Agnostic Learning and Single Hidden Layer Neural Networks

Wee Sun Lee

BE. University of Queensland

February 1996

A thesis submitted for the degree of Doctor of Philosophy

of the Australian National University

Department of Systems Engineering

Research School of Information Sciences and Engineering

Australian National University

Declaration

These doctoral studies were conducted with Dr Robert C. Williamson and Dr Peter L. Bartlett as supervisors and Dr Rodney A. Kennedy as advisor.

The work presented in this thesis is the result of original research carried out by myself, in collaboration with others, whilst enrolled in the Department of Systems Engineering as a candidate for the degree of Doctor of Philosophy. This work has not been submitted for any other degree or award in any other university or educational institution.

Wee Sun Lee

February 1996

List of Publications

A number of papers resulting from this work have been submitted to refereed journals or are in preparation.

- Lee, W. S., Bartlett, P. L. and Williamson, R. C., 'Efficient agnostic learning of neural networks with bounded fan-in'. Accepted for publication in *IEEE Trans. on Information Theory*.
- Lee, W. S., Bartlett, P. L. and Williamson, R. C., 'The importance of convexity in learning with squared loss'. Submitted to *IEEE Trans. on Information Theory*.
- Lee, W. S., 'Learning smooth functions in high dimensions'. *In preparation*.
- Lee, W. S., 'Fixed basis functions for approximation in high dimensions'. *In preparation*.

A number of papers on this work has also been presented at or submitted to conferences. Some of the material overlaps with that covered in the journal submissions.

- Lee, W. S., Bartlett, P. L. and Williamson, R. C. (1995), 'Efficient agnostic learning of neural networks with bounded fan-in', *in* 'Proc. 6th Australian Conference on Neural Networks', pp. 201–204.
- Lee, W. S., Bartlett, P. L. and Williamson, R. C. (1995), 'On efficient agnostic learning of linear combinations of basis functions', *in* 'Proc. 8th Annu. Workshop on Comput. Learning Theory', ACM Press, New York, NY, pp. 369–376.
- Lee, W. S., Bartlett, P. L. and Williamson, R. C. (1995), 'The importance of convexity in learning with squared loss'. Submitted to the 9th Annu. Workshop on Comput. Learning Theory.

During my doctoral studies, I did some work listed below which is not covered in this thesis.

- Lee, W. S., Bartlett, P. L. and Williamson, R. C. (1994), 'The Vapnik-Chervonenkis dimension of neural networks with restricted parameter ranges', *in* 'Proc. 5th Australian Conference on Neural Networks', pp. 198–201.
- Lee, W. S., Bartlett, P. L. and Williamson, R. C. (1994), 'Lower bounds on the VC-dimension of smoothly parametrized function classes, *in* 'Proc. 7th Annu. ACM Workshop on Comput. Learning Theory', ACM Press, New York, NY, pp. 362–367.
- Lee, W. S., Bartlett, P. L. and Williamson, R. C. (1995), 'Lower bounds on the VC-dimension of smoothly parametrized function classes', *Neural Computation* **7**(5), 1040–1053.
- Lee, W. S., Bartlett, P. L. and Williamson, R. C. (1995), 'Correction to "Lower bounds on the VC-dimension of smoothly parametrized function classes"'. Submitted to *Neural Computation*.
- Lee, W. S. and Telford, A. (1995), 'On the application of recursive systems identification techniques to blast furnace thermal condition index estimation'. Technical report, BHP Research Laboratories.

Acknowledgements

I would like to thank my supervisors Dr Robert C. Williamson and Dr Peter L. Bartlett for their supervision, support and guidance throughout my studies. Much of what I learned about research work is due to my interaction with them. I would also like to thank Professor John B. Moore, Head of the Department of Systems Engineering for his advice and encouragements and for providing a wonderful working environment. I have also benefited from my involvement with the Co-operative Research Centre for Robust and Adaptive Systems.

For many useful suggestions on the work done in this thesis, I would like to thank Professor Andrew R. Barron.

Thanks to Dr John Shawe-Taylor for hosting my visit to Royal Holloway, University of London and to Dr Jonathan Baxter and Dr Martin Anthony for their friendships while I was there. Thanks also to Dr Andrew Telford for providing me with the opportunity to experience research in industry at BHP Research Laboratories in Melbourne.

For teaching interesting classes, thanks to Dr Iven Mareels, Professor Bob Bitmead and Professor Brian Anderson.

My fellow students have made the department a pleasant place to work in. Some have also become good friends. Thanks particularly to Subhra, Dragan, Leonardo, Natasha, Andrew, Jason, Dan-chi, Jeremy, Anton and Ari.

I've made many good friends in the last three years. I'd like to thank Klaus, Jennifer, Karthika, Alba, Gan, Pinky and many others for the times we shared together.

Last but not least, I'd like to thank my family for their support through all these years.

Abstract

This thesis is concerned with some theoretical aspects of supervised learning of real-valued functions. We study a formal model of learning called agnostic learning. The agnostic learning model assumes a joint probability distribution on the observations (inputs and outputs) and requires the learning algorithm to produce an hypothesis with performance close to that of the best function within a specified class of functions. It is a very general model of learning which includes function learning, learning with additive noise and learning the best approximation in a class of functions as special cases.

Within the agnostic learning model, we concentrate on learning functions which can be well approximated by single hidden layer neural networks. Artificial neural networks are often used as black box models for modelling phenomena for which very little prior knowledge is available. Agnostic learning is a natural model for such learning problems. The class of single hidden layer neural networks possesses many interesting properties, which we explore in this thesis, within the agnostic learning model.

Two main aspects of learning studied here are the amount of information required (the sample complexity) and the amount of computation required (computational complexity) for agnostic learning. We determine the sample complexity for agnostic learning based on properties of the function class such as the pseudo-dimension and the fat-shattering function and show that for certain function classes, if the closure of the function class is not convex, the sample complexity for agnostic learning (with squared loss) can be worse than the sample complexity for learning with additive noise if we are restricted to hypotheses from the same class. We also show that if the closure of the function class is convex, then the sample complexity bound is similar to that for learning with noise. This motivates learning convex hulls of non-convex function classes. For many function classes, the convex hull can be represented by single hidden layer neural networks with an unbounded number of hidden units and a bound on the sum of the absolute values of

the output layer weights. We show that for certain function classes, agnostic learning of the convex hull gives better approximation to the target function (conditional expectation) without much penalty to the order of the sample complexity.

