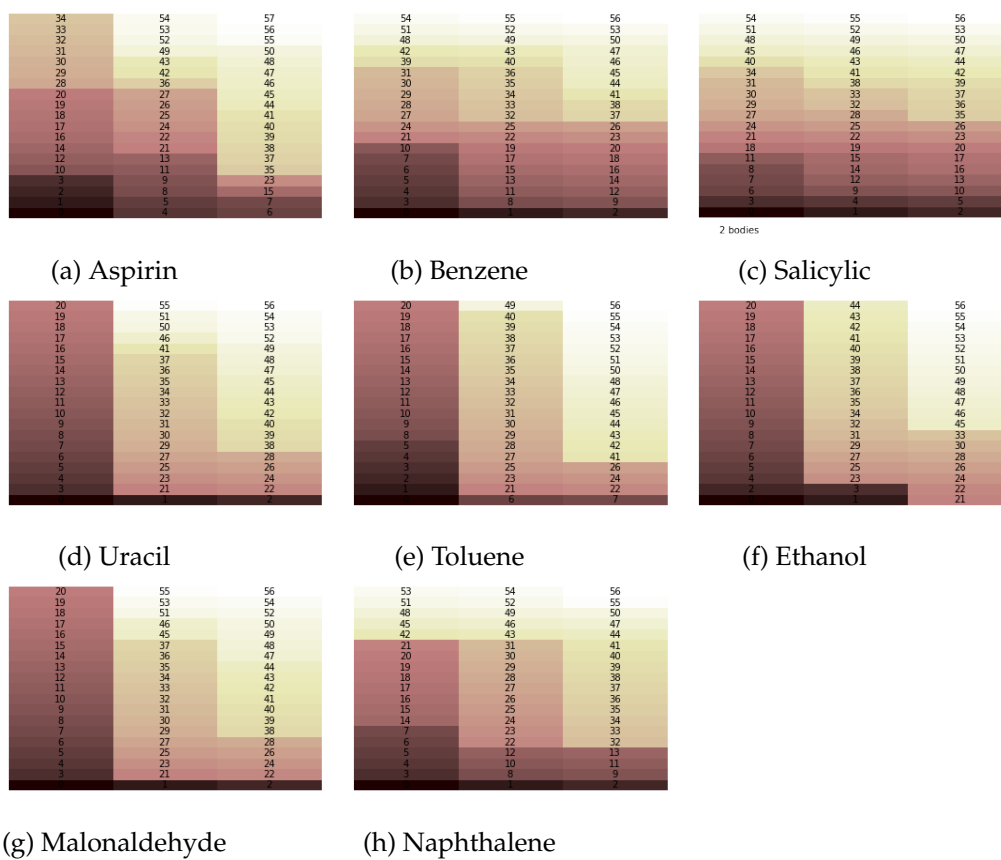


Figure 8.18: An evolution of the chosen indices during an adaptive sparse grid algorithm.

		Benzene		Uracil		Naphthalene		Aspirin	
$V_p$	$k_{\max}$	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
5	2	21.19	3.70	20.59	3.55	23.67	3.78	28.16	4.12
10	2	14.31	3.04	15.52	3.12	20.18	3.55	24.11	3.93
15	2	13.95	2.98	14.42	2.98	20.26	3.55	23.57	3.85
5	3	8.49	1.42	18.02	3.34	26.24	3.98	33.02	4.61
10	3	2.20	1.08	5.34	1.79	7.33	1.99	16.35	3.17
15	3	2.01	1.08	2.77	1.30	6.27	1.71	6.61	2.01
5	4	2.47	0.79	14.07	2.93	27.35	4.18	30.17	4.37
10	4	0.28	0.35	2.36	1.19	15.52	3.08	18.49	3.40
15	4	0.19	0.31	0.59	0.59	1.86	1.06	4.22	1.54
		Ethanol		Malonaldehyd		Salicylic		Toluene	
$V_p$	$k_{\max}$	RMSE	MAE	RMSE	MAE	RMSE	MAE	RMSE	MAE
5	2	17.69	3.31	13.61	2.94	28.54	4.24	19.27	3.49
10	2	16.48	3.21	12.20	2.80	26.59	4.08	15.61	3.07
15	2	9.28	2.38	7.63	2.16	23.25	3.81	14.92	3.00
5	3	15.42	3.08	13.06	2.94	26.49	4.09	16.75	3.27
10	3	10.91	2.58	8.53	2.30	15.56	3.18	3.89	1.54
15	3	2.61	1.24	3.19	1.37	4.92	1.75	2.65	1.27
5	4	14.54	2.98	13.02	2.87	24.73	3.97	17.68	3.37
10	4	8.73	2.26	7.82	2.15	9.41	2.45	4.34	1.59
15	4	2.45	0.94	1.87	1.05	1.67	0.97	1.04	0.75

