A Learning Hierarchy for Classification and Regression

by

Ganesh Ajjanagadde

S.B., Massachusetts Institute of Technology (2015)

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2016

Author .....
Department of Electrical Engineering and Computer Science
August 20, 2016

Certified by .....
Gregory Wornell
Professor
Thesis Supervisor

Accepted by .....
Christopher Terman
Chairman, Masters of Engineering Thesis Committee

A Learning Hierarchy for Classification and Regression

by

Ganesh Ajjanagadde

Submitted to the Department of Electrical Engineering and Computer Science
on August 20, 2016, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

This thesis explores the problems of learning analysis of variance (ANOVA) decompositions over $\mathbf{GF}(2)$ and $\mathbb{R}$, as well as a general regression setup. For the problem of learning ANOVA decompositions, we obtain fundamental limits in the case of $\mathbf{GF}(2)$ under both sparsity and degree structures. We show how the degree or sparsity level is a useful measure of the complexity of such models, and in particular how the statistical complexity ranges from linear to exponential in the dimension, thus forming a "learning hierarchy". Furthermore, we discuss the problem in both an "adaptive" as well as a "one-shot" setting, where in the adaptive case query choice can depend on the entire past history. Somewhat surprisingly, we show that the "adaptive" setting does not yield significant statistical gains. In the case of $\mathbb{R}$, under query access, we demonstrate an approach that achieves a similar hierarchy of complexity with respect to the dimension.

For the general regression setting, we outline a viewpoint that captures a variety of popular methods based on locality and partitioning of some kind. We demonstrate how "data independent" partitioning may still yield statistically consistent estimators, and illustrate this by a lattice based partitioning approach.

Thesis Supervisor: Gregory Wornell
Title: Professor

Acknowledgments

First, I would like to thank my advisor Prof. Gregory Wornell for his brilliance, intellectual curiosity, and guidance. When I first joined MIT, based off my rather vague interests at the time, my freshman advisor Prof. Marc Baldo had a hunch that Greg would be a great person to work with. I am happy to say that he was right.

It never ceases to amaze me how some of Greg's ideas, so compactly expressed, can be so fruitful. In particular, in one of our very first meetings, Greg expressed the idea of using quantization for classification and regression problems. This idea has remained as an anchor for some of the research explored in this thesis, and I look forward to examining it further.

I also thank Prof. Yury Polyanskiy for providing a wonderful undergraduate research opportunity. It is not an exaggeration to say that most of the information theory I know is from Yury and his notes [27].

I would next like to thank the EECS department as a whole for providing a fantastic learning environment. Indeed, I liked it sufficiently enough as an undergraduate to continue here for doctoral studies, in spite of the Boston winters.

Part of what makes the EECS department wonderful are the people, be they faculty or friends. They are too many to list here, and so I will restrict myself to a few remarks. Prof. John Tsitsiklis's course in probability (6.041) was a fantastic first look at the world of uncertainty, and led to a wonderful first research experience with Prof. Alan Willsky. I had the great fortune of taking multiple courses with Prof. Alexandre Megretski (6.003, 6.241), who showed me the value of intellectual clarity and rigor. Prof. George Verghese is extremely kind and patient, and I thus view him as a wonderful mentor. Prof. Michael Sipser's course on the theory of computation (6.840) was the best I took here till date, and allowed me to explore a topic that I felt was extremely lacking in my EECS education.

Friends are one of the most interesting aspects of life. At MIT, I have had wonderful interactions with Anuran Makur, James Thomas, Govind Ramnarayan, Nirav Bhan, Adam Yedidia, Eren Kizildag, Tarek Lahlou, Tuhin Sarkar, Pranav Kaundinya,

Deepak Narayanan, and Harihar Subramanyam among others. In particular, I am grateful to James Thomas for thoughtful observations on the high level formulations of this thesis.

I would also like to express my gratitude to all of the members of the Signals, Information, and Algorithms (SIA) Laboratory. The group meetings were nice opportunities to get a broader feel for research outside our own core interests. Special thanks goes to Atulya Yellepeddi for being a fantastic co-TA, Gauri Joshi for giving me my first SIA tour, Xuhong Zhang for her interesting perspectives on statistical learning problems and a proof in Chapter 2, Joshua Ka Wing Lee for being a wonderful office-mate, and Tricia Mulcahy O'Donnell for being a fantastic admin for our group.

No human enterprise can succeed in isolation. I am grateful to the free software community for their wonderful tools, and the FFmpeg community in particular for making me understand free software development as a total package, with its highs and lows.

All of the above remarks serve as a mere index for very special kinds of support, and are thus really inadequate in describing the full measure of support and gratitude. Nonetheless, I close by thanking my parents, extended family, and the salt of the earth [30]:

> There are some who are really the salt of the earth in every country and who work for work's sake, who do not care for name, or fame, or even to go to heaven. They work just because good will come of it.

Contents

Chapter 1

Introduction

Statistical learning is a discipline with numerous applications across science and engineering. The work and motivations differ across communities, and thus one encounters a variety of terms: machine learning, data science, deep learning, statistics, inference, etc.

For some idea of the range of applications, consider the following taken from [15]:

1. Diagnosis of patients from various clinical measurements.

2. Prediction of the price of an asset in the future from economic data.

3. Character recognition from handwriting samples.

One can immediately recognize that 1 and 3 are discrete in character, since a diagnosis is either healthy or not, and characters of natural language come from a finite alphabet. On the other hand, 2 is continuous in character, since although monetary instruments have a granularity determined by the lowest denomination, such granularity is conveniently viewed as a infrequent quantization, with all intermediate computations occurring with infinite precision. This is analogous to the distinction between real numbers and their quantized counterparts on computers. The first category are termed *classification* problems and the second *regression* problems.

It should be emphasized that structural assumptions of some kind are essential to formulate well defined problems here. Moreover, assumptions on structure are what

makes learning "interesting", since without such assumptions the problems are either statistically or computationally hard or infeasible. We give some trivial examples of this, to clarify (heuristically) what we mean by the notions of statistical and computational hardness and infeasibility. In these examples and subsequent development, d will denote the dimensionality of the data, n the number of samples, and $\hat{X}$ will denote an estimator of X.

Example 1. (One-time pad, statistical infeasibility) *Consider the class of functions*

$$f_k : \{0,1\}^d \to \{0,1\}^d, \quad f_k(x_1^d) = x_1^d \oplus k_1^d, \quad k_1^d \sim U(\{0,1\}^d),$$

where $\oplus$ denotes bit-wise XOR, U refers to a uniform distribution. Suppose the user sends a message $m \sim U(\{0,1\}^d)$, the adversary observes a ciphertext $y = f(m)$, and forms a guess $\hat{m}$ of the sent message. Then, $\mathbb{P}[\hat{m} = m] = 2^{-d}$, no matter what guessing rule the adversary employs.

Example 2. (Black swan, statistical infeasibility) *Consider a function $f : \{0,1\}^d \to \{0,1\}$, drawn uniformly from the space of all Boolean functions on d-dimensional Boolean vectors. Suppose one observes samples $(x_i, f(x_i)), 1 \leq i \leq n$, where the x_i are drawn uniformly over all the $2^d - 1$ non-zero Boolean vectors. The task is to estimate $f_0 = f(\vec{0})$; call the estimate $\hat{f}_0$. Then regardless of the value of n and choice of $\hat{f}_0$, $\mathbb{P}[\hat{f}_0 = f_0] = 0.5$. Note that as long as the learner is still confined to the non-zero vectors and f is still drawn uniformly, the same conclusion applies even if the learner used a non-uniform distribution over the samples or picked samples. This is a mathematical example of what philosophers call the "black swan problem", otherwise known as "the problem of induction" [24].*

Example 3. (coupon collector over the hypercube, statistical and computational hardness) *We use the same setup as in the previous example, except here x_i are drawn from all the 2^d possible d-dimensional Boolean vectors. Here we add a slight adversarial element: the adversary gets to pick the point x at which $f(x)$ needs to be estimated and does not reveal this choice to the learner. Alternatively, one could*

consider setups where the cost measure is $\sum_{x\in\{0,1\}^d} |f(x) - \hat{f}(x)|$ *and no adversarial element. The core conclusions do not really change; the adversarial element is for ease of exposition. Then, in the case where the learner can freely pick samples, we have:*

$$\mathbb{P}[\hat{f}_x \neq f_x] = \begin{cases} 0.5 & \text{if } n < 2^d, \\ 0 & \text{if } n \geq 2^d. \end{cases}$$

In the case where the learner draws $x_i \sim U(\{0,1\}^d)$ *i.i.d, we may use the results of the classical coupon collector's problem (see e.g. [21, Prop. 2.4]). As such, we get:*

$$\mathbb{P}[\hat{f}_x \neq f_x] = \frac{1}{2} \quad \text{if } n < 2^d$$

and

$$\mathbb{P}[\hat{f}_x \neq f_x] \leq \frac{e^{-c}}{2} \quad \text{if } n > (d\ln(2) + c)2^d.$$

Thus in both setups, an exponential (in d) samples are required for learning accurately.

1.1 Common Structural Assumptions

As seen in the above examples, learning problems in their full generality are hard, and there is thus a need for structural assumptions. We give a brief review of some fruitful structural assumptions here.

1.1.1 Linear Regression

Perhaps the oldest and best understood structure is that of linear regression. For a fascinating historical development of this method and its role in statistics, we refer the reader to [29]. The idea is to restrict the class of predictors to linear (or more precisely affine) models, i.e., given input vectors $X^T = (X_1, X_2, \ldots, X_d)$, the goal is to predict the output Y via $\hat{Y} = \hat{\beta}_0 + \sum_{i=1}^d X_i\hat{\beta}_i$. Note that one can prepend a 1 to X to form $X' = (1, X_1, X_2, \ldots, X_d)$, thus allowing one to restrict study to linear models $\hat{Y} = X'^T\hat{\beta}$.

One of their great advantages is that they are simple to describe and hence score well in terms of "interpretability". They have simple convex optimization problems associated with them for obtaining β from the training data, the most basic approach being least squares. Furthermore, in certain sparse or data limited regimes, they can often do better than fancier methods.

One of the main drawbacks is that the standard setup is ill-suited to nonlinearities. Nevertheless, this can often be overcome by appropriate basis expansions, yielding predictors of the form $\hat{Y} = \sum_{i=1}^{d'} h_i(X)\hat{\beta}_i$ where d' is not necessarily the same as d, and h_i is a suitable basis. Common bases include coordinate projections (the standard linear setup), multivariate polynomials of various degrees, indicators, and splines. Basically, the idea is that although Y is not linear in the original data X, there is some transformed input space (represented by $(h_1(X), h_2(X), \ldots, h_{d'}(X))$) where Y is linear.

1.1.2 ANOVA Decomposition

Analysis of variance (ANOVA) decomposition is a popular method for imposing structure on regression functions. According to [29], there were precursors to this method explored hundreds of years earlier by Laplace, but the initial modern development is primarily due to Fisher, who included it in his classic book [11]. An ANOVA decomposition has the form:

$$f(X_1, X_2, \ldots, X_d) = g_0 + \sum_i g_i(X_i) + \sum_{i<j} g_{ij}(X_i, X_j) + \ldots g_{12\ldots d}(X_1, X_2, \ldots, X_d). \tag{1.1}$$

In (1.1), a variety of structural constraints can be imposed, making it quite attractive. For example, one may choose to eliminate high-order terms as one may not have enough data to justify their inclusion, invoking arguments like Occam's razor, or arguments involving the difficulty of estimating such influences, making any inference of them suspicious at best. This could also be justified in terms of local approximation ideas, since keeping the low degree terms is analogous to using only the first few terms

from a Taylor expansion. Furthermore, this is a very popular idea for exploratory data analysis, a simple illustration of which is estimating dependencies on single variables first (the "main effect" or "first-order effect"), then pairs on the residual formed by subtracting off the single variable terms, and so on. One could also envision a scheme of successive refinement, and one popular approach is the backfitting algorithm for additive models, introduced in [6].