We show that the class of single hidden layer neural networks is efficiently (polynomial-time) agnostically learnable if and only if the class of hidden units is efficiently agnostically learnable. However, we also show that many classes of single hidden layer neural networks (including that with linear threshold units as hidden units) are unlikely to be efficiently agnostically learnable. This leads to the study of subclasses of functions which are efficiently agnostically learnable. We show that function classes which can be well approximated by single hidden layer neural networks with bounded fan-in are efficiently agnostically learnable. We also show that if the functions in a class are smooth enough in a certain sense and have small L_1 norms, then the class is efficiently agnostically learnable using single hidden layer neural networks.

As many of the function classes considered in this thesis are nonparametric (infinite dimensional), we also consider the rate of approximation for these function classes. We give an iterative approximation result for finding the best approximation in the convex hull of a function class when the target is outside the class. We also show the existence of a small (polynomial sized) set of fixed basis functions for the approximation of certain smooth functions in high dimensions.

Contents

So I took thought, and invented what I conceived
to be an appropriate title of "agnostic".

— Thomas Henry Huxley
"Agnosticism" in Gordon Stein, editor,
An Anthology of Atheism and Rationalism.

Chapter 1

Introduction

Machine learning is concerned with computer programs which improve their performance through experience. In this thesis, we investigate certain aspects of machine learning through a formal model of learning called agnostic learning. Agnostic learning is a fairly realistic model of learning which makes virtually no assumptions about the target function which a learning algorithm is trying to learn. Instead, it assumes that the learning algorithm will search a limited space of hypothesis functions in an attempt to find the "best" approximation to the target function.

The agnostic learning model we use is based on the models introduced by Haussler (1992) and Kearns, Schapire & Sellie (1994) to address the shortcomings of of the Probably Approximately Correct (PAC) learning model (Valiant 1984). The PAC model is a formal model for learning $\{0, 1\}$-valued functions which assumes that the target function is known to be in a particular class of functions. While these assumptions have permitted rigorous study of the complexity of learning as a function of the representational complexity of the target function, it diverges from the typical setting encountered in empirical machine learning. In practical machine learning problems, observations are often noisy and very little is usually known about the target function. Because of limited resources, practitioners usually attempt to find a useful hypothesis from a small class of functions which may not necessarily contain the best possible function for the problem. Agnostic learning generalizes the PAC learning model to encompass these situations and also to allow learning real valued functions.

Within the agnostic learning framework, we study the function classes represented by artificial neural networks, particularly single hidden layer neural networks. Artificial neural networks form a flexible class of functions which has been used successfully for various machine learning problems such as learning to play backgammon (Tesauro & Sejnowski 1989, Tesauro 1990), learning to recognize hand-written zip codes (Le Cun, Boser, Denker, Henderson, Howard, Hubbard & Jackel 1989) and learning to navigate a car (Pomerleau 1989). Investigating particular classes of functions allows us to elucidate some of the fundamental properties of the agnostic learning framework. It also gives considerable insights into the properties of the function class being investigated (in this case, artificial neural networks). Such insights are useful for deciding when it is appropriate to use the function class. Such insights may also give useful information about the nature of the learning task when the hypothesis produced by the learning algorithm fails to perform well. Knowing the properties of the function class can also help us to design learning algorithms, to choose input representations, to decide on proper preprocessing of the data and to partition the learning problems into appropriate subproblems.

In agnostic learning, the only assumption made about the phenomenon that is being learned is that it can be represented by a joint probability distribution on $\mathcal{X} \times \mathcal{Y}$ where $\mathcal{X}$ is the domain and $\mathcal{Y}$ is a bounded subset of $\mathbb{R}$. The term *agnostic* reflects the fact that the learning algorithm has no a priori "beliefs" regarding the structure of the phenomenon. As such, we cannot expect the learning algorithm to always produce a model with small error. Instead, we demand that with high probability, the model produced is at least close to the optimal function in a certain class of functions. In this thesis, we use the expected squared error as a measure of the performance of the functions. (A more formal definition of agnostic learning, along with many of the terms used in this chapter, is given in Chapter 2.)

The agnostic learning model provides a very general framework for investigating learning problems. Any algorithm which performs well in the agnostic learning framework will also perform well in the following special cases:

- **Function Learning**. The inputs X_i are independently drawn from an unknown probability distribution but there is a deterministic relationship between the inputs and targets $Y_i = f(X_i)$, such that the function f belongs to the class used for learning. PAC learning is just a special case where the class of functions have $\{0, 1\}$-valued outputs.

- **Learning the Best Approximation**. This is the same as function learning except that the target function f does not have to be in the class used for learning. This represents the cases where the target function may be very complex but we are content to have a good approximation from a small approximating class.

- **Learning with Noise**. The observations (X_i, Y_i) are independently drawn from an unknown probability distribution on $\mathcal{X} \times \mathcal{Y}$ and the optimal function (in the sense of minimum mean squared error), $f^*(x) = \mathbf{E}(Y|X = x)$ is inside the class. This is the same as learning functions corrupted by additive bounded zero mean noise.

- **Approximating the Conditional Expectation**. As in the case of function learning, when the conditional expectation is not in the class used for learning, the aim becomes finding the best approximation to the conditional expectation. Learning the best approximation to the conditional expectation is the most general problem considered in this thesis.

- **Learning Probabilistic Concepts**. The targets $Y_i \in \{0, 1\}$ are noisy labels for a binary classification problem. The conditional expectation is again the optimal function. If the mean squared error for a function is small, the function will perform well when thresholded to give a classification function.