Figure 8.19: Overview of the achieved error on the root mean squared error (RMSE) and the mean absolute error (MAE) on the energy. Here,  $k_{\max}$  describes the maximal order of incorporated bodies. One sees, that a lower maximal body order than four is in general not sufficient to describe the energy of the incorporated molecules in MD-17.



## Chapter 9

# Conclusion

In this thesis we analysed the penalized least squares regression to approximate the high-dimensional Born-Oppenheimer Potential energy surface. To examine large physical systems and simplify the problem, we assumed the PES to decompose in local atomic neighbourhoods which do not interact. We specified a maximal distance  $r_{\text{cut}} > 0$  under which an interaction of two particles is possible. The many-body decomposition further decomposes those atomic neighbourhoods in lower dimensional parts, i.e. into single nucleus, pairs of nuclei and so on. Based on these locality assumptions, we investigated two approaches in order to optimize the penalized least squared regression.

To cope with the decomposition of the physical space in lower dimensional parts, we also constructed a search space that is decomposed into lower dimensional parts which opens up the possibility of an adaptive sparse grid approach. Here, we decided step by step which body order to incorporate and with which accuracy to approximate each body. In chapter 8 we applied the penalized least squares approach on a data set incorporating tungsten in a periodic environment and on small molecules. Here, we were able to achieve a root-mean squared error of  $0.12 \text{ eV}\text{\AA}^{-1}$  on the energies and of  $0.14 \text{ eV}$  on the forces. On the molecular dynamics data set we were able to achieve a mean-absolute error of chemical accuracy, i.e. of less than  $1 \text{ kcal/mol}$  for all molecules, except aspirin as the largest one. Moreover, we showed that we in the MD-17 data set are able to achieve the same error with up to 3000 degrees of freedom less than in the full grid case.

As a second approach, we considered an optimization regarding the training data rather than the search set. Following [49] we did that by minimizing the models variance. This way, we choose the training data in a way which provides for the most certainty about the approximation. We explained the one to one correspondence to the Fisher information. Minimizing the variance in the penalized least squares model broke down to maximizing the determinant of a matrix. We stated the Cramér-Rao bound and even made it explicit for the case of regularized linear regression. Furthermore we saw that the linear regression estimator even achieves the Cramér-Rao bound and we are actu-

ally directly minimizing our prediction variance when maximizing the Fisher information. In chapter 8 we demonstrated that maximizing the Fisher information is equivalent to choosing the training data set in such a way that it covers as much of the domain as possible. The bigger the random data set the smaller the benefit from active learning got, since it is harder to find new data points exceeding the extrapolation threshold.

What makes active learning especially interesting for the case of approximating the Born-Oppenheimer potential energy surface is that obtaining labeled data is very expensive. The active D-optimality approach we chose does not need labeled data to make the choice of which data points are important for a good estimate. It solely relies on the unlabeled data and can be used as a tool to decide which new data points should be labeled next.