Another area where structure can be imposed is for the choice of basis in basis expansions of the $g_i, g_{ij}, \ldots$ used for the reduction of the problem of learning the model to a linear regression. Lastly, one can impose sparsity constraints on the g functions, by setting a bunch of them to 0. The topic of sparsity in regression is reviewed next.

1.1.3 Sparse Regression

Sparse regression has a rich history, and has appeared in the literature under a variety of terms: compressive sensing, compressed sensing, compressive sampling, and sparse sampling. The literature on this topic is too vast to properly review here. Among many possible sources, we refer the reader to [16] for the usage of these methods in statistics. Nevertheless, we provide an outline of what to expect. The basic idea is that knowledge of the sparsity of a signal can be used to recover it from far fewer measurements than in the unconstrained case. This allows one to solve inverse problems such as recovering $\mathbf{x}$ from $\mathbf{y} = \mathbf{A}\mathbf{x}$ in classically underdetermined situations. This principle is widely applicable and is observed in both noiseless and noisy settings. Indeed, the results in this topic are powerful enough to drive changes in the design of the feature space for $\mathbf{x}$, or the measurement matrix $\mathbf{A}$ in order to enable the harnessing of these techniques effectively. Thus, it is interesting to study its incorporation into the ANOVA and basis expansion framework as an additional level of structure. We explore this in the $\mathbf{GF}(2)$ case in Chapter 2.

1.1.4 Probabilistic Graphical Models

A probabilistic graphical model or simply a graphical model is a probabilistic model where a graph is used to encode the conditional dependence structure between random variables. They are commonly classified into directed graphical models and undirected graphical models, with numerous specialized forms such as factor graphs being occasionally useful. Generally, an increase in the number of edges increases the dependencies in the model, with the most extreme case in the undirected scenario being the fully connected graph. The size of the maximal embedded clique in the graph gives the maximal degree of interaction between the variables. Observed in the log-likelihood domain, the problem of learning a graphical model is an ANOVA decomposition problem with non-negativity and normalization constraints.

1.1.5 Other Structures

There are many other kinds of structures useful for the purposes of learning, such as low-rank assumptions for the task of matrix completion and the emerging topic of sum-product networks [28] as an alternative to standard deep neural network architectures. As the ideas from these kinds of structures have not played a significant role in the problem formulations of this thesis, we do not discuss these interesting topics in greater detail.

1.2 Outline of the Thesis

In general, the primary goal of this thesis is to learn an ANOVA decomposition of the form (1.1). There are a couple of aspects here that to the best of our knowledge have not been explored:

1. We develop theory for classification in a completely Boolean setting, where all additions in (1.1) are over $\mathbf{GF}(2)$. This is motivated by the problem of classification of categorical data, where the data is drawn from a discrete alphabet and the output is a binary label (in the case of binary classification). There is

some prior work on using ANOVA decompositions in the context of representing Boolean functions [26, Sec. 6.2,8.3]. Indeed, the decomposition described in [26, Sec. 6.2] is the one we focus on in this thesis. However, the primary focus in [26] is on real-valued Boolean functions, for which an alternative decomposition is of greater utility.

2. We discuss the setting of learning with chosen examples and learning with feedback, also known as adaptive learning. Learning with chosen examples as opposed to drawing examples uniformly over the space is not unrealistic in applications where there is significant understanding of the feature space. Moreover, it tends to simplify the exposition without changing the essential conclusions, something we illustrated in Example 3. Perhaps more importantly, this allows exploration of the topic of "adaptive learning". Learning with feedback is a topic that is not as well explored as the classical "one-shot" statistical setting. Although there is a danger of creating unreasonable beliefs through reinforcement via feedback, there is no inherent reason to avoid exploring this broadening of the problem setup. Indeed, in [26], this is often the default assumption in the "query access model".

   Note that the problem of adaptive learning as described above is different from a classical online learning setup. A critical difference is that in adaptive learning, the learner is free to pick samples based on the history, while in online learning, the samples are not in control of the learner. Indeed, in this sense the problem is closer to that of reinforcement learning, where the learner is called an "agent" and is free to choose its actions based on the history of rewards and states. Nevertheless, in the adaptive learning setup described above, there is no notion of intermediate rewards associated for every action, where in this case action refers to a sample choice. Instead, the "reward" is measured by some sort of loss function and is computed once all samples have been picked or drawn.

3. We demonstrate how the optimal risk varies depending on the degree of the ANOVA decomposition from constant to exponential in d for the problem over

GF(2). In the case of the problem over $\mathbb{R}$, [19] provides some upper bounds on the excess risk as compared to the optimal value (under l_2 achieved by the conditional expectation) under certain assumptions. We provide a simple upper bound analysis here when the query points can be chosen by the learner. The advantage of our analysis is that it applies to a less restrictive set of assumptions than that of [19], at the cost of needing query access. This gives a "learning hierarchy" of complexity that ranges from constant to exponential in d, as expressed in the title of this thesis.

We discuss the above topics related to the ANOVA decomposition and its learning in Chapter 2 of this thesis. In addition, in Chapter 3, we demonstrate how quantizers can be used for regression. This demonstration is done by using lattice quantizers. Finally, in Chapter 4, we present some ideas for future work.

Chapter 2

Learning of ANOVA Decompositions

2.1 Introduction

In this chapter, we focus primarily on the problem of learning a function $f : \{0,1\}^d \to \{0,1\}$. We also briefly examine the problem of learning a function $f : [0,1]^d \to \mathbb{R}$. As already noted in Example 3, in the absence of structure this requires exponential (in d) samples in the $\mathbf{GF}(2)$ setting. In the real case, the same applies, see e.g. [14, Thm. 3.2]. We first consider the Boolean case by examining the structure imposed by (1.1) adapted to a $\mathbf{GF}(2)$ setting:

$$f(X_1, X_2, \ldots, X_d) = g_0 \oplus \sum_i g_i(X_i) \oplus \sum_{i<j} g_{ij}(X_i, X_j) \oplus \ldots g_{12\ldots d}(X_1, X_2, \ldots, X_d), \tag{2.1}$$

where $\oplus$ is used instead of $+$ to highlight the fact that the addition occurs over $\mathbf{GF}(2)$, and not $\mathbb{R}$. The idea is that all non-linearities of f are captured within the g_i. It is not immediately clear what, if any, connection there is to the classical Fourier expansion of Boolean functions described in [26, 1.2]. To shed light on this, we have the following proposition (also given as [26, Prop. 6.18]).

Proposition 1. *$f_S = AND_S = \prod_{i \in S} x_i$ ranging over $S \subseteq \{1, 2, \ldots, d\}$ serves as a basis for the vector space of $f : \{0,1\}^d \to \{0,1\}$. Note that the "empty" AND_ϕ is defined as the identically 1 function. In other words, there is a unique collection $\mathcal{S}$*

(possibly empty) of distinct sets $S_i \subseteq \{1, 2, \ldots, d\}$, $1 \leq i \leq l$, *such that*

$$f \equiv \sum_{i=1}^{l} AND_{S_i}. \tag{2.2}$$

Proof. The number of $f : \{0,1\}^d \to \{0,1\}$ is 2^{2^d}, and the number of collections $\mathcal{S} = \{S_1, \ldots, S_l\}$ is equal to the cardinality of the power set of the set of all subsets of $\{1, 2, \ldots, d\}$, namely 2^{2^d} as well. It thus suffices to show that the mapping from collections of subsets to Boolean functions is injective. Suppose not, and let $f = \sum_i g_i = \sum_j h_j$ be two distinct representations as a sum of AND_S over a collection of subsets S. Adding, we get $0 = \sum_i g_i + \sum_j h_j$, yielding a non-trivial representation of the identically zero function (call it f_0) of the form:

$$f_0 \equiv \sum_{S_i \in \mathcal{S}} AND_{S_i}. \tag{2.3}$$

For this, the idea is to simply look at the Boolean vectors in increasing order of Hamming weight. More precisely, consider first $0 = f_0(0, \ldots, 0)$. (2.3) immediately gives $\phi \notin \mathcal{S}$. Considering $0 = f_0(1, \ldots, 0) = f_0(0, 1, \ldots, 0) = \cdots = f_0(0, 0, \ldots, 1)$ successively in turn, we see that $\{x\} \notin \mathcal{S}$ for any singleton set $\{x\}$. Repeating the above over all Boolean vectors of Hamming weight 2, we see that $S \notin \mathcal{S}$ for any S with $|S| = 2$. This may clearly be repeated in turn over vectors of Hamming weight 3, 4, and so on up to d, yielding $\mathcal{S} = \phi$. This gives the desired contradiction, implying that the mapping between collections of subsets to Boolean functions is injective. This demonstrates that AND_S forms a basis for the vector space of Boolean functions. □

Proposition 1 gives an idea as to what the ANOVA decomposition (2.1) looks like. For instance,

$$g_{123}(X_1, X_2, X_3) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 X_3 + \beta_4 X_1 X_2 + \beta_5 X_2 X_3 + \beta_6 X_1 X_3 + \beta_7 X_1 X_2 X_3.$$

Furthermore, it shows that there is a close relation between the expansion described here and the one proposed in [26, Thm. 1.1]. The difference between the two lies

chiefly in the fact that the addition in [26, Thm. 1.1] is over the reals, with real coefficients as opposed to addition over $\mathbf{GF}(2)$ here. The approach using $\mathbf{GF}(2)$ exclusively is briefly examined in [26, Sec. 6.2]. The main advantage of [26, Thm. 1.1] is that it gives the flexibility to consider real-valued Boolean functions, i.e. $f : \{0,1\}^d \to \mathbb{R}$ in addition to $f : \{0,1\}^d \to \{0,1\}$. However, this is achieved at the cost of embedding $\{0,1\}$ in $\mathbb{R}$ regardless of the setting. Indeed, it turns out to be convenient in such a setup to consider $f : \{-1,1\}^d \to \{-1,1\}$ via the identification $0 \leftrightarrow -1, 1 \leftrightarrow 1$. We follow this identification here as well in our illustrative comparisons between the two representations.

Two kinds of structure that help with learning Boolean functions efficiently are low degree expansions and sparse expansions. We define these formally below, for the expansion (2.2).

Definition 1. *The **degree** is* $\max(|S_1|, |S_2|, \ldots, |S_l|)$. *We also denote this by* $deg(f) = \max(|S_1|, \ldots, |S_l|)$. *Sometimes, we also use the term **order** as a synonym for **degree**.*

Definition 2. *The **sparsity level** is* l. *We also refer to this by saying that the function* f *is* l*-sparse, or* $sparsity(f) = l$.

We now examine some common Boolean functions and their expansions, with particular emphasis on the degree and sparsity level. We also compare this with their counterparts based on the classical Fourier expansion [26, Thm. 1.1]. We remark that in the classical Fourier expansion, there is a lot more freedom regarding the notion of sparsity, as one can define the notion of "approximately sparse" with real coefficients for the basis expansion in a number of ways. For simplicity, in the examples below, we will list the level of sparsity as the "exact sparsity level", namely the number of non-zero coefficients. In many cases, judging the "approximate" sparsity level of the function may be easily done once the Fourier expansion coefficients have been explicitly computed.

Before proceeding further, we make a simple observation regarding the degree with respect to (2.2). This is also given as [26, Cor. 6.22], left as an exercise there.