The generality of the agnostic learning framework has its disadvantages. The algorithms are required to perform well under all probability distributions on $\mathcal{X} \times \mathcal{Y}$. Results showing the difficulties of learning certain classes of functions in this framework do not mean that learning these classes of functions under certain other realistic conditions is also difficult. We also require the algorithms to perform well uniformly over the function class (that is, the best approximation may be provided by any function in the class). This is a more severe requirement than allowing the performance of the algorithms to depend on the target function. However, this framework provides an extreme condition which is useful for comparisons with other frameworks which make more assumptions.

In this thesis, we concentrate on two aspects of learning: the sample complexity and the computational complexity. For learning to be feasible, we require the sample size and the amount of computation time used by the learning algorithm to be small. As our computational and data gathering capabilities improve, the size of the problems which we can solve increases. However, if the sample size and the amount of computation required grow exponentially with the problem size, the size of the problems we can solve will grow very slowly. As such, we delineate the boundary of what is feasibly learnable by the existence of algorithms which use a polynomial sample size and run in polynomial time. We say that a function class is *efficiently (agnostically) learnable* if the sample size and the amount of computation grow at most polynomially with $1/\epsilon$, $1/\delta$ and the complexity parameters, where ϵ is the accuracy required and δ is the probability that the algorithm will fail. The complexity parameters are the parameters of interest which affect the complexity of the learning problem. For example, the dimension of the input space is a parameter which is particularly relevant in neural networks learning since applications using neural networks tend to have large input dimensions.

1.1 Related Work

The agnostic learning framework that we use follows closely that used by Kearns et al. (1994), who studied the computational complexity of agnostic learning for boolean functions and some simple real-valued functions. The computational complexity of learning has been extensively studied since the seminal work of Valiant (1984). Most of these works involve boolean functions in the PAC model (see e.g. (Pitt & Valiant 1988, Kearns, Li, Pitt & Valiant 1987, Kearns 1989)). Function classes which have been found to be efficiently PAC learnable include the classes of monomials, linear threshold functions, k-CNF and k-decision lists. Classes of functions which are not efficiently PAC learnable (unless $RP = NP$) include the class of k-term DNF, certain classes of neural networks (Judd 1990) and classes of two hidden unit neural networks (Blum & Rivest 1992, DasGupta, Siegelmann & Sontag 1995). For a survey of results on PAC learning, see (Kearns 1989). For agnostic PAC learning (agnostic learning with the values of the targets and hypotheses restricted to $\{0, 1\}$), considerably fewer function classes are efficiently learnable. Certain function classes which are efficiently PAC learnable, such as the class of monomials and the class of linear threshold functions, are not efficiently agnostically PAC learnable unless $RP = NP$ (Kearns et al. 1994, Höffgen & Simon 1992). For these hardness results (both PAC and agnostic learning), the learning algorithms are restricted to producing hypotheses from the same class as the target function. Restricting the hypothesis to be from the same class as the target function makes learning harder in some cases. For example, Pitt & Valiant (1988) have shown that the class of k-term DNF is efficiently learnable if the class of k-CNF is used as the hypothesis class, even though it is not efficiently learnable if the hypothesis is restricted to be a k-term DNF. This shows that choosing the appropriate representation for a learning problem is important. Representation independent hardness results based on cryptographic assumptions are also available for some function classes (Kearns 1989).

The computational complexity of learning real-valued functions is not as well studied. Maass (1995) has shown that fixed architecture neural networks with piecewise-polynomial activation functions are efficiently agnostically learnable (in the sense of being polynomial-time with respect to $1/\epsilon$ and $1/\delta$). Koiran (1994) has shown that single hidden layer neural networks with fixed input dimension, piecewise linear activation functions, an unbounded number of hidden units, and bounded sum of absolute values of output weights is efficiently learnable with small bounded noise.

The sample complexity of learning has been studied in various areas such as pattern recognition, statistics and computational learning theory. The agnostic learning framework is based on a general framework proposed by Haussler (1992). Within this framework, Haussler (1992) has given upper bounds on the sample complexity of learning based on properties of the function class such as the pseudo-dimension and the covering number. Function classes examined by Haussler (1992) include classes of neural networks with a variety of activation functions. More recent works on the sample complexity of agnostic learning include (Bartlett, Long & Williamson 1994) and (Bartlett & Long 1995) where upper and lower bounds on the sample complexity of agnostic learning are given based on a property of the function class called the fat-shattering function. These works are based on results on uniform convergence of empirical estimates which has been studied in the empirical process literature (Vapnik & Chervonenkis 1971, Pollard 1984, Pollard 1990, Dudley 1978). Earlier works on sample complexity based on uniform convergence results can be found in (Vapnik 1982, Blumer, Ehrenfeucht, Haussler & Warmuth 1989).

A lot of related work has also been done in the area of nonparametric statistical theory of curve estimation and classification. In nonparametric estimation, the target function is only usually restricted by some general smoothness properties and is not assumed to be a member of a finite dimensional function class. To learn such functions, it is necessary to consider the approximation error as well as the estimation error from using a finite sample size. Various nonparametric estimators can be used to learn such function classes (see (Silverman 1986, Eubank 1988, Hardle 1990)). These classes can also be learned by using sequences of parametric function classes where the dimensions of the parametric function classes grow as a function of the sample size so that arbitrarily good approximation to the target function can be obtained. The results in the nonparametric statistics literature are usually given in the form of the risk of an estimator (learning algorithm) as a function of the sample size and not in the form of sample complexity as is common in the computational learning theory literature. Although the framework used is different, these results are similar to those in the computational learning theory literature in the sense that they give the rate at which the sample size must grow as we require greater accuracy from the learning algorithms. It is known that the minimax rate of convergence of the mean integrated squared error for functions with all partial derivatives of order s square-integrable is of the order $(1/m)^{2s/(2s+n)}$ where m is the sample size and n is the input dimension (Ibragimov & Hasminskii 1980, Pinsker 1980, Stone 1982, Nussbaum 1986). These rates suggest that efficient learning (in terms of sample size) for these function classes is not possible unless s grows linearly with n. Another way to restrict the function class so that the sample size required for learning does not grow exponentially with the input dimension is given in (Barron 1994). There he shows that functions with bounded first absolute moment of the Fourier transform can be learned using single hidden layer neural networks with mean integrated squared error of $O(\sqrt{\log m/m})$.