# Bibliography

- [1] Bond lengths and energies. <http://www.science.uwaterloo.ca/~cchieh/cact/c120/bondel.html>. Accessed: 2018-07-25.
- [2] Datasets, a qm/mm resource. <https://qmml.org/datasets.html#solids>. Accessed: 2018-10-03.
- [3] The nobel prize in chemistry 2013. <https://nobelprize.org>. Accessed: 2018-10-18.
- [4] A. P. Bartók, R. Kondor, and G. Csányi. Erratum: On representing chemical environments [phys. rev. b 87, 184115 (2013)]. *Physical Review B*, 96(1):019902, 2017.
- [5] A. P. Bartók, M. C. Payne, R. Kondor, and G. Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical review letters*, 104(13):136403, 2010.
- [6] J. Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of chemical physics*, 134(7):074106, 2011.
- [7] J. Behler. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Physical Chemistry Chemical Physics*, 13(40):17930–17955, 2011.
- [8] R. Bellman. Curse of dimensionality. *Adaptive control processes: a guided tour*. Princeton, NJ, 1961.
- [9] C. Bishop, C. M. Bishop, et al. *Neural networks for pattern recognition*. Oxford university press, 1995.
- [10] R. Biswas and D. Hamann. Interatomic potentials for silicon structural energies. *Physical review letters*, 55(19):2001, 1985.
- [11] R. Biswas and D. Hamann. New classical models for silicon structural energies. *Physical Review B*, 36(12):6434, 1987.
- [12] T. Blank. Tb blank, sd brown, aw calhoun, and dj doren, j. chem. phys. 103, 4129 (1995). *J. Chem. Phys.*, 103:4129, 1995.
- [13] B. Bohn. Error analysis of regularized and unregularized least-squares regression on discretized function spaces. 2016.

- [14] B. Bohn. On the convergence rate of sparse grid least squares regression. 2017.
- [15] D. W. Brenner, O. A. Shenderova, J. A. Harrison, S. J. Stuart, B. Ni, and S. B. Sinnott. A second-generation reactive empirical bond order (rebo) potential energy expression for hydrocarbons. *Journal of Physics: Condensed Matter*, 14(4):783, 2002.
- [16] H.-J. Bungartz and M. Griebel. Sparse grids. *Acta numerica*, 13:147–269, 2004.
- [17] K. Chaloner and I. Verdinelli. Bayesian experimental design: A review. *Statistical Science*, pages 273–304, 1995.
- [18] H. Chen and C. Ortner. Qm/mm methods for crystalline defects. part 1: Locality of the tight binding model. *Multiscale Modeling & Simulation*, 14(1):232–264, 2016.
- [19] S. Chmiela, H. E. Sauceda, K.-R. Müller, and A. Tkatchenko. Towards exact molecular dynamics simulations with machine-learned force fields. *arXiv preprint arXiv:1802.09238*, 2018.
- [20] S. Chmiela, A. Tkatchenko, H. E. Sauceda, I. Poltavsky, K. T. Schütt, and K.-R. Müller. Machine learning of accurate energy-conserving molecular force fields. *Science advances*, 3(5):e1603015, 2017.
- [21] A. Cohen, M. A. Davenport, and D. Leviatan. On the stability and accuracy of least squares approximations. *Foundations of computational mathematics*, 13(5):819–834, 2013.
- [22] A. Cohen and G. Migliorati. Optimal weighted least-squares methods. *arXiv preprint arXiv:1608.00512*, 2016.
- [23] F. Cucker and D. X. Zhou. *Learning theory: an approximation theory viewpoint*, volume 24. Cambridge University Press, 2007.
- [24] M. S. Daw and M. I. Baskes. Embedded-atom method: Derivation and application to impurities, surfaces, and other defects in metals. *Physical Review B*, 29(12):6443, 1984.
- [25] S. Even. *Graph algorithms*. Cambridge University Press, 2011.
- [26] V. V. Fedorov. *Theory of optimal experiments*. Elsevier, 1972.
- [27] P. Flaherty, A. Arkin, and M. I. Jordan. Robust design of biological experiments. In *Advances in neural information processing systems*, pages 363–370, 2006.
- [28] G. Friesecke. The multiconfiguration equations for atoms and molecules: charge quantization and existence of solutions. *Archive for Rational Mechanics and Analysis*, 169(1):35–71, 2003.