Proposition 2. *Consider a function $f : \{0,1\}^d \to \{0,1\}$, where ranging over all possible inputs, the number of output 1's is odd. In other words, $\sum_{x\in\{0,1\}^d} f(x) = 1$. Then the degree of f with respect to (2.2) is d. The converse also holds.*

Proof. The proof is immediate from the observation that

$$\sum_{x\in\{0,1\}^d} AND_S(x_1, x_2, \ldots, x_d) = \mathbb{1}_{S=\{1,2,\ldots,d\}}.$$

This is because there are $2^{d-|S|} \equiv \mathbb{1}_{S=\{1,2,\ldots,d\}} \pmod 2$ Boolean vectors x for which $AND_S(x) = 1$. □

This offers a quick necessary and sufficient condition for checking whether the degree is the maximum possible, namely d.

2.1.1 AND Function

One of the simplest Boolean functions is $AND_{\{1,2,\ldots,d\}}$. With (2.2), the expansion of $AND_{\{1,2,\ldots,d\}}$ it is simply the function itself, yielding a degree of d and a sparsity level of 1. With the Fourier expansion, it is $2\left(\prod_{i=1}^d \frac{1+x_i}{2}\right) - 1$, yielding a degree of d and a sparsity level of 2^d.

2.1.2 Indicator Functions

A slight generalization of $AND_{\{1,2,\ldots,d\}}$ are indicator functions. With (2.2), the expansion of $\mathbb{1}_a(x_1, x_2, \ldots, x_d)$ where $a \in \{0,1\}^d$ is $\prod_{i=1}^d (1 + a_i + x_i)$. This yields a degree of d and a sparsity level of $2^{d-HW(a)}$, where $HW(a)$ denotes the Hamming weight of a. With the Fourier expansion, it is $2\left(\prod_{i=1}^d \frac{1+a_i x_i}{2}\right) - 1$, yielding a degree of d and a sparsity level of 2^d. We shall use the sparsity level expression later in Theorem 5.

2.1.3 OR Function

With (2.2), the expansion of $OR_{\{1,2,\ldots,d\}}$ is $1 + \prod_{i=1}^d (1 + x_i)$, yielding a degree of d and a sparsity level of $2^d - 1$. With the Fourier expansion, it is $1 - 2\left(\prod_{i=1}^d \frac{1-x_i}{2}\right)$,

yielding a degree of d and sparsity level of 2^d.

2.1.4 Complement of a Function

The above expansions bear a lot of similarity to each other, and this is no accident. Basically, the complement of f is $1 \oplus f$ with respect to (2.2). In the Fourier expansion, it is $-f$. From this, we see that the degree and sparsity level of the complement of a function are identical to that of the original function up to a possible ± 1 correction, regardless of the expansion type.

2.1.5 XOR Function

With (2.2), the expansion of $XOR(x_1, x_2, \ldots, x_d)$ is $\sum_{i=1}^{d} x_i$, yielding a degree of 1 and a sparsity level of d. With the Fourier expansion, it is $(-1)^{d+1} \prod_{i=1}^{d} x_i$, yielding a degree of d and a sparsity level of 1.

2.1.6 Majority Function

For simplicity, we examine the case of d odd, as this allows an unambiguous definition of the majority function. For $d > 1$ odd, we define $Maj_d(x_1, x_2, \ldots, x_d) = \mathbb{1}(HW(x) \geq \frac{d+1}{2})$. With the Fourier expansion, a closed form expression for the Fourier expansion weights is given in [26, Thm. 5.19], and asymptotics in [26, 5.11]. The degree is d, and sparsity is 2^{d-1}. For (2.2), the expansion is trickier, with the non-zero coefficient pattern depending on the greatest power of 2 dividing $d-1$. We give the exact expansion only for $d = 2^r + 1$, as this turns out to elicit the worst case in terms of sparsity. More precisely, we have the following proposition.

Proposition 3. *For $d = 2^r + 1$, $Maj_d = \sum_{S: \frac{d+1}{2} \leq |S| < d} AND_S$. Furthermore, for any odd $d > 1$, sparsity$(Maj_d) \leq 2^{d-1} - 1$, with equality precisely when $d = 2^r + 1$ for some $r \geq 1$. In general, deg$(Maj_d) \leq d-1$, and equality occurs for $d = 2^r + 1$, among other cases.*

For this, we first prove the following lemma that generalizes Proposition 2. This is stated as [26, Prop. 6.21] and left as an exercise.

Lemma 1. *Consider* (2.2). *S_i occurs in this expansion precisely when* $\sum_{x:supp(x)\subseteq S_i} f(x) = 1$.

Proof. Consider $\sum_{x:\mathrm{supp}(x)\subseteq S_i} f(x)$. The coefficient of AND_S in this sum is $|\{x : S \subseteq \mathrm{supp}(x) \subseteq S_i\}|$. Thus if $S \nsubseteq S_i$, the coefficient is 0. Furthermore, if $S \subsetneq S_i$, the coefficient is $2^{|S\setminus S_i|} \equiv 0 \pmod 2$. Lastly, if $S = S_i$, the coefficient is 1. This gives the desired result. □

Proof of Proposition 3. By Lemma 1, and the fact that Maj_d is permutation symmetric for all odd d, the occurrence of AND_S depends solely on $|S|$ and not its elements. Consider S with $|S| = k \geq \frac{d+1}{2}$. Then AND_S occurs precisely when $\sum_{i=\frac{d+1}{2}}^{k} \binom{k}{i} \equiv 1 \pmod 2$. But we have:

$$\begin{aligned}\sum_{i=\frac{d+1}{2}}^{k} \binom{k}{i} &\equiv \sum_{i=\frac{d+1}{2}}^{k} \binom{k-1}{i} + \binom{k-1}{i-1} \pmod 2 \\ &\equiv \binom{k-1}{\frac{d-1}{2}} + \binom{k-1}{k} \pmod 2 \\ &\equiv \binom{k-1}{\frac{d-1}{2}} \pmod 2. \end{aligned} \tag{2.4}$$

For analyzing congruence relations of binomial coefficients modulo a prime, the following theorem due to Édouard Lucas [22, Eqn. 137] is useful:

Theorem 1 (Lucas' Theorem)**.** *For non-negative integers m, n and a prime p, we have the congruence relation:*

$$\binom{m}{n} \equiv \prod_{i=0}^{k} \binom{m_i}{n_i} \pmod p,$$

where

$$m = (\overline{m_k m_{k-1} \ldots m_1 m_0})_p = \sum_{i=0}^{k} m_i p^i,$$

and

$$n = (\overline{n_k n_{k-1} \ldots n_1 n_0})_p = \sum_{i=0}^{k} n_i p^i.$$

Remark: For our purposes here this is sufficient. However, to analyze congruence relations of binomial coefficients modulo an arbitrary residue, we may first reduce to the case of prime powers by the Chinese remainder theorem (see e.g. [12, 4.6,4.7] for a simple explanation of this idea). For prime powers, Kummer's theorem provides a partial answer [20]. A more general result in this direction is provided by [13, Thm. 1].

Returning to the proof, we see that it suffices to study (2.4). For convenience, let $s = \frac{d-1}{2}$. Clearly, when $k = d = 2s+1$, $\binom{k-1}{s} = \binom{2s}{s} \equiv 0 \pmod 2$ either via a symmetry argument invoking $\binom{2s}{s+i} = \binom{2s}{s-i}$ or by direct application of Theorem 1. Thus, for $|S| = d$, AND_S won't occur in the expansion, giving the desired degree bound. Furthermore for $k \leq s$, AND_S for $|S| = k$ won't occur. Thus, we see that

$$\begin{aligned}\text{sparsity}(Maj_d) &\leq \sum_{i=s+1}^{2s} \binom{2s+1}{i} \\ &= 2^{d-1} - 1.\end{aligned}$$

Furthermore, equality holds iff $\binom{i}{s} \equiv 1 \pmod 2 \quad \forall s \leq i < 2s$. By examining the binary representation of s, it is clear by Theorem 1 that this happens iff $s = 2^{r-1}$ for some $r \geq 1$. More precisely, let $2^m \parallel s$, where $p^r \parallel n$ denotes $p^r | n, p^{r+1} \nmid n$. In other words, $s = 2^m + 2^{m+1}j$. If $s \neq 2^{r-1}$ for some r, $j \neq 0$. Let $i = 2^{m+2}j$. Then $i - s = 2^{m+1}j - 2^m > 0$, and $i < 2s = 2^{m+1} + 2^{m+2}j$. Then $\binom{i}{s} \equiv 0$ by Theorem 1, because the $(m+1)^{\text{st}}$ least significant bit (LSB) is 1 in s and 0 in i. Conversely, if $s = 2^{r-1}$, any number $s \leq i < 2s$ has the r^{th} LSB set to 1 yielding equality in $\text{sparsity}(Maj_d) = 2^{d-1} - 1$. For such s, $\binom{2s-1}{s} \equiv 1 \pmod 2$, demonstrating that $\deg(Maj_d) = d - 1$ in such cases. □

2.1.7 Mod_3 Function

Define $\text{Mod}_3(x) = \mathbb{1}(HW(x) \equiv 0 \pmod 3)$[1]. [26, Ex. 6.21] gives the Fourier expansion of this function. The degree is d or $d-1$, and sparsity is 2^{d-1}. For (2.2), in

[1]Examination of the representation of this function was suggested by Govind Ramnarayan.

similar fashion to the majority function by permutation symmetry, it suffices to focus on the cardinality of subsets S; call $|S| = k$. Then, AND_S occurs in the expansion precisely when $A_k = \sum_{i=0}^{\lfloor \frac{k}{3} \rfloor} \binom{k}{3i} \equiv 1 \pmod 2$ by Lemma 1. Evaluation of A_k is well known and is easily accomplished by a "roots of unity" trick:

$$\begin{aligned} A_k &= \frac{1}{3}\left((1+1)^k + (1+\omega)^k + (1+\omega^2)^k\right) \\ &= \frac{2^k + (-1)^k(\omega^k + \omega^{2k})}{3}, \end{aligned} \tag{2.5}$$

where $\omega = e^{\frac{2\pi i}{3}}$ denotes a cube root of unity. Since $1 + \omega + \omega^2 = 0$, we have by (2.5), $A_k \equiv 0 \pmod 2 \iff k \equiv 0 \pmod 3$. Thus, the degree is d or $d-1$, and sparsity is $\frac{2^{d+1}}{3} + O(1)$.

2.2 Low Degree Exact Learning of Boolean Functions

The above has given some idea as to what to expect in terms of sparsity and degree for Boolean functions with respect to (2.2). We now develop exact learning algorithms that make use of low degree assumptions and also derive fundamental limits on their performance. Such understanding gives insight into the possibilities and limitations of learning ANOVA decompositions, a main goal of this thesis. Before proceeding, a few clarifications are in order for the problem setup:

1. We focus on the noiseless case, or in other words the problem of exact recovery.

2. We consider the problem when all input points have to be chosen before hand (corresponding to the "one-shot" learning scenario), and one where the input points can be chosen based on the points chosen before and the responses received from a query oracle (corresponding to a "query access", "adaptive learning", or "learning with feedback" setup). The first problem allows one to bound the performance of models where the input points are chosen randomly from a distribution independent of the responses. The second problem allows one to

observe the gains (if any) obtained by breaking down the independence assumption.