In order to obtain bounds on the sample complexity for learning functions defined by smoothness constraints using a sequence of parametric function classes, bounds on the rate of approximation by the sequence of function classes are required. For functions with all partial derivatives of order s square integrable, the best approximation rate for the integrated squared error achievable by basis function expansions using order r^n parameters is of order $O(1/r^{2s})$ for $r = 1, 2, \ldots$, e.g. see (Pinkus 1985) (r is the degree of the polynomials for polynomial methods and the number of knots per coordinate for spline methods). The number of parameters used for approximation grows exponentially with the dimension of the input space for these methods. For functions with bounded first absolute moment of the Fourier transform, an approximation rate of order $O(1/k)$, where k is the number of basis functions, is achievable if the basis functions are adaptable (Barron 1993). However, it is not known if there is a polynomial time algorithm for adapting the basis functions to provide the approximation.

1.2 Contributions of the Thesis

In this thesis, we consider agnostic learning with the squared loss functions. We review known bounds on the sample complexity of agnostic learning based on properties of the function class such as the covering number, pseudo-dimension and fat-shattering function. We then compare the sample complexity of agnostic learning with the sample complexity of function learning and learning with noise. We show that if the closure of a function class is not convex, the sample complexity for agnostically learning the function class can be worse than the sample complexity for learning with noise if we are restricted to hypotheses from the same class. We also show that if the function class is convex, sample complexity similar to that for learning with noise can be provided for agnostic learning.

This motivates agnostically learning the convex hull of the function class instead of the function class itself. As an example, the class of fixed sized single hidden layer neural networks is not convex but the class of networks with an unbounded number of hidden units is. We show that for some function classes, the order of the sample complexity for agnostically learning the convex hull of the function class is similar to that for agnostically learning the function class if we are restricted to hypotheses from the same class. In some cases, the order of the sample complexity is better for agnostically learning the convex hull. Since the convex hull usually gives better approximation than the original function class, our results shows that in many situations it may be advantageous to learn the convex hull of a function class instead of the function class itself. The convex hull can be learned by increasing the number of functions in a convex combination of functions from the class as a function of the sample size. The convex hull of a function class can be thought of as the class of single hidden layer neural networks with an unbounded number of hidden units and a bound on the sum of absolute values of the output weights.

We also extend an iterative approximation result of Barron (1993) and Jones (1992) to the case when the target function does not belong to the convex hull of the function class. Then we explore the relationship between the computational complexity of learning a single hidden unit and the computational complexity of learning a single hidden layer neural network. In agnostic learning, a pleasing relationship exists because of the iterative approximation result. We show that the class of single hidden layer neural networks is efficiently agnostically learnable if and only if a single hidden unit is efficiently agnostically learnable. We also give some evidence showing that agnostically learning the class of single hidden neural networks is computationally difficult as the dimension of the input space grows.

Since agnostic learning for the class of single hidden layer neural networks is likely to be computationally difficult, we examine subclasses of functions which can be learned efficiently by single hidden layer neural networks. We show that low order function classes (function classes which can be approximated arbitrarily closely by a single hidden layer neural network with bounded fan-in) are efficiently agnostically learnable. We also show that classes of functions in n dimensions which are local in space (have small L_1 norm) and have finite q-th absolute moment of the Fourier transform are efficiently agnostically learnable as long as q increases linearly with n. These function classes can be learned using networks with sinusoidal hidden units as well as sigmoidal ones. This shows that sufficiently smooth functions which are well localised are learnable. While the capabilities of these learnable function classes are quite restricted, we believe they might still be useful in practice. Smoothness is a natural assumption for many learning problems and functions with small L_1 norm often appear in pattern recognition problems. For example, for character recognition problems, a single character (e.g. 'A'), without allowing noise and invariances is only one possible bitmap out of 2^n where n is the size of the bitmap. Allowing some noise and invariances, such as translation and some rotation, will increase the L_1 norm of the functions polynomially with n as long as the amount of noise is small. Similarly, low order functions can be practically useful as shown in (Boser, Guyon & Vapnik 1992) where good results for handwritten digit recognition is achieved using low degree polynomials (low order functions).

A set of basis functions which can be used to approximate one function to a certain accuracy may not necessarily be able to approximate another function from the same class to the same accuracy. For functions with uniformly bounded q-th moment of the Fourier series and uniformly bounded L_1 norm, we show that the size of a set of basis functions which can approximate all functions in the class to the same accuracy, need only grow polynomially (instead of exponentially) with the input dimension if q grows linearly with the input dimension. Since this set of basis functions can be used to approximate all functions in the class, it can be fixed in advance before any learning procedure. This result also gives a bound on the number of basis function needed for learning multi-output functions (with arbitrarily many outputs).

1.3 Outline of the Thesis

In Chapter 2, we give technical definitions and describe the agnostic learning model and the function classes used in the thesis.

Chapter 3 provides results for bounding the number of examples (sample complexity) required for learning classes of functions. We also show that if the function class is convex then we can provide a sample complexity bound for agnostic learning which is of the same order as the sample complexity bound for learning with noise.

In Chapter 4 we give a partial converse. That is, if the closure of the function class is not convex, the sample complexity for agnostic learning can be worse than the sample complexity for learning with noise if we are restricted to hypotheses from the same class.

Chapter 5 provides sample complexity bounds for learning the convex hull of function classes.

In Chapter 6, we consider the computational complexity of learning the class of single hidden layer neural networks. We show that this class is efficiently agnostically learnable if and only if the class of hidden units is efficiently agnostically learnable. We give some evidence indicating that some classes of single hidden layer neural networks are unlikely to be efficiently agnostically learnable. Finally, we show that single hidden layer neural networks with bounded fan-in are efficiently agnostically learnable.

In Chapter 7, we show that the class of functions with finite q-th absolute moment of the Fourier transform is efficiently agnostically learnable if q grows linearly with n, the input dimension. We also show that a smooth enough function class can be well approximated with a polynomial sized fixed set of basis functions.

We give the conclusions of the thesis in Chapter 8.

Learning – to acquire knowledge of or skill in by study, instruction, or experience.