- [29] F. Galton. Regression towards mediocrity in hereditary stature. *The Journal of the Anthropological Institute of Great Britain and Ireland*, 15:246–263, 1886.
- [30] C. F. Gauss. *Theoria motus corporum coelestium in sectionibus conicis solem ambientium*, volume 7. Perthes et Besser, 1809.
- [31] T. Gerstner and M. Griebel. Dimension-adaptive tensor-product quadrature. *Computing*, 71(1):65–87, 2003.
- [32] T. Gerstner and M. Griebel. Sparse grids. *Encyclopedia of Quantitative Finance*, 2010.
- [33] M. Griebel, J. Hamaekers, and F. Heber. Bossanova: A bond order dissection approach for efficient electronic structure calculations. *INS Preprint*, 704, 2008.
- [34] M. Griebel, J. Hamaekers, and F. Heber. A bond order dissection anova approach for efficient electronic structure calculations. In *Extraction of Quantifiable Information from Complex Systems*, pages 211–235. Springer, 2014.
- [35] M. Griebel, S. Knapek, and G. Zumbusch. Numerical simulation in molecular dynamics. numerics, algorithms, parallelization, applications, volume 5 of texts in computational science and engineering, 2007.
- [36] M. Griebel, M. Schneider, and C. Zenger. *A combination technique for the solution of sparse grid problems*. Citeseer, 1990.
- [37] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.
- [38] K. Hallatschek. Fouriertransformation auf dünnen gittern mit hierarchischen basen. *Numerische Mathematik*, 63(1):83–97, 1992.
- [39] K. Hansen, F. Biegler, R. Ramakrishnan, W. Pronobis, O. A. Von Lilienfeld, K.-R. Müller, and A. Tkatchenko. Machine learning predictions of molecular properties: Accurate many-body potentials and nonlocality in chemical space. *The journal of physical chemistry letters*, 6(12):2326–2331, 2015.
- [40] K. Hansen, G. Montavon, F. Biegler, S. Fazli, M. Rupp, M. Scheffler, O. A. Von Lilienfeld, A. Tkatchenko, and K.-R. Müller. Assessment and validation of machine learning methods for predicting molecular atomization energies. *Journal of Chemical Theory and Computation*, 9(8):3404–3419, 2013.
- [41] M. Y. Hayes, B. Li, and H. Rabitz. Estimation of molecular properties by high-dimensional model representation. *The Journal of Physical Chemistry A*, 110(1):264–272, 2006.
- [42] M. Hirn, N. Poilvert, and S. Mallat. Quantum energy regression using scattering transforms. *arXiv preprint arXiv:1502.02077*, 2015.