2.2.1 One-shot Low Degree Exact Learning

Theorem 2. *Let function f be an element of the set $\mathcal{F}_k = \{f : deg(f) \leq k\}$. The learner picks $\{x_1, x_2, \ldots, x_n\}$, and sends these points to a query oracle which responds with $\{(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_n, f(x_n))\}$. The task of the learner is to identify f via a guess f', and is said to succeed if $f' \equiv f$. Then, in order to guarantee success, $n \geq \sum_{i=0}^{k} \binom{d}{i}$. Equality may be obtained by choosing $\{x_1, x_2, \ldots, x_n\} = \{x : HW(x) \leq k\}$.*

Sufficiency. We show that $\{x_1, x_2, \ldots, x_n\} = \{x : HW(x) \leq k\}$ suffices for reconstruction. Necessity follows from the necessity bound in the adaptive case, provided in Theorem 3. Since $\deg(f) \leq k$, by Lemma 1 it suffices to obtain $\sum_{x:\text{supp}(x)\subseteq S} f(x)$ for $|S| \leq k$. But if $\text{supp}(x) \subseteq S$ and $|S| \leq k$, $HW(x) \leq k$ as well. Thus, obtaining the values of $f(x)$ for $\{x : HW(x) \leq k\}$ suffices for exact identification of f. □

We now examine the asymptotics of $A(d, k) = \sum_{i=0}^{k} \binom{d}{i}$ to get an idea of the statistical complexity of such a low degree reconstruction procedure. This is easy to handle when $k = \tilde{\Omega}(d)$ via the exponential approximation $\binom{d}{k} \doteq 2^{h\left(\frac{k}{d}\right)}$, where $h(x)$ denotes the binary entropy function $h(x) = -x \log_2(x) - (1 - x) \log_2(1 - x)$. In particular, for such k, we get exponential complexity in d. This should not be surprising, since any Boolean function over the first k bits has degree less than or equal to k. Learning such a function (regardless of adaptivity) should take 2^k samples, which is exponential in d. However, for smaller $k = o(d)$, there is a need for better bounds. We have the following:

Proposition 4. $\binom{d}{k} \leq \sum_{i=0}^{k} \binom{d}{i} \leq \binom{d}{k} \frac{d-k+1}{d-2k+1}$.

Proof. The lower bound is trivial. For the upper bound, consider:

$$\begin{aligned}\sum_{i=0}^{k}\binom{d}{i} &= \binom{d}{k}\left(1+\frac{k}{d-k+1}+\frac{k(k-1)}{(d-k+1)(d-k+2)}+\cdots+\frac{k!}{\prod_{j=1}^{k} d-k+j}\right)\\ &\leq \binom{d}{k}\left(1+\frac{k}{d-k+1}+\left(\frac{k}{d-k+1}\right)^2+\dots\right)\\ &= \binom{d}{k}\frac{d-k+1}{d-2k+1}.\end{aligned}$$

□

Note that for k sufficiently small, this gives tighter bounds than Chernoff. For a simple application of Proposition 4, we get $A(d,k) = \Theta(d^k)$ for constant k and $d \to \infty$.

2.2.2 Adaptive Low Degree Exact Learning

Theorem 3. *Let function f be an element of the set $\mathcal{F}_k = \{f : deg(f) \leq k\}$. The learner picks x_1, sends it to a query oracle which responds with $f(x_1)$. It then picks x_2 (which might depend on $(x_1, f(x_1))$), and gets a response $f(x_2)$. The process is repeated until n points have been picked, namely $(x_1, x_2, \dots, x_n)$. Note that the choice of x_i can depend upon $\{(x_1, f(x_1)), (x_2, f(x_2)), \dots, (x_{i-1}, f(x_{i-1}))\}$. The task of the learner is to identify f via a guess f', and is said to succeed if $f' \equiv f$. Then, in order to guarantee success, $n \geq \sum_{i=0}^{k}\binom{d}{i}$. Equality may be obtained by choosing $\{x_1, x_2, \dots, x_n\} = \{x : HW(x) \leq k\}$.*

Proof. Sufficiency is obvious, since those samples sufficed in the one-shot setting by Theorem 2. For necessity, we first observe that $|\mathcal{F}_k| = 2^{A(d,k)}$. Next, we use the decision tree idea used in the algorithms literature. This is used for instance in the proof of a lower bound on sorting via comparisons. For an excellent exposition of this idea, see e.g. [8, 8.1]. The leaves of the decision tree correspond to the elements $f \in \mathcal{F}_k$. The nodes of the tree correspond to the points x_i. The choice of right versus left child reflects the Boolean response $f(x_i)$. The number of samples n corresponds to the depth of the decision tree. Thus, we have $n \geq \log_2 |\mathcal{F}_k| = A(d,k)$ as desired. □

2.3 Sparse Exact Learning of Boolean Functions

Here, we develop exact learning algorithms that make use of sparsity assumptions and obtain fundamental limits on their performance. First, we define sets $\mathcal{G}_l = \{f : \text{sparsity}(f) \leq l\}$, and the goal is to make statements on the lines of Theorems 2 and 3.

2.3.1 One-shot Sparse Exact Learning

Note that the problem of one-shot sparse exact learning can be given a linear algebraic interpretation in order to bring it to a form more reminiscent of the standard compressed sensing setups. We define a $n \times 2^d$ matrix $\mathbf{A}$ whose i^{th} row consists of the vector $[x_i(1), x_i(2), x_i(3), \ldots, x_i(d), x_i(1)x_i(2), x_i(1)x_i(3), \ldots, x_i(d-1)x_i(d), \ldots, \prod_{j=1}^{d} x_i(j)]'$ corresponding to a d-dimensional Boolean vector x_i, where $x_i(j)$ denotes the j^{th} bit of x_i. The outputs are collected in a vector $\mathbf{b} = [f(x_1), f(x_2), \ldots, f(x_n)]'$. Then, solving the equation $\mathbf{Az} = \mathbf{b}$ over $\mathbf{GF}(2)$ is equivalent to the problem of inferring the function f, by Proposition 1. In particular, the problem of sparse exact learning is equivalent to exact sparse recovery of $\mathbf{z}$.

At first glance, one may hope to leverage the work of [9] which obtains compressive sensing methods and bounds for finite fields via a connection to error correcting codes, a connection first noted in [10] and [7, p. 92]. However, in these works the $\mathbf{A}$ matrix had design leeway, allowing one to use a good error correcting code in its design. Here, there are considerable constraints on $\mathbf{A}$ that do not map into conventional constraints on codes. Indeed, it turns out that these constraints are strong enough to rule out efficient methods for one-shot sparse exact learning.

We follow the same strategy as with the low degree exact learning case, first proving an achievability bound for the one-shot case, and then a converse bound for the adaptive case. It is intriguing that these turn out to be essentially equivalent to each other even in the sparsity case, as developed below.

Theorem 4. *Let function f be an element of the set $\mathcal{G}_l = \{f : \text{sparsity}(f) \leq l\}$. The learner picks $\{x_1, x_2, \ldots, x_n\}$, and sends these points to a query oracle which responds with $\{(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_n, f(x_n))\}$. The task of the learner is to identify*

f via a guess f', and is said to succeed if $f' \equiv f$. Then, for $n \geq \sum_{i=0}^{\lfloor \log_2(l) \rfloor + 1} \binom{d}{i}$, there is a reconstruction procedure that guarantees success. Equality may be obtained by choosing $\{x_1, x_2, \ldots, x_n\} = \{x : HW(x) \geq d - 1 - \lfloor \log_2(l) \rfloor\}$.

First, we prove a useful lemma relating sparse recovery to null-space constraints.

Lemma 2. *Consider the problem of exact sparse recovery over the class $\mathcal{G}_l$, $l \geq 1$. Then to guarantee successful recovery over the entire class, $HW(\mathbf{z}) \geq 2l + 1$ for all $\{\mathbf{z} \neq \mathbf{0} : \mathbf{Az} = \mathbf{0}\}$. Conversely, if $HW(\mathbf{z}) \geq 2l + 1$ for all $\{\mathbf{z} \neq \mathbf{0} : \mathbf{Az} = \mathbf{0}\}$, we can ensure successful recovery.*

Proof. For the forward direction, suppose not, and let $\mathbf{Az} = \mathbf{0}$ for $\mathbf{z} = [j\vec{1}', (n-j)\vec{0}']'$ where $1 \leq j \leq 2l$. The assumption of the support being the first j entries may be made without loss of generality by reordering the rows of $\mathbf{A}$. Then, for $\mathbf{x} = [l\vec{1}', (n-l)\vec{0}']'$, $\mathbf{Ax} = \mathbf{A}(\mathbf{x} + \mathbf{z})$. Moreover, sparsity$(\mathbf{x}) = l$, and sparsity$(\mathbf{x} + \mathbf{z}) \leq l$ as well, contradicting the hypothesis of exact sparse recovery. For the reverse direction, if $HW(\mathbf{x}) \leq l$ and $HW(\mathbf{z}) \geq 2l + 1$, we have $HW(\mathbf{x} + \mathbf{z}) \geq 2l + 1 - l = l + 1$. Thus there can be only one l-sparse $\mathbf{x}$ satisfying $\mathbf{Ax} = \mathbf{b}$ in such a scenario, and it may be recovered combinatorially by iterating over all $\{\mathbf{x} : HW(\mathbf{x}) \leq l\}$ and checking whether $\mathbf{Ax} = \mathbf{b}$ or not. □

We now propose what we call a "high Hamming weight" reconstruction scheme. This scheme uses all $\mathcal{T}_k = \{x : HW(x) \geq d - k\}$ for some $k \geq 0$. Obviously if one chooses $k = d$, then regardless of l, we can guarantee exact recovery. But this uses an exponentially large (in d) number of samples. We now give a much tighter dependence of k on l.

We prove a lemma which reduces such a tightening to a combinatorial set cover type of question.

Lemma 3. *We say that $\mathcal{T}$ covers $\mathcal{S}$ precisely when $\mathcal{S} \subseteq \mathcal{T}$. Let k be a number such that for any collection of $j \leq 2l$ subsets $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_j \subseteq \{1, 2, \ldots, d\}$, there is a $\mathcal{T} \subseteq \{1, 2, \ldots, d\}$ with $|\mathcal{T}| \geq d - k$ such that $\mathcal{T}$ covers an odd number of the $\mathcal{S}_i$. For such a k (which depends on d, l), via a high Hamming weight reconstruction corresponding to this k, we can recover exactly any $f \in \mathcal{G}_l$.*

Proof. By Lemma 2, it suffices to show that for any $\mathbf{z} \in \{\mathbf{z} \neq \mathbf{0} : \mathbf{Az} = \mathbf{0}\}$, $HW(\mathbf{z}) \geq 2l+1$. Here $\mathbf{A}$ corresponds to the high Hamming weight reconstruction that uses $\mathcal{T}_k$. Suppose not, and assume $\mathbf{Az} = \mathbf{0}, \mathbf{z} \neq \mathbf{0}$ and $HW(\mathbf{z}) \leq 2l$. Equivalently, there is a f (corresponding to $\mathbf{z}$) with sparsity$(f) \leq 2l$ and $f(t) = 0$ for all $t \in \mathcal{T}_k$. Let $\mathcal{S}_i$ over $1 \leq i \leq HW(\mathbf{z})$ be the elements in the expansion of f via (2.2). Then by hypothesis there is a $\mathcal{T}$ with $|\mathcal{T}| \geq d-k$ covering an odd number of the $\mathcal{S}_i$. Let $\mathbf{a_i}'$ be the row of $\mathbf{A}$ corresponding to the indicator of $\mathcal{T}$. Thus by the hypothesis and (2.2), we have $f(\mathbf{a_i}) = \mathbf{a_i}'\mathbf{z} = 1$. This contradicts $\mathbf{Az} = \mathbf{0}$. □

The next lemma addresses the expression of k in terms of d and l, thereby addressing the combinatorial question raised in Lemma 3.