— The Macquarie Dictionary

Chapter 2

Definitions and Learning Model

In this chapter, we define the agnostic learning model, the function classes used and give other relevant definitions and notation. The agnostic learning model presented here is based on the model used by Kearns et al. (1994) and Haussler (1992).

2.1 Agnostic Learning

Domain and Range. Let $\mathcal{X}$ be a set called the *domain* and let a point in $\mathcal{X}$ be called an *instance*, denoted x. In this thesis, $\mathcal{X}$ is usually $\mathbb{R}^n$ or a subset of $\mathbb{R}^n$. Let $\mathcal{Y} \subset \mathbb{R}$ be the *observed range*. We restrict ourselves to bounded ranges with $|y| \leq T$ for every $y \in \mathcal{Y}$. We call the pair $Z = (X, Y)$, randomly sampled according to some probability distribution on $\mathcal{X} \times \mathcal{Y}$, an observation. For learning problems, we are interested in finding a mapping from $\mathcal{X}$ to $\mathcal{Y}$ that will perform well.

Probability Distributions. For agnostic learning, we require that the algorithm be able to perform well over the class of all probability distributions on $\mathcal{X} \times \mathcal{Y}$. By restricting the class of probability distributions, we obtain the following special cases.

For *function learning*, we have an arbitrary distribution on $\mathcal{X}$ and $Y = f(X)$ for some f restricted to be in some class $\mathcal{F}$ which is used by the learning algorithms. The case when $\mathcal{F}$ consists of $\{0, 1\}$-valued functions and $Y \in \{0, 1\}$ is the situation in the well-known *Probably Approximately Correct (PAC)* learning model.

Learning the best approximation is the same as function learning except that the function f need not be from $\mathcal{F}$.

For *learning with noise*, Y is allowed to be a random variable but the conditional expectation $f^*(x) = \mathbf{E}(Y|X = x)$ must belong to $\mathcal{F}$.

Learning probabilistic concepts is the same as learning with noise except that $\mathcal{Y}$ is restricted to be $\{0, 1\}$.

Hypothesis and Target Class. We would often like to learn functions in one class using functions from another class. For example, the function class we are trying to learn may only be defined in terms of some smoothness properties but we may want to use single hidden layer neural networks to learn the functions. We call the function class which we use for learning the *hypothesis class* $\mathcal{H}$, and the function produced by the learning algorithm, the hypothesis. The function class we wish to learn is called the *target class* $\mathcal{T}$. In agnostic learning, we judge the performance of the algorithm by how well it performs relative to the best function in the target class instead of the hypothesis class. Using a hypothesis class which includes the target class (or can approximate the target class arbitrarily closely) makes it possible to do as well as the best function in the target class. Using a larger hypothesis class is also computationally advantageous in some situations (for an example in PAC learning, see (Pitt & Valiant 1988)).

Parametrized Classes. To study the complexity of learning, we will parametrize the classes of functions and classes of probability distributions by several measures of complexity. We will consider the classes of all probability distributions with bounded range indexed by the bound on the range T. We will index the target classes by a vector p of *complexity parameters*. The complexity parameters used in this thesis includes the dimension of the input space, measures of smoothness and other parameters described later in this chapter.

Loss Functions. To measure the performance of a hypothesis h on an observation (X, Y), we use a loss function $L(h(X), Y)$ also denoted $L_h(X, Y)$. In this thesis, we concentrate on the *quadratic loss function* $Q(h(X), Y) = (h(X) - Y)^2$. Other loss functions used include the *absolute loss function* $\Lambda(h(X), Y) = |h(X) - Y|$ and the *discrete loss function* $Z(h(X), Y) = 0$ if $h(X) = Y$ and $Z(h(X), Y) = 1$ otherwise.

Empirical and Expected loss. Let $\boldsymbol{Z} = (\boldsymbol{Z}_1, \boldsymbol{Z}_2, \ldots, \boldsymbol{Z}_m)$ be a sequence of observations. Call the probability distribution formed by giving each of $\boldsymbol{Z}_1, \ldots, \boldsymbol{Z}_m$ equal weighting the *empirical distribution*. The *empirical loss* of a hypothesis h is $\hat{\mathbf{E}}_{\boldsymbol{Z}}(L_h) = \frac{1}{m}\sum_{i=1}^{m} L(h(\boldsymbol{X}_i), \boldsymbol{Y}_i)$, denoted $\hat{\mathbf{E}}(L_h)$ when the meaning is clear from the context. Given observations drawn according to a probability distribution, the *expected loss* is $\mathbf{E}(L_h(X,Y))$ which we denote $\mathbf{E}(L_h)$ when the meaning is clear from the context. For a class $\mathcal{T}$, we define $opt(\mathcal{T}) = \inf_{h\in\mathcal{T}} \mathbf{E}[L_h]$ and $\hat{opt}(\mathcal{T}) = \inf_{h\in\mathcal{T}} \hat{\mathbf{E}}[L_h]$.

Agnostic Learning. Let p be a vector of complexity parameters parametrizing a function class $\mathcal{T}$. We say that $\mathcal{T}$, parametrized by p, is *agnostically learnable* (with respect to loss function L) if there exists a function class $\mathcal{H}$, a function $m(\epsilon, \delta, T, p)$ and an algorithm A such that for any probability distribution on $\mathcal{X} \times \mathcal{Y}$ with $\mathcal{Y} \subseteq [-T, T]$, for every $0 < \delta \leq 1$, $\epsilon > 0$, $T > 0$ and p, the algorithm draws $m(1/\epsilon, 1/\delta, T, p)$ independent observations and gives a hypothesis $h \in \mathcal{H}$ such that with probability at least $1-\delta$, $\mathbf{E}[L_h] \leq opt(\mathcal{T}) + \epsilon$.

We leave out the parametrization p when the meaning is clear from the context. Despite the name "algorithm", A is just a mapping from a sequence of observations to a function in $\mathcal{H}$. We will impose computational requirements only when we consider issues of computational complexity. With appropriate restrictions on the probability distributions, we define *function learnable* and *learnable with noise* in the same way. In all cases , if the hypothesis class $\mathcal{H}$ is restricted to be the same as the target class $\mathcal{T}$, we say that the function class is *properly* learnable.