- [43] S. Knapek. Approximation und kompression mit tensorprodukt-multiskalen-approximationsräumen. *Doktorarbeit, Universität Bonn, Institut für Angewandte Mathematik*, 2000.
- [44] W. Kohn. Nobel lecture: Electronic structure of matter—wave functions and density functionals. *Reviews of Modern Physics*, 71(5):1253, 1999.
- [45] A. M. Legendre. *Nouvelles méthodes pour la détermination des orbites des comètes*. F. Didot, 1805.
- [46] D. D. Lewis and W. A. Gale. A sequential algorithm for training text classifiers. In *Proceedings of the 17th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 3–12. Springer-Verlag New York, Inc., 1994.
- [47] G. Montavon, M. Rupp, V. Gobre, A. Vazquez-Mayagoitia, K. Hansen, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld. Machine learning of molecular electronic properties in chemical compound space. *New Journal of Physics*, 15(9):095003, 2013.
- [48] F. Q. Nazar and C. Ortner. Locality of the thomas–fermi–von weizsäcker equations. *Archive for Rational Mechanics and Analysis*, 224(3):817–870, 2017.
- [49] E. V. Podryabinkin and A. V. Shapeev. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science*, 140:171–180, 2017.
- [50] C. E. Rasmussen. Gaussian processes in machine learning. In *Advanced lectures on machine learning*, pages 63–71. Springer, 2004.
- [51] N. Roy and A. McCallum. Toward optimal active learning through monte carlo estimation of error reduction. *ICML, Williamstown*, pages 441–448, 2001.
- [52] M. Rupp, A. Tkatchenko, K.-R. Müller, and O. A. Von Lilienfeld. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters*, 108(5):058301, 2012.
- [53] A. L. Samuel. Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 3(3):210–229, 1959.
- [54] B. Scholkopf and A. J. Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2001.
- [55] E. Schrödinger. An undulatory theory of the mechanics of atoms and molecules. *Physical review*, 28(6):1049, 1926.
- [56] K. T. Schütt, F. Arbabzadah, S. Chmiela, K. R. Müller, and A. Tkatchenko. Quantum-chemical insights from deep tensor neural networks. *Nature communications*, 8:13890, 2017.

- [57] F. Schwabl. *Quantenmechanik*. Springer, 4:183, 2002.
- [58] B. Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.
- [59] H. S. Seung, M. Opper, and H. Sompolinsky. Query by committee. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 287–294. ACM, 1992.
- [60] A. V. Shapeev. Moment tensor potentials: A class of systematically improvable interatomic potentials. *Multiscale Modeling & Simulation*, 14(3):1153–1173, 2016.
- [61] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE transactions on Information Theory*, 44(5):1926–1940, 1998.
- [62] S. Smolyak. Quadrature and interpolation formulas for tensor products of certain classes of functions. In *Soviet Math. Dokl.*, volume 4, pages 240–243, 1963.
- [63] W. J. Szlachta, A. P. Bartók, and G. Csányi. Accuracy and transferability of gaussian approximation potential models for tungsten. *Physical Review B*, 90(10):104108, 2014.
- [64] J. Tersoff. New empirical approach for the structure and energy of covalent systems. *Physical Review B*, 37(12):6991, 1988.
- [65] J. Tersoff. Modeling solid-state chemistry: Interatomic potentials for multicomponent systems. *Physical Review B*, 39(8):5566, 1989.
- [66] R. Todeschini and V. Consonni. *Molecular descriptors for chemoinformatics: volume I: alphabetical listing/volume II: appendices, references*, volume 41. John Wiley & Sons, 2009.
- [67] J. A. Tropp. User-friendly tail bounds for sums of random matrices. *Foundations of computational mathematics*, 12(4):389–434, 2012.
- [68] V. N. Vapnik. An overview of statistical learning theory. *IEEE transactions on neural networks*, 10(5):988–999, 1999.
- [69] M. Yuan, T. T. Cai, et al. A reproducing kernel hilbert space approach to functional linear regression. *The Annals of Statistics*, 38(6):3412–3444, 2010.
- [70] E. Zeidler. *Nonlinear functional analysis and its applications: III: variational methods and optimization*. 2013.
- [71] N. Zettili. *Quantum mechanics: concepts and applications*. John Wiley & Sons, 2009.
- [72] G. M. Zhislin. Discussion of the spectrum of schrödinger operators for systems of many particles. *Trudy Moskovskogo matematicheskogo obschestva*, 9:81–120, 1960.

- [73] X. Zhu, Z. Ghahramani, and J. D. Lafferty. Semi-supervised learning using gaussian fields and harmonic functions. In *Proceedings of the 20th International conference on Machine learning (ICML-03)*, pages 912–919, 2003.