Lemma 4. *For any collection $\mathcal{S}$ of $l \geq 1$ distinct subsets $\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_l \subseteq \{1, 2, \ldots, d\}$, there is a $\mathcal{T} \subseteq \{1, 2, \ldots, d\}$ with $|\mathcal{T}| \geq d - \lfloor \log_2(l) \rfloor$ such that $\mathcal{T}$ covers an odd number of the $\mathcal{S}_i$. We call such a $\mathcal{T}$ an odd covering set.*

Proof. [2] We prove by induction on d. For $d = 1$, the cases $l = 1$ and $l = 2$ are obvious. Suppose $d > 1$, and assume the result holds for $1, 2, \ldots, d-1$. Suppose an element i is contained in all of the l subsets. Then we may appeal to the induction hypothesis, getting an odd covering set $\mathcal{T}'$ with $|\mathcal{T}'| \geq d - 1 - \lfloor \log_2(l) \rfloor$ on $\{1, 2, \ldots, d\} \setminus \{i\}$. Adding $\{i\}$ to $\mathcal{T}'$, we get the desired $\mathcal{T}$. Likewise, if there is an element i not contained in any of the l subsets, we may follow the same procedure. Thus, without loss of generality, we may assume each element of $\{1, 2, \ldots, d\}$ appears in at most $l-1$ of the $\mathcal{S}_i$ and at least 1 of them. We may also assume without loss that l is even, since otherwise we may simply use $\mathcal{T} = \{1, 2, \ldots, d\}$. We prove by contradiction, and hence suppose that there is a collection $\mathcal{S} = \{\mathcal{S}_1, \mathcal{S}_2, \ldots, \mathcal{S}_l\}$ with $|\mathcal{S}| = l$ such that for all $\mathcal{T}, |\mathcal{T}| \geq d - \lfloor \log_2(l) \rfloor, \mathcal{T}$ covers an even number of the $\mathcal{S}_i$.

We now view the setting through the complement $\mathcal{T}^{\mathsf{c}}$ as this turns out to simplify the argument. Specifically, we first introduce the notion of "influencing": a set $\mathcal{T}$ is

[2] Proof due to Xuhong Zhang.

said to influence a set $\mathcal{S}$ iff $S \cap T \neq \phi$. Observe that:

$$\mathcal{T} \text{ covers } \mathcal{S} \Leftrightarrow \mathcal{T}^{\mathsf{c}} \text{ does not influence } \mathcal{S}. \tag{2.6}$$

Using (2.6) and that l is even, we see that:

$$\begin{aligned} &\forall \mathcal{T}, \quad |\mathcal{T}| \geq d-k, \quad \mathcal{T} \text{ covers an even number of the } \mathcal{S}_i \Leftrightarrow \\ &\forall \mathcal{T}^{\mathsf{c}}, \quad |\mathcal{T}^{\mathsf{c}}| \leq k, \quad \mathcal{T}^{\mathsf{c}} \text{ influences an even number of the } \mathcal{S}_i. \end{aligned} \tag{2.7}$$

By the inclusion-exclusion principle, we have:

$$\begin{aligned} &\forall \mathcal{T}, \quad |\mathcal{T}| \leq j, \quad \mathcal{T} \text{ influences an even number of the } \mathcal{S}_i \Rightarrow \\ &\forall \mathcal{T}, \quad |\mathcal{T}| \leq j, \quad |\{\mathcal{S}_i \in \mathcal{S} : \mathcal{T} \subseteq \mathcal{S}_i\}| \text{ is even.} \end{aligned} \tag{2.8}$$

Indeed, consider $\{x_1, x_2, \ldots, x_j\}$. Define sets

$$\mathcal{A}_i = \{\mathcal{S}_j \in \mathcal{S} : x_i \in \mathcal{S}_j\}.$$

Then, since all $\mathcal{T}$ with $|\mathcal{T}| \leq j$ influence an even number of the $\mathcal{S}_i$, we have

$$|\mathcal{A}_i|, |\mathcal{A}_i \cup \mathcal{A}_j|, \ldots, |\bigcup_{i=1}^{j} \mathcal{A}_i|$$

are all even. By inclusion exclusion, we get that

$$|\mathcal{A}_i|, |\mathcal{A}_i \cap \mathcal{A}_j|, \ldots, |\bigcap_{i=1}^{j} \mathcal{A}_i|$$

are all even, yielding (2.8). In fact, the above argument yields more. Since the universe where the $\mathcal{A}_i$ live is $\mathcal{S}$, and $|\mathcal{S}|$ is even, the cardinality of any of the regions formed by the Venn diagram of the $\mathcal{A}_i$ for $1 \leq i \leq j$ [3] is also even. In particular, we

[3]For the reader familiar with the term, formally this is the algebra generated by $\mathcal{A}_i$.

have:

$$\begin{aligned}&\forall \mathcal{T}, \quad |\mathcal{T}| \leq j, \quad \mathcal{T} \text{ influences an even number of the } \mathcal{S}_i \Rightarrow \\ &\forall \text{ partitions } \mathcal{T} = \mathcal{A} \cup \mathcal{B}, \mathcal{A} \cap \mathcal{B} = \phi, |\{\mathcal{S}_i \in \mathcal{S} : \mathcal{A} \subseteq \mathcal{S}_i, \mathcal{B} \cap \mathcal{S}_i = \phi\}| \text{ is even.}\end{aligned} \tag{2.9}$$

Now consider a tree construction procedure as follows. We start with a root node consisting of all l elements of $\mathcal{S}$. At each node containing more than one set of the collection $\mathcal{S}$, we can find an element i such that at least one of the sets at this node contains i, and at least one does not. This may be done since the $\mathcal{S}_i$ are distinct. We then split the node into two child nodes, the left child corresponding to the sets containing i and the right child corresponding to the sets lacking i. We grow the tree in this fashion. The path to a node at depth j corresponds to a $\mathcal{T} = \{x_1, x_2, \ldots, x_j\} = \mathcal{A} \cup \mathcal{B}$. Here, $\mathcal{A}$ corresponds to the left children and $\mathcal{B}$ corresponds to the right children. By (2.9) and our hypothesis, a node at depth $j \leq \lfloor \log_2(l) \rfloor$ has an even number of elements. Thus, the tree may be grown till a depth of $\lfloor \log_2(l) \rfloor$, and nodes at that depth have at least 2 sets. Thus, we have:

$$\begin{aligned}2^{\lfloor \log_2(l) \rfloor} 2 &\leq l \\ \Rightarrow \lfloor \log_2(l) \rfloor &\leq \log_2(l) - 1,\end{aligned}$$

a contradiction to $\lfloor x \rfloor > x - 1$. This completes the proof of Lemma 4. □

We may now prove Theorem 4.

Proof of Theorem 4. By Lemmas 3 and 4, we see that for $\mathcal{G}_l$, we have a high Hamming weight reconstruction for $k = \lfloor \log_2(2l) \rfloor = \lfloor \log_2(l) \rfloor + 1$. This immediately gives the result. □

2.3.2 Adaptive Sparse Exact Learning

We first give a very simple lower bound on the number of samples required with quite a large gap from the achievability Theorem 4. Essentially, we mimic the proof of Theorem 3.

Proposition 5. *Let function f be an element of the set $\mathcal{G}_l = \{f : sparsity(f) \leq l\}$. The learner picks x_1, sends it to a query oracle which responds with $f(x_1)$. It then picks x_2 (which might depend on $(x_1, f(x_1))$), and gets a response $f(x_2)$. The process is repeated until n points have been picked, namely $(x_1, x_2, \ldots, x_n)$. Note that the choice of x_i can depend upon $\{(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_{i-1}, f(x_{i-1}))\}$. The task of the learner is to identify f via a guess f', and is said to succeed if $f' \equiv f$. Then, in order to guarantee success, $n \geq \log_2 \left(\sum_{i=0}^{l} \binom{2^d}{i} \right)$.*

Proof. We know that $|\mathcal{G}_l| = \sum_{i=0}^{l} \binom{2^d}{i}$. As in Theorem 3, we construct a decision tree. The leaves of the tree correspond to the elements $f \in \mathcal{G}_l$. The nodes of the tree correspond to the points x_i. The choice of right versus left child reflects the Boolean response $f(x_i)$. The number of samples n corresponds to the depth of the decision tree. Thus, $n \geq \log_2 |\mathcal{G}_l|$ as desired. □

To see the gap between the lower bound on n of the above Proposition 5 and the upper bound from Theorem 4, consider l constant, and $d \to \infty$. Then, the lower bound of Proposition 5 is $\Theta(dl) = \Theta(d)$ via Proposition 4, while the upper bound from Theorem 4 is $\Theta(d^{\log_2(l)})$, a difference between linear and polynomial complexity.

It is thus desirable to improve upon the above crude bound. The intuition is that the bound assumes "optimistic" equal (or constant ratio if one only cares about asymptotics) splitting of $\frac{j}{2} : \frac{j}{2}$ where j denotes the number of functions corresponding to a node in the decision tree. This may not be a realistic assumption, and in the worst case of $1 : j-1$ splitting, the number of nodes in a tree of depth k changes from exponential in k to linear in k. The question now is how to elicit such a worst case at every stage of the decision tree. The answer is provided in the following improved lower bound, which essentially matches the upper bound of Theorem 4.

Theorem 5. *We use the same adaptive learning setup as in Proposition 5. Then, in order to guarantee success, $n \geq \sum_{i=0}^{\lfloor \log_2(l) \rfloor} \binom{d}{i}$.*

Proof. Consider the class of functions $\mathcal{F} = \{\mathbb{1}_x : HW(x) \geq d - \lfloor \log_2(l) \rfloor\}$. By the discussion in 2.1.2, we know that for all $f \in \mathcal{F}$, $\text{sparsity}(f) \leq 2^{\lfloor \log_2(l) \rfloor} \leq l$. Furthermore, at any stage of the decision tree, querying any point x can at best do a

$1 : j-1$ split, where j is the number of functions at the current node in the decision tree. Thus, $n \geq |\mathcal{F}| = \sum_{i=0}^{\lfloor \log_2(l) \rfloor} \binom{d}{i}$. □

2.4 Low Degree Learning of Real Functions

The theorems developed in the previous sections give a useful summary of the problem of noiseless learning of Boolean functions with respect to the ANOVA decomposition (2.2). We now turn to a similar question regarding the learning of functions $f : [0,1]^d \to \mathbb{R}$. We note that this problem has been very recently independently proposed and examined in [19]. The approach developed there is named SALSA (**S**hrunk **A**dditive **L**east **S**quares **A**pproximation). The primary focus in [19] is on the case where f has a suitable kernel decomposition in a reproducing kernel Hilbert space (RKHS). This allows the use of nice, computationally efficient, linear algebraic techniques for the solution of the regression problem. Furthermore, [19] utilizes product kernels $k(x, x') = \prod_{i=1}^{j} k_i(x(i), x'(i))$ and their symmetric counterparts over other j element subsets of $\{1, 2, \ldots, d\}$, corresponding to a j^{th} degree ANOVA decomposition. This reduces the task of computing the kernel matrix to that of evaluating symmetric polynomials up to j^{th} order among d variables. This kernel structure allows the employment of the Girard-Newton identities (see e.g. [23, Eqn. 2.11']) to reduce a naive $O\left(j\binom{d}{j}\right)$ to a $O(j^2 d)$ complexity for evaluation of a single entry of the matrix.

However, there are numerous limitations of this work and thus opportunities for further exploration, some of which have been noted by the authors themselves. In this section, we partially address one such direction as described below. The main theorems of [19] deal with the case where the true f ($f(x) = E[Y|X = x]$ in the standard l_2 setup) is drawn from a class with finite norm with respect to the same RKHS as used for the regression, among other more technical assumptions detailed in their work. This may be unreasonable, and so the authors present another theorem for dealing with violations to this assumption. Nevertheless, this theorem is unsatisfactory in that in general it can fall back to exponential complexity in d from the general hierarchy captured by $\binom{d}{j}$ which they obtain in the more restrictive setting.