Efficient Learning. We say that $\mathcal{T}$ is *efficiently* (agnostically) learnable if it satisfies a few other conditions in addition to the requirements for (agnostic) learnability. The sample size $m(1/\epsilon, 1/\delta, T, p)$ used by the learning algorithm must be bounded by a fixed polynomial in $1/\epsilon$, $1/\delta$, T and the components of p. The computation time of the algorithm A must be bound by a fixed polynomial in $1/\epsilon$, $1/\delta$, T and the components of p. The hypothesis produced must also be evaluable in time polynomial in $1/\epsilon$, $1/\delta$, T and the components of p. Such an algorithm is called an efficient (agnostic) learning algorithm.

Sample Complexity. The sample complexity for agnostically learning a function class $\mathcal{T}$ is the smallest number of observations $m(1/\epsilon, 1/\delta, T, p)$ necessary for the existence of a learning algorithm that can learn to accuracy ϵ with probability at least $1-\delta$ without regard to computational requirements. The sample complexity for efficiently agnostically learning a function class $\mathcal{T}$ is the smallest number of observations $m(1/\epsilon, 1/\delta, T, p)$ necessary for the existence of an efficient agnostic learning algorithm that can learn to accuracy ϵ with probability at least $1 - \delta$. The sample complexity of an agnostic learning algorithm is the number of observation used by that algorithm for learning to accuracy ϵ with probability at least $1 - \delta$. The sample complexity is defined similarly for function learning and learning with noise.

Computational Model. For simplicity, we work in the uniform cost model of computation (see (Aho, Hopcroft & Ullman 1974)). In the uniform cost model, real numbers occupy one unit of space and standard arithmetic operations (addition, multiplication etc.) take one unit of time. Where appropriate we also make comments on using the logarithmic cost model where numbers are represented in finite precision and operations on them are charged time proportional to the number of bits of precision (Aho et al. 1974).

2.2 *Function Classes*

Basis Functions. A class of real-valued functions $\mathcal{G}$ is an *admissible* class of basis functions if $\mathcal{G}$ is *permissible* and there exists $b > 0$ such that $|g(x)| \leq b$ for all $g \in \mathcal{G}$, $x \in \mathcal{X}$. Permissibility is a mild measurabililty condition which will be discussed in Section 2.3. We will also call a class of basis functions a class of *hidden units*. A class of basis functions is called *symmetric* if $-g \in \mathcal{G}$ for all $g \in \mathcal{G}$.

Single Hidden Layer Neural Networks. Let $\mathcal{G}$ be an admissible class of basis functions. Then for every $K > 0$, let

$$\mathcal{N}^{\mathcal{G}}_{K,k} = \left\{ x \mapsto w_0 + \sum_{i=1}^{k} w_i g_i(x) \colon g_i \in \mathcal{G}, w_i \in \mathbb{R}, \sum_{i=0}^{k} |w_i| \leq K \right\}.$$

Then $\mathcal{N}^{\mathcal{G}}_{K} = \bigcup_{k=1}^{\infty} \mathcal{N}^{\mathcal{G}}_{K,k}$ is the class of *linear combinations* of functions from $\mathcal{G} \bigcup \{x \mapsto 1\}$ with the sum of magnitudes of weights bounded by K. We will also call such function classes *single hidden layer neural networks*. The class of single hidden layer neural networks is the convex hull of the symmetric function class $\mathcal{G}^1_K = \{x \mapsto Kg(x), x \mapsto -Kg(x), x \mapsto K, x \mapsto -K \colon g \in \mathcal{G}\}$.

We will often use the following function class (indexed by k) to approximate $\mathcal{N}^{\mathcal{G}}_{K}$.

$$\mathcal{A}^{\mathcal{G}}_{K,k} = \left\{ x \mapsto \frac{1}{k} \sum_{i=1}^{k} g_i(x) \colon g_i \in \mathcal{G}^1_K \right\}.$$

Let S be the class of all one-to-one mappings from $\{1, \ldots, \tau\}$ to $\{1, \ldots, n\}$. Let $\mathcal{P} := \{(x_1, \ldots, x_n) \mapsto (x_{s(1)} \ldots, x_{s(\tau)}) : s \in S\}$. We say $\mathcal{G}$ has *fan-in* τ if there exists a class of functions $\mathcal{G}'$ mapping from $\mathbb{R}^\tau$ into $\mathbb{R}$ and $\mathcal{G} = \{x \mapsto g \circ p(x) : x \in \mathbb{R}^n, p \in \mathcal{P}, g \in \mathcal{G}'\}$. We say a single hidden layer neural networks with hidden units from $\mathcal{G}$ has fan-in τ if $\mathcal{G}$ has fan-in τ. We also call classes of functions which can be approximated arbitrarily closely by a single hidden layer neural network (with linear threshold hidden units) with fan-in τ (relative to the metric) classes of functions of *order* τ.

Hidden Units. The following classes of hidden units are used in this thesis. The input space is a subset of $\mathbb{R}^n$ and for $v, x \in \mathbb{R}^n$, $v \cdot x = \sum_{i=1}^n v_i x_i$.

- Linear threshold units: $\{g(x) = h(v \cdot x + v_0) : v \in \mathbb{R}^n, v_0 \in \mathbb{R}\}$ where $h(u) = 1$ for $u \geq 0$ and $h(u) = 0$ otherwise.
- Standard sigmoid functions: $\{g(x) = \sigma(v \cdot x + v_0) : v \in \mathbb{R}^n, v_0 \in \mathbb{R}\}$ where $\sigma(u) = 1/(1 + e^{-u})$.
- Sinusoidal basis functions: $\{g(x) = \sin(v \cdot x), g(x) = \cos(v \cdot x) : v \in \mathbb{R}^n, v_0 \in \mathbb{R}\}$.