In spite of this, the practical performance of this method is very good, as substantiated by a quite thorough examination of synthetic as well as real data from a variety of sources. This may suggest that the $\binom{d}{j}$ hierarchy holds even when the restrictive "same-kernel" assumptions are lifted.

Here, we offer a glimpse of this possibility in a chosen query model. The idea is that if the function f has a low degree decomposition of degree k, one can reconstruct f on a scaled cubic lattice via querying points with number of nonzero components going from 0 to k. Reproducing f on the scaled cubic lattice is sufficient for reconstruction of f under most reasonable locality/smoothness assumptions due to the equivalence of norms on finite dimensional vector spaces (here $[0,1]^d \subseteq \mathbb{R}^n$). Normally, such a procedure requires querying f at an exponential number of points in order to cover the lattice. However, here we can obtain the values at all points of the scaled cubic lattice via a reconstruction procedure from the points containing a small (0 to k) number of nonzero components. Thus, we effectively obtain a reduction from 2^d to $A(d,k) = \sum_{i=0}^{k} \binom{d}{i}$ in terms of statistical complexity.

We develop these ideas more rigorously below. First, we state and prove the core reconstruction lemma.

Lemma 5. *Let $f : [0,1]^d \to \mathbb{R}$ have a k^{th} order decomposition:*

$$f(x_1, x_2, \ldots, x_d) = \sum_{i=1}^{\binom{d}{k}} f_i(x_{i_1}, x_{i_2}, \ldots, x_{i_k}). \tag{2.10}$$

Then, we have the following identity for such f:

$$f(x_1, x_2, \ldots, x_d) = \sum_{i=0}^{k} (-1)^i \binom{d-k-1+i}{i} \sum_{sym} f(x_1, x_2, \ldots, x_{k-i}, 0, 0, \ldots, 0). \tag{2.11}$$

Here, $\sum_{sym}$ is a shorthand for symmetric summation, e.g. for $d = 4$:

$$\sum_{sym} f(x_1, 0, 0, 0) = f(x_1, 0, 0, 0) + f(0, x_2, 0, 0) + f(0, 0, x_3, 0) + f(0, 0, 0, x_4),$$

and

$$\sum_{sym} f(x_1, x_2, 0, 0) = f(x_1, x_2, 0, 0) + f(x_1, 0, x_3, 0) + f(x_1, 0, 0, x_4) + f(0, x_2, x_3, 0) + f(0, x_2, 0, x_4) + f(0, 0, x_3, x_4).$$

Proof. By the symmetry of (2.11), it suffices to focus on the coefficient of f_1 of (2.10) on the left and right hand sides. The coefficient of $f_1(x_1, \ldots, x_k)$ on both the left and right hand sides of (2.11) is 1 since $(-1)^0\binom{d-k-1}{0} = 1$. Out of all vectors x passed into f_1 with i zeros for $i \geq 1$ and x_j's populating the rest, again by symmetry it suffices to examine the coefficient of $f_1(x_1, \ldots, x_{k-i}, 0, 0, \ldots, 0)$ on the left and right hand side. On the left hand side, it is clearly 0. On the right hand side, the number of terms in the inner symmetric summation for the jth index is $\binom{d-k}{i-j}$. This is because the first k entries are fixed to $(x_1, x_2, \ldots, x_{k-i}, 0, 0, \ldots, 0)$, and we need to pick out of the remaining $d-k$ positions $i-j$ of them to populate with the respective x_l's in order to get a vector with $k-j$ entries populated by x_l's, the rest populated by zeros. Thus, the coefficient on the right hand side is:

$$\sum_{j=0}^{i}(-1)^j\binom{d-k-1+j}{j}\binom{d-k}{i-j} = \sum_{j=0}^{i}\binom{k-d}{j}\binom{d-k}{i-j} \tag{2.12}$$

$$= \binom{0}{i} \tag{2.13}$$

$$= \mathbb{1}(i=0). \tag{2.14}$$

Here we have used the extended definition of binomial coefficients: $\binom{r}{k} = \frac{\prod_{i=0}^{k-1} r-i}{k!}$, valid for any $r \in \mathbb{R}$, $k \geq 0 \in \mathbb{N}$. For a complete definition of binomial coefficients, see e.g. [12, Eqn. 5.1]. (2.12) follows from a standard identity $\binom{r}{k} = (-1)^k\binom{k-r-1}{k}$ (see e.g. [12, Eqn. 5.14]). (2.13) follows from the Vandermonde convolution $\sum_k \binom{r}{k}\binom{s}{n-k} = \binom{r+s}{n}$ (see e.g. [12, Eqn. 5.27]). Note that this proof handles $i = 0$, i.e. the coefficient of $f_1(x_1, \ldots, x_k)$ as well, which was treated as a special case at the beginning of the proof. Nevertheless, we feel that there is value in presenting the special case separately as it allows a good sanity check on the more complicated binomial coefficient

summation (2.12). □

We now prove a general lemma regarding the usage of a cubic lattice for learning of $f : [0,1]^d \to \mathbb{R}$ under a query access model.

Lemma 6. *Suppose $f : [0,1]^d \to \mathbb{R}$ belongs to the class of 1-Lipschitz functions with respect to the 2-norm, with Lipschitz constant M. In other words, $|f(x) - f(y)| \leq M||x-y||_2 \quad \forall x, y \in [0,1]^d$. Let $\mathcal{F}$ denote the set of all such 1-Lipschitz functions with Lipschitz constant M. Suppose that a learner obtains the values of $f(x)$ over all x belonging to a "s-scaled" cubic lattice: $\{x : x = \frac{1}{s}(x'_1, x'_2, \ldots, x'_d), 0 \leq x'_i \leq s\}$, and its goal is to estimate f via an estimator $\hat{f}$. Then, $\hat{f}$ may be chosen so that $||f - \hat{f}||_\infty \leq \frac{M\sqrt{d}}{2s}$.*

Proof. The proof simply follows by choosing $\hat{f}$ to be a nearest neighbor decoder. Maximum distance to the nearest neighbor (NN) in the lattice is $\frac{\sqrt{d}}{2s}$. Thus, for any $x \in [0,1]^d$,

$$\begin{aligned} |f(x) - \hat{f}(x)| &\leq |f(x) - f(NN(x))| + |\hat{f}(x) - f(NN(x))| \\ &\leq M|x - NN(x)| + 0 \\ &\leq \frac{M\sqrt{d}}{2s}. \end{aligned}$$

□

Combining Lemma 5 and Lemma 6, we easily get the following proposition:

Proposition 6. *Suppose $f : [0,1]^d \to \mathbb{R}$ belongs to the class of 1-Lipschitz functions with respect to the 2-norm, with Lipschitz constant M. Suppose f also has a k^{th} order decomposition 2.10. Suppose a learner picks $\{x_1, x_2, \ldots, x_n\}$, and sends these points to a query oracle which responds with $\{(x_1, f(x_1)), (x_2, f(x_2)), \ldots, (x_n, f(x_n))\}$. Let the learner's estimate of f be $\hat{f}$. Then, for $A(d,k,s) = \sum_{i=0}^{k} \binom{d}{i} s^i$, $n \geq A(d,k,s)$, the learner can respond with a $\hat{f}$ satisfying:*

$$||f - \hat{f}||_\infty \leq \frac{M\sqrt{d}}{2s}.$$

Proof. The learner can simply pick the x_i to be the points of the s-scaled cubic lattice with number of non-zeros not exceeding k. Then, by Lemma 5, it can obtain the values of $f(x)$ over all x belonging to the s-scaled cubic lattice. Finally, by applying Lemma 6, we get the desired result. □

In order to understand the asymptotics of the number of samples n versus the tolerance level (upper bound on $||f-\hat{f}||_\infty$), it is helpful to obtain bounds on $A(d,k,s)$. These may be easily obtained via the same argument used for Proposition 4, and are given in the following proposition:

Proposition 7. $s^k\binom{d}{k} \leq \sum_{i=0}^{k}\binom{d}{i}s^i \leq s^{k+1}\binom{d}{k}\frac{(d-k+1)}{s(d-k+1)-k}$.

Proof. The lower bound is trivial. For the upper bound, consider:

$$\begin{aligned}
\sum_{i=0}^{k}\binom{d}{i}s^i &= s^k\binom{d}{k}\left(1+\frac{k}{s(d-k+1)}+\frac{k(k-1)}{s^2(d-k+1)(d-k+2)}+\cdots+\frac{k!}{s^k\prod_{j=1}^{k}d-k+j}\right)\\
&\leq s^k\binom{d}{k}\left(1+\frac{k}{s(d-k+1)}+\left(\frac{k}{s(d-k+1)}\right)^2+\dots\right)\\
&= s^{k+1}\binom{d}{k}\frac{(d-k+1)}{s(d-k+1)-k}.
\end{aligned}$$

□

2.5 Conclusions

1. We characterized the fundamental limits of exact learning in a noiseless Boolean setting with respect to an ANOVA inspired decomposition. In order to precisely define questions of interest, we focused on sparsity and low degree assumptions. We also explored the question of gains via an adaptive model, where a learner gets to pick samples based on the query-response history. At a high level, the statistical complexity of learning a k degree Boolean function over $\{0,1\}^d$ is $O(d^k)$, while that of learning an l sparse Boolean function over $\{0,1\}^d$ is $O(d^{\log_2(l)+1})$. At some level, this is a pleasing result since a degree k function is necessarily at most 2^k sparse. Thus we have shown that enlarging to the

set of sparse functions from the set of low degree functions does not incur a greater than $O(d)$ multiplicative penalty factor. This is reminiscent of the classical $O(k \log(\frac{n}{k}))$ type of results in classical compressed sensing, where the multiplicative penalty factor is $O(\log(n)) = O(\log(2^d)) = O(d)$ in our setup. This reinforces the current wisdom of the literature that sparsity is a very useful notion for good learning properties.

2. We also showed that, somewhat surprisingly, allowing adaptive learning does not improve upon the statistical complexity, beyond a possible $O(d)$ multiplicative factor in the sparse case. Note that the setup discussed here allows access to the complete history. In the case of the popular topic of streaming algorithms, first studied in [25] and popularized as well as formalized in [1], there is access to only a fraction of the history due to storage constraints. In that sense, the above represents the most optimistic scenario for adaptive learning.

   Nevertheless, we still believe that the adaptive learning paradigm is a useful one, and deserves further exploration. For instance, the results of this chapter do not tackle the issues of computational complexity. In particular, it is possible that allowing adaptive learning might allow greater computational efficiency. This would be especially interesting, since a common use case of adaptive learning is in real-time (hard or soft) applications. Consider for instance a robot wishing to explore its environment. In many cases, it will be free to position and orient its sensors in specific locations (a query access model), and can choose future queries based on the past and its responses. However, it might also need to make decisions in real-time, and thus developing fast algorithms for adaptive learning in such settings could be useful. Our primary inspiration is from the benefits of sequential hypothesis testing over the classical binary hypothesis testing setup, compare for instance [27, Thm. 13.3] versus [27, Thm. 13.2]. Thus, we are also interested in understanding other formulations for which statistical complexity gains are possible, and translating such statistical gains into computational gains.

3. The results of this chapter primarily deal with the noiseless case. Indeed, the Boolean case is particularly appealing for this purpose in that it allows one to formulate clean, exact learning problems that can serve as a guideline on what to expect in other spaces. We believe that by analogy with the results in the literature for the real case, it should be possible to achieve robustness against noise of various kinds with minimal loss of complexity.

4. We provided a glimpse into the possibilities of ANOVA learning in the real case, under a query access model in a noiseless setting in Proposition 6. In particular, we obtain sub-exponential statistical complexity for the problem, compared with exponential complexity in the unstructured case. We remark that under the standard assumption of a uniform bound on the conditional variance $Var[Y|X = x] \leq \sigma^2$, one may simply repeat the queries and use Chebyshev's inequality to obtain sub-exponential statistical complexity for the problem of estimating $E[Y|X = x]$. We omitted this analysis as it is unlikely that this is an order-optimal strategy for the estimation of $E[Y|X = x]$. Furthermore, there are other more serious limitations of Proposition 6, such as the assumption of query access being provided. As such, we leave such developments that extend the work of [19] and this section to future work.

5. On a more technical front, there is a gap ($O(d)$ multiplicative) between the converse Theorem 5 for the adaptive case and the achievability for the one-shot setting in Theorem 4. However, when confined to the one-shot setting, the achievability bound in Theorem 4 is in fact tight, with a matching converse. The astute reader may notice that this may be shown by using the proof technique of the adaptive learning converse, and making use of the $2l$ versus l discrepancy.

Chapter 3

Usage of Lattice Quantizers for Regression

3.1 Introduction

In this chapter, we discuss an approach towards regression based on the ideas of classification and regression trees (CART), first introduced in [5], as well as the ideas of quantization.

The general setup is that one draws n samples i.i.d from P_X, say $X_1, \ldots, X_n$. One observes noisy realizations of a function f at these points, call these $Y_1, Y_2, \ldots, Y_n$. The goal is to estimate f by means of an estimator $\hat{f}$. In this chapter, we let X live in $\mathbb{R}^d$, and Y live in $\mathbb{R}$.

The idea of a classification and regression tree is to construct a cuboidal partitioning of $\mathbb{R}^d$, or more commonly a bounded subset of $\mathbb{R}^d$, typically taken as $[0,1]^d$ in the literature. The partitioning is constructed by recursively splitting along coordinate axes, choosing the axis to split upon and the location of the split based on the training data $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_d, Y_d)\}$ according to some rule. This is then iterated until a certain stopping condition is met. A typical stopping condition is when there are too few points left in an element of the partition, so that if the partitioning were to continue, some cells would become empty. For more information on typical splitting and stopping rules, we refer the reader to [15, Sec. 9.2]. Following this training

procedure, a test data point x is first mapped onto its containing cell, and then an aggregation procedure based on the training data within that cell is employed to decide the response $\hat{f}(x)$. In the case of regression, a common aggregation rule is to simply take the average of the training responses within the cell. In the case of classification, a common rule is to take a majority vote on the training labels within the cell. The "tree" term arises because the mapping of a point x to a cell is accomplished via a tree traversal, each decision being made by thresholding a coordinate.

There is a wide range of literature on this topic. Here, we mention that this method can be extended to the training of multiple trees over different subsets of the data, constructing what is referred to as a "random forest" [4]. Although quite popular in practice [15, Chap. 15], there are still many open questions regarding theoretical understanding of the method, such as those outlined in [3].

Part of the original motivation for the CART methodology was the notion of "interpretability". Although "interpretability" is a loose term, generally it refers to a classification or regression procedure with simple, easy to understand, and easy to modify rules. The coordinate splittings and tree structure are very helpful in this respect. Nevertheless, upon extending to random forests, some of this is lost.

Here, we present an abstract generalization of the CART idea that focuses on the key principle of defining a partition and using locality. We then specialize it and study one such specialization based on lattices in greater detail.

Definition 3. *Let $\mathcal{M}, \mathcal{N}$ be subsets of normed vector spaces. Consider drawing n samples i.i.d from P_X, say $\{X_1, X_2, \ldots, X_n\} \subseteq \mathcal{M}$. One then obtains from a query oracle $\{(X_1, Y_1), \ldots, (X_n, Y_n)\}$, where $Y_i = f(X_i) + Z_i$. Here, $f : \mathcal{M} \to \mathcal{N}$, and $Z_i \in \mathcal{N}$ are i.i.d noise instantiations of Z. The goal is to estimate f by an estimator $\hat{f}$, so that $|f(x) - \hat{f}(x)|_{\mathcal{N}}$ is as small as possible for all $x \in \mathcal{M}$. Note that this by itself does not specify an objective function, as there is ambiguity to how the different $x \in \mathcal{M}$ are treated relative to each other. Nevertheless, this serves as an overall goal for the schemes detailed here.*

Consider a estimator $\hat{f}$ constructed as follows. It first partitions the "root" $\mathcal{M}_0 = \mathcal{M}$ into the "first level" $\{\mathcal{M}_{00}, \ldots, \mathcal{M}_{0i}\}$. The "second level" is constructed from the

first by partitioning some of the $\mathcal{M}_{0i}$, forming sets labeled $\mathcal{M}_{0ij}$. This procedure is repeated a finite number of times until a stopping condition is met. Both the partitioning and the stopping condition may depend on the training data.

On test data point x, the estimator $\hat{f}$ outputs $\hat{f}(x) = A(\{(X_i, Y_i) : (X_i) \in \mathcal{M}_x\})$, where $\mathcal{M}_x$ denotes the "neighborhood" of point x, i.e. it is the smallest set constructed by the partitioning procedure containing x. Here, A denotes an "aggregation procedure". If the neighborhood contains no training points, there are a variety of reasonable estimates that can be formed. For simplicity, in our subsequent discussion we output the average over all points $\hat{f}(x) = \frac{\sum_{i=1}^{n} Y_i}{n}$ in such a case. Note that in more sophisticated analysis, one can try climbing up the tree towards the root until at least one point is found. However, this can result in very high variance if the first such node has just a single point. The above simple rule keeps the variance low by averaging over all the noise realizations, at the cost of greater bias. In fact, the proof of Lemma 7 shows that this term does not dominate the expression in our setup and analysis, which intuitively makes sense as it should be a low probability event compared to other sources of error. We call such an estimator a ***locality based estimator****.*

We note that the family of locality based estimators is very large. In particular, nearest neighbor methods, weighted neighborhood methods, CART, and linear classifiers are special cases of this family. The differences lie primarily in the partitioning and stopping condition, though sometimes the aggregation procedure also varies. It is intuitively clear that the performance of these methods crucially depends on some sort of locality assumption on f, a popular choice being a Lipschitz assumption.

All of the above examples use a "data dependent" partitioning scheme of $\mathcal{M}$. By this we mean that the partitioning depends on the actual (X_i, Y_i) pairs. On the other hand, in many other lines of research, "data independent" partitioning is used. For example, standard quantizers in information theory and communications are data independent. Locality sensitive hash functions, introduced in [18] for the problem of approximate nearest neighbors are also data independent, though recent work [2] uses a data dependent hash family for this problem.

In the subsequent discussion, we explore the possibilities of a "data independent"

partitioning scheme for a locality based estimator.

3.2 A Data Independent Partitioning Scheme

All theoretical analysis of locality based estimators requires structure on P_X. For instance, if one wants a uniform bound on $l(x) = E[|f(x) - \hat{f}(x)|]$, one needs a probability density (or pmf) bounded away from zero. Otherwise, one runs into "black swan" problems, such as Example 2. Structure on f can be used to alleviate some of this, but in general is insufficient on its own in the formation of consistent estimators, i.e. estimators where the loss function goes to 0 uniformly as $n \to \infty$ keeping d and other parameters of the setup fixed.

Typically, such analysis is carried out for the case when P_X is uniform for simplicity. Usually, such analysis extends easily to scenarios where $p_{min} \leq P_X \leq p_{max}$, with p_{min} and p_{max} showing up in the expressions. We shall assume P_X is uniform over a region $\mathcal{R} \subseteq \mathbb{R}^d$ here.

Consider a data independent scheme where the region $\mathcal{R}$ is partitioned into k sub-regions of equal measure. Then, we have the following lemma:

Lemma 7. *Consider $\mathcal{M} = \mathcal{R} \subseteq \mathbb{R}^d$ with the l_2 norm and $\mathcal{N} = \mathbb{R}$ with the l_1 norm in Definition 3. Consider a partitioning scheme that stops at level 1, resulting in a partition of $\mathcal{R}$ into $\mathcal{R}_i, 1 \leq i \leq k$ with $vol(\mathcal{R}_i) = \frac{vol(\mathcal{R})}{k}$ and $diam(\mathcal{R}_i) \leq s$ for all $1 \leq i \leq k$. Here, vol denotes the Lebesgue measure, and diam the diameter of a set. Assume that the noise Z satisfies $|Z| \leq z_0$ almost surely and $\mathbb{E}[Z] = 0$. Assume $P_X \sim U(\mathcal{R})$. Assume $|f(x) - f(y)| \leq M||x - y||_2$, or in other words f is 1-Lipschitz with Lipschitz constant M. Then by using a locality based estimator $\hat{f}$ for this setup, we have for any $x \in \mathcal{R}$:*

$$\mathbb{E}[|f(x)-\hat{f}(x)|] < \left[Mdiam(\mathcal{R}) + \frac{z_0\sqrt{2\pi}}{\sqrt{n}} \right] e^{-\frac{n}{k}} + Ms + z_0\sqrt{2\pi} \left(e^{\frac{-n}{2k^2}} + \sqrt{\frac{k}{2n}} \right). \quad (3.1)$$

Proof. Let $\mathcal{N}_x$ denote the neighborhood of x as described in Definition 3. Then, $|\mathcal{N}_x| \sim Binom(n, \frac{1}{k})$. Conditioning on $|\mathcal{N}_x|$ and using the law of total expectation,

and using subscripts x_i to denote the indices of random variables corresponding to $\mathcal{N}_x$, we have:

$$\begin{aligned}
\mathbb{E}[|f(x) - \hat{f}(x)|] &= \left(1 - \frac{1}{k}\right)^n \mathbb{E}\left[\left|f(x) - \frac{\sum_{i=1}^n f(X_{x_i})}{n} - \frac{\sum_{i=1}^n Z_{x_i}}{n}\right|\right] + \\
&\sum_{l=1}^n P(|\mathcal{N}_x| = l)\mathbb{E}\left[\left|f(x) - \frac{\sum_{i=1}^l f(X_{x_i})}{l} - \frac{\sum_{i=1}^l Z_{x_i}}{l}\right|\right] \\
&\leq \left(1 - \frac{1}{k}\right)^n \left(\mathbb{E}\left[\left|f(x) - \frac{\sum_{i=1}^n f(X_{x_i})}{n}\right|\right] + \mathbb{E}\left[\left|\frac{\sum_{i=1}^n Z_{x_i}}{n}\right|\right]\right) + \\
&\sum_{l=1}^n P(|\mathcal{N}_x| = l)\left(\mathbb{E}\left[\left|f(x) - \frac{\sum_{i=1}^l f(X_{x_i})}{l}\right|\right] + \mathbb{E}\left[\left|\frac{\sum_{i=1}^l Z_{x_i}}{l}\right|\right]\right) \\
&\leq \left[Mdiam(\mathcal{R}) + \mathbb{E}\left[\left|\frac{\sum_{i=1}^n Z_{x_i}}{n}\right|\right]\right]\left(1 - \frac{1}{k}\right)^n + \\
&Ms\left[1 - \left(1 - \frac{1}{k}\right)^n\right] + \sum_{l=1}^n P(|\mathcal{N}_x| = l)\mathbb{E}\left[\left|\frac{\sum_{i=1}^l Z_{x_i}}{l}\right|\right]. \qquad (3.2)
\end{aligned}$$

But by Hoeffding's inequality [17], we have:

$$\begin{aligned}
&P\left(\left|\frac{\sum_{i=1}^l Z_i}{l}\right| > t\right) \leq 2e^{\frac{-t^2 l}{2z_0^2}} \\
&\Rightarrow \mathbb{E}\left[\left|\frac{\sum_{i=1}^l Z_i}{l}\right|\right] \leq \frac{z_0\sqrt{2\pi}}{\sqrt{l}}. \qquad (3.3)
\end{aligned}$$