2.3 *Other Definitions and Notations*

Function Norm. The L_1 norm of a real-valued function f defined on $\mathcal{X}$ is $\int_{\mathcal{X}} |f(x)| dx$. The L_2 norm of a real-valued function f defined on $\mathcal{X}$ is $\sqrt{\int_{\mathcal{X}} f(x)^2 dx}$. The sup-norm of a real-valued continuous function f defined on $\mathcal{X}$ is $\sup_{x \in \mathcal{X}} |f(x)|$.

Metric Spaces and Covering Number. A *pseudo-metric* on a set S is a function ρ from $S \times S$ into $\mathbb{R}^+$ (the set of real numbers greater than or equal to zero) such that for all $x, y, z \in S$, $x = y \Rightarrow \rho(x, y) = 0$, $\rho(y, x) = \rho(x, y)$ (symmetry), and $\rho(x, z) \leq \rho(x, y) + \rho(y, z)$ (triangle inequality). If $\rho(x, y) = 0 \Rightarrow x = y$, then ρ is a *metric*. (S, ρ) is a *(pseudo-) metric space*.

Let $T \subseteq S$. For a given ρ, for any $\epsilon > 0$, an ϵ-cover for T is a finite set $N \subseteq S$ (not necessarily contained in T) such that for all $x \in T$ there is a $y \in N$ with $\rho(x, y) \leq \epsilon$. The ϵ *covering number*, denoted $N(\epsilon, T, \rho)$ is the size of the smallest ϵ-cover for T using the (pseudo-) metric ρ. A set $R \subseteq T$ is ϵ*-separated* if for all distinct $x, y \in R$, $\rho(x, y) > \epsilon$. We denote the size of the largest ϵ-separated subset of T by $M(\epsilon, T, \rho)$ and refer to it as a *packing number*.

Some of the metrics and pseudo-metrics used in this thesis are described here.

- For a class of continuous functions $\mathcal{F}$ with $f, g \in \mathcal{F}$, $d_{L_\infty}(f,g) = \sup\{|f(x) - g(x)| : x \in \mathcal{X}\}$ is a metric on $\mathcal{F}$.
- Let P be a probability distribution on $\mathcal{X}$. For $f, g \in \mathcal{F}$, $d_{L_1(P)}(f,g) = \int |f(x) - g(x)| dP(x)$ is a pseudo-metric on $\mathcal{F}$.
- Let $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ and let $f_{|\boldsymbol{x}} = (f(\boldsymbol{x}_1), \ldots, f(\boldsymbol{x}_m))$ for $f \in \mathcal{F}$. Let $\mathcal{F}_{|\boldsymbol{x}} = \{f_{|\boldsymbol{x}} : f \in \mathcal{F}\}$. The following are metrics on the subset of $\mathbb{R}^m$ induced by $\boldsymbol{x}$ and $\mathcal{F}$. For $u, v \in \mathcal{F}_{|\boldsymbol{x}}$, $d_{l_1}(u,v) = \frac{1}{m}\sum_{i=1}^m |u_i - v_i|$ and $d_{l_\infty}(u,v) = \max\{|u_i - v_i| : i = 1, \ldots, m\}$.

For notational convenience, we denote these (pseudo-)metrics by their subscripts when used for covering and packing numbers e.g. $N(\epsilon, T, d_{L_\infty})$ is denoted as $N(\epsilon, T, L_\infty)$.

Closure-Convex Function Classes. Suppose $d_{L_2(P)}(f,g) = \sqrt{\int (f(x) - g(x))^2 dP(x)}$ is the pseudo-metric induced by the probability distribution P on $\mathcal{X}$. $\mathcal{F}$ is *closure-convex* if for all P on $\mathcal{X}$, the closure of $\mathcal{F}$ under the pseudo-metric $d_{L_2(P)}$ is convex. Let $\bar{\mathcal{F}}$ denote the closure of $\mathcal{F}$.

VC-dimension. Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to $\{0,1\}$ and let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathcal{X}$. We say $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are *shattered* by $\mathcal{F}$ if for each $b = (b_1, \ldots, b_m) \in \{0,1\}^m$, there is an $f \in \mathcal{F}$ such that for each i,

$$f(\boldsymbol{x}_i) = \begin{cases} 1 & \text{if } b_i = 1 \\ 0 & \text{if } b_i = 0. \end{cases}$$

The *VC-dimension* is defined as

$$\text{VCdim}(\mathcal{F}) = \max\{m \in \mathbb{N} : \exists \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m,\ \mathcal{F} \text{ shatters } \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$$

if such a maximum exists, and ∞ otherwise.

The VC-dimension was used in (Vapnik & Chervonenkis 1971) for the study of uniform convergence of relative frequencies to their probabilities.

Pseudo-dimension. Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ and let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathcal{X}$. We say $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are *pseudo-shattered* by $\mathcal{F}$ if there exists $r \in \mathbb{R}^m$ such that for each

$b = (b_1, \ldots, b_m) \in \{0,1\}^m$, there is an $f \in \mathcal{F}$ such that for each i,

$$f(\boldsymbol{x}_i) \begin{cases} \geq r_i & \text{if } b_i = 1 \\ < r_i & \text{if } b_i = 0. \end{cases}$$

The *pseudo-dimension* is defined as

$$\text{Pdim}(\mathcal{F}) = \max\{m \in \mathbb{N} \colon \exists \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m,\ \mathcal{F} \text{ pseudo-shatters } \boldsymbol{x}_1, \ldots, \boldsymbol{x}_m\}$$

if such a maximum exists, and ∞ otherwise. The pseudo-dimension is a useful generalization of the VC-dimension to real-valued functions. It is defined in this form in (Haussler 1992) and (Pollard 1990).

Fat-shattering dimension. Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ and let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m \in \mathcal{X}$. We say $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are γ-*shattered* by $\mathcal{F}$ if there exists $r \in \mathbb{R}^m$ such that for each $b = (b_1, \ldots, b_m) \in \{0,1\}^m$, there is an $f \in \mathcal{F}$ such that for each i,

$$f(\boldsymbol{x}_i) \begin{cases} \geq r_i + \gamma & \text{if } b_i = 1 \\ \leq r_i - \gamma & \text{if } b_i = 0. \end{cases}$$

For each γ, let $\text{fat}_{\mathcal{F}}(\gamma) = \max\{d \in \mathbb{N} \colon \exists \boldsymbol{x}_1, \ldots, \boldsymbol{x}_d,\ \mathcal{F}\ \gamma\text{-shatters } \boldsymbol{x}_1, \ldots, \boldsymbol{x}_d\}$ if such a maximum exists and ∞ otherwise.

The fat-shattering function was introduced in (Kearns & Schapire 1994) for the purpose of constructing lower bounds on the sample complexity for learning probabilistic concepts.

Permissible Classes of Functions. Some of the results used in this thesis requires certain measurability assumptions to be made concerning the function class $\mathcal{F}$. Following Pollard (1984) and Haussler (1992), we have indicated this by requiring these classes to be permissible.

Let $\mathcal{F}$ be a class of real-valued functions on a set $\mathcal{X}$ and let $\mathcal{A}$ be a σ-algebra of subsets of $\mathcal{X}$ such that each function in $\mathcal{F}$ is measurable. We say that the function class $\mathcal{F}$ is *permissible* if it can be indexed by a set T such that

1. T is a Borel subspace of a compact metric space S and
2. the function $f \colon \mathcal{X} \times T \to \mathbb{R}$ that indexes $\mathcal{F}$ by T is measurable with respect to the σ-algebra $\mathcal{A} \times \mathcal{B}(T)$ where $\mathcal{B}(T)$ is the σ-algebra of Borel sets on T.

More details on these conditions can be found in (Pollard 1984, Haussler 1992).

Asymptotics. Given functions f and g of p variables, we say that $f(a_1, \ldots, a_p) = O(g(a_1, \ldots, a_p))$ if there exist constants $K, \alpha_1, \ldots, \alpha_p$ such that $f(a_1, \ldots, a_p) \leq Kg(a_1, \ldots, a_p)$ for all $a_i > \alpha_i, i = 1, \ldots, p$. We say that $f(a_1, \ldots, a_p) = \Omega(g(a_1, \ldots, a_p))$ if there exist constants $K, \alpha_1, \ldots, \alpha_p$ such that $f(a_1, \ldots, a_p) \geq Kg(a_1, \ldots, a_p)$ for all $a_i > \alpha_i$, $i = 1, \ldots, p$.

A little learning is a dang'rous thing;

Drink deep, or taste not the Pierian spring;

There shallow draughts intoxicate the brain,

And drinking largely sobers us again.

— Alexander Pope,

An Essay on Criticism, 215.

Chapter 3

Upper Bounds for Sample Complexity

The sample complexity is arguably the most important component of many learning problems. Observations associated with a learning problem are often time consuming and difficult to obtain. As such it is desirable for the number of observations used to be as small as possible.

In this chapter, we study how the sample complexity scales as we require better performance from the learning algorithm. We also study how the sample complexity relates to the complexity of the function class used for learning. We review known bounds which depend on various measures of complexity of the function classes such as the covering number, pseudo-dimension and fat-shattering function. The pseudo-dimension and fat-shattering function are useful measures of complexity because they can sometimes be more easily bounded than the covering number. We also review examples of function classes with known bounds on these complexity measures. These bounds show that for many of the function classes used in practice, such as linear functions and fixed sized multilayer neural networks, the sample complexity scales reasonably (polynomially) with many of the parameters of interest such as the number of parameters parametrizing the classes and the input dimension.

We also examine the sample complexity required for agnostic learning compared to some special cases such as function learning and learning with noise. We find that better bounds can be given on the sample complexity for function learning and learning with noise when compared to the available bounds for agnostic learning. In fact, if we are restricted to hypotheses from the

Learning Problem	*Sample Complexity*
Function learning	$O\left(\frac{1}{\epsilon}\left(\ln \max_{\boldsymbol{x}\in\mathcal{X}^{2m}} N\left(\epsilon, \mathcal{F}_{\lvert\boldsymbol{x}}, l_1\right) + \ln\frac{1}{\delta}\right)\right)$
Learning with noise	$O\left(\frac{1}{\epsilon}\left(\ln \max_{\boldsymbol{x}\in\mathcal{X}^{2m}} N\left(\epsilon, \mathcal{F}_{\lvert\boldsymbol{x}}, l_1\right) + \ln\frac{1}{\delta}\right)\right)$
Agnostic learning	$O\left(\frac{1}{\epsilon^2}\left(\ln \max_{\boldsymbol{x}\in\mathcal{X}^{2m}} N\left(\epsilon, \mathcal{F}_{\lvert\boldsymbol{x}}, l_1\right) + \ln\frac{1}{\delta}\right)\right)$
Agnostic learning ($\mathcal{F}$ closure-convex)	$O\left(\frac{1}{\epsilon}\left(\ln \max_{\boldsymbol{x}\in\mathcal{X}^{2m}} N\left(\epsilon, \mathcal{F}_{\lvert\boldsymbol{x}}, l_1\right) + \ln\frac{1}{\delta}\right)\right)$
Proper Agnostic learning ($\mathcal{F}$ not closure-convex)	$\Omega\left(\frac{\ln(1/\delta)}{\epsilon^2}\right)$

Table 3.1: Sample complexity $m(1/\epsilon, 1/\delta)$ for learning with squared loss (assuming that the covering number grows polynomially with $1/\epsilon$ and does not grow with m).

target class, the better sample complexity achievable for function learning and learning with noise cannot be achieved for agnostic learning for some function classes (as we will show in Chapter 4). However, we show that if the function class satisfies the property that it is closure-convex (which is implied by convexity, see Chapter 2 for the definition of closure-convex), then a sample complexity bound similar to that for function learning and learning with noise can be achieved for agnostic learning.

All the sample complexity bounds given in this chapter can be achieved for proper learning (where the hypotheses are restricted to be from the target class). We will consider allowing learning with other hypothesis classes in Chapter 5. The constants in various bounds on the sample complexity are not the best possible. The bounds are meant to be used to relate the dependence of the sample complexity on various parameters and are not tight enough for practical purposes.

In Section 3.1, we review Haussler's work (Haussler 1992) which shows how uniform convergence of the empirical loss to the expected loss can be used with optimization algorithms to