Using (3.3) and applying Hoeffding again, we get:

$$\begin{aligned}
\sum_{l=1}^n P(|\mathcal{N}_x| = l)\mathbb{E}\left[\left|\frac{\sum_{i=1}^l Z_{x_i}}{l}\right|\right] &\leq z_0\sqrt{2\pi}\left(\sum_{l=1}^n \frac{P(|\mathcal{N}_x| = l)}{\sqrt{l}}\right) \\
&\leq z_0\sqrt{2\pi}\left(P\left(|\mathcal{N}_x| \leq \frac{n}{2k}\right) + \sqrt{\frac{k}{2n}}P\left(|\mathcal{N}_x| > \frac{n}{2k}\right)\right) \\
&< z_0\sqrt{2\pi}\left(e^{\frac{-n}{2k^2}} + \sqrt{\frac{k}{2n}}\right). \qquad (3.4)
\end{aligned}$$

Plugging in (3.3) and (3.4) in (3.2), and using $(1 - \frac{1}{k})^k < e^{-1}$, we get the desired:

$$\mathbb{E}[|f(x) - \hat{f}(x)|] < \left[Mdiam(\mathcal{R}) + \frac{z_0\sqrt{2\pi}}{\sqrt{n}}\right]e^{-\frac{n}{k}} + Ms + z_0\sqrt{2\pi}\left(e^{\frac{-n}{2k^2}} + \sqrt{\frac{k}{2n}}\right).$$

□

3.3 A Lattice Based Partitioning Scheme

Lemma 7 offers a path towards understanding data independent partitioning and their associated locality based estimators. Nevertheless, it is somewhat abstract in that there are numerous degrees of freedom, such as how k can be picked in relation to n, what the regions $\mathcal{R}_i$ can look like, etc. Here, we offer a specialization which uses a nested partitioning scheme based on lattices in order to establish a family of consistent estimators. We refer the reader to [31] for a beautiful introduction to the topic of lattices and their use in engineering applications.

The issue of using a partition scheme based on nested lattices directly is that with Voronoi cells, one may need to reduce modulo Λ_2, where $\Lambda_2 \subseteq \Lambda_1$ denotes the coarse lattice. This can introduce nuisance terms into the loss analysis. In particular, if one wants a uniform bound on $l(x)$ over all x, problems arise with x in the cells that "wrap-around" during the modulo reduction, as these do not have good locality properties. If one is interested simply in an expectation over x with respect to some reasonable distribution, this is fine as such "wrap-arounds" are not too frequent. Nevertheless, [31, Sec. 8.4.1] offers an alternative where one considers parallelepiped cells, in which case these issues do not arise.

Here, we give an illustration of this using a scaled cubic lattice in the following proposition.

Proposition 8. *Let $\mathcal{R} = [0,1]^d$, R_i for $1 \leq i \leq k = q^d$ be a parallelepiped partition generated by the q-scaled cubic lattice $\{x : x = \frac{1}{q}(x_1, x_2, \ldots, x_d), 0 \leq x_i \leq q\}$. Let $n = k^{2+\epsilon}$ for a fixed $\epsilon > 0$ (one can handle non-integral n by taking the ceiling, or imposing restrictions on the form of q, such that it must be a perfect square for the case of $\epsilon = \frac{1}{2d}$). Then under the assumptions of Lemma 7, and using the locality based*

estimator outlined there, we have:

$$\mathbb{E}[|f(x)-\hat{f}(x)|] < \left[M\sqrt{d} + z_0\sqrt{2\pi}q^{-\frac{(2+\epsilon)d}{2}}\right] e^{-q^{(1+\epsilon)d}} + \frac{M\sqrt{d}}{q} + z_0\sqrt{2\pi}\left(e^{-\frac{q^{\epsilon d}}{2}} + \frac{1}{\sqrt{2}}q^{-\frac{(1+\epsilon)d}{2}}\right). \tag{3.5}$$

Thus, as $q \to \infty$, we have consistency.

Proof. Proof is immediate from Lemma 7. Indeed, applying Lemma 7 and substituting the appropriate values, we have:

$$\mathbb{E}[|f(x)-\hat{f}(x)|] < \left[M\sqrt{d} + z_0\sqrt{2\pi}q^{-\frac{(2+\epsilon)d}{2}}\right] e^{-q^{(1+\epsilon)d}} + \frac{M\sqrt{d}}{q} + z_0\sqrt{2\pi}\left(e^{-\frac{q^{\epsilon d}}{2}} + \frac{1}{\sqrt{2}}q^{-\frac{(1+\epsilon)d}{2}}\right).$$

In particular, letting $q \to \infty$, we have $\mathbb{E}[|f(x) - \hat{f}(x)|] \to 0$ uniformly over x, i.e we have consistency as desired. □

3.4 Conclusions

1. We outlined a general viewpoint on a variety of popular classification and regression methodologies, including but not limited to nearest neighbor methods, CART, and linear classifiers. All of these estimators fundamentally rely on some locality assumption in a metric space and a partitioning scheme which relies on such locality.

2. We discussed how these popular methods rely on a data dependent partitioning, and address the question of whether data independent partitioning can be used to get consistency. Lemma 7 and its specialization via scaled cubic lattices in Proposition 8 demonstrate that consistent estimators can be obtained.

3. As can be seen from (3.5), the number of samples used is exponential in d as a function of the loss level. This is a fundamental limitation under Lipschitz assumptions on f. In particular, random forests and nearest neighbor methods also suffer from this curse of dimensionality. It would thus be of interest to see whether locality based estimators can be adapted to make use of structure on f

and remove this exponential complexity. A stronger question would be whether data independent partitioning is sufficient for this purpose.

Chapter 4

Conclusion

In this thesis we primarily addressed the question of learning ANOVA decompositions over $\mathbf{GF}(2)$ and $\mathbb{R}$. The order of the ANOVA decomposition serves as a useful measure of the complexity of a model that is easily interpretable in terms of the degree of interactions between different coordinates. In the context of ANOVA decompositions over $\mathbf{GF}(2)$, we obtained fundamental limits on the performance of learning algorithms. In particular, we demonstrated a learning hierarchy of statistical complexity from linear to exponential in the dimension, justifying the title of this thesis. We also demonstrated the usefulness of the sparsity paradigm in this context. Furthermore, we discussed this problem in both an "adaptive" and "one-shot" setting. We showed that, somewhat surprisingly, the increased freedom from adaptivity does not result in significant changes to the statistical complexity. It will be interesting to see whether this statement holds up with respect to the computational complexity of learning algorithms for these tasks. In the context of $\mathbb{R}$, we obtained a glimpse into the possibilities of learning beyond the kernel assumptions of [19]. In future work, we hope to refine this further to account for noise and a lack of query access, thus making the methods more robust and general. Moreover, the construction of low computational complexity algorithms for the above problems is another interesting direction for future work.

We also developed a general viewpoint on a wide class of popular regression methods. We introduced the concept of data independent partitioning as a specialization

of this general viewpoint. By using a lattice based partitioning scheme, we demonstrated that this can achieve statistical consistency. Nevertheless, this method also suffers from the fundamental curse of dimensionality in higher dimensions. It would thus be interesting to figure out whether restrictions on the class of functions can be naturally incorporated into these methods in order to reduce the complexity.

Bibliography

[1] Noga Alon, Yossi Matias, and Mario Szegedy. The space complexity of approximating the frequency moments. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*, pages 20–29. ACM, 1996.

[2] Alexandr Andoni and Ilya Razenshteyn. Optimal data-dependent hashing for approximate near neighbors. In *Proceedings of the Forty-Seventh Annual ACM on Symposium on Theory of Computing*, pages 793–801. ACM, 2015.

[3] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

[4] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[5] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[6] Leo Breiman and Jerome H. Friedman. Estimating optimal transformations for multiple regression and correlation. *Journal of the American statistical Association*, 80(391):580–598, 1985.

[7] Venkat Bala Chandar. *Sparse graph codes for compression, sensing, and secrecy*. PhD thesis, Massachusetts Institute of Technology, 2010.

[8] Thomas H. Cormen, Charles Eric Leiserson, Ronald L. Rivest, and Clifford Stein. *Introduction to algorithms*, volume 6. MIT press Cambridge, 2001.

[9] Abhik Kumar Das and Sriram Vishwanath. On finite alphabet compressive sensing. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 5890–5894. IEEE, 2013.

[10] Stark C Draper and Sheida Malekpour. Compressed sensing over finite fields. In *Proceedings of the 2009 IEEE international conference on Symposium on Information Theory-Volume 1*, pages 669–673. IEEE Press, 2009.

[11] Ronald Aylmer Fisher. *Statistical methods for research workers*. Genesis Publishing Pvt Ltd, 1925.

[12] Ronald L. Graham, Donald E. Knuth, and Oren Patashnik. *Concrete Mathematics: A Foundation for Computer Science (2nd Edition)*. Addison-Wesley Professional, 2 edition, 3 1994.

[13] Andrew Granville. Arithmetic properties of binomial coefficients. i. binomial coefficients modulo prime powers. organic mathematics (burnaby, bc, 1995), 253–276. In *CMS Conf. Proc*, volume 20, pages 151–162, 1997.

[14] László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Science & Business Media, 2006.

[15] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics)*. Springer, 2nd ed. 2009. corr. 7th printing 2013 edition, 4 2011.

[16] Trevor Hastie, Robert Tibshirani, and Martin Wainwright. *Statistical learning with sparsity: the lasso and generalizations*. CRC Press, 2015.

[17] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American statistical association*, 58(301):13–30, 1963.

[18] Piotr Indyk and Rajeev Motwani. Approximate nearest neighbors: towards removing the curse of dimensionality. In *Proceedings of the thirtieth annual ACM symposium on Theory of computing*, pages 604–613. ACM, 1998.

[19] Kirthevasan Kandasamy and Yaoliang Yu. Additive approximations in high dimensional nonparametric regression via the salsa. *arXiv preprint arXiv:1602.00287*, 2016.

[20] Ernst Eduard Kummer. Über die ergänzungssätze zu den allgemeinen reciprocitätsgesetzen. *Journal für die reine und angewandte Mathematik*, 44:93–146, 1852.

[21] David Asher Levin, Yuval Peres, and Elizabeth Lee Wilmer. *Markov chains and mixing times*. American Mathematical Soc., 2009.

[22] Édouard Lucas. Théorie des fonctions numériques simplement périodiques.[continued]. *American Journal of Mathematics*, 1(3):197–240, 1878.

[23] Ian Grant Macdonald. *Symmetric functions and Hall polynomials*. Oxford university press, 1998.

[24] John Stuart Mill. *A system of logic ratiocinative and inductive: Being a connected view of the principles of evidence and the methods of scientific investigation*. Harper, 1884.

[25] J Ian Munro and Mike S Paterson. Selection and sorting with limited storage. *Theoretical computer science*, 12(3):315–323, 1980.

[26] Ryan O'Donnell. *Analysis of boolean functions*. Cambridge University Press, 2014.

[27] Yury Polyanskiy and Yihong Wu. Lecture Notes on Information Theory. `http://people.lids.mit.edu/yp/homepage/data/itlectures_v4.pdf`, 2016. [Online; accessed 16-June-2016].

[28] Hoifung Poon and Pedro Domingos. Sum-product networks: A new deep architecture. In *Computer Vision Workshops (ICCV Workshops), 2011 IEEE International Conference on*, pages 689–690. IEEE, 2011.

[29] Stephen M Stigler. *The history of statistics: The measurement of uncertainty before 1900*. Harvard University Press, 1986.

[30] Swami Vivekananda. *Karma Yoga: The yoga of action*. Advaita Ashrama, 2015.

[31] Ram Zamir. *Lattice Coding for Signals and Networks: A Structured Coding Approach to Quantization, Modulation, and Multiuser Information Theory*. Cambridge University Press, 2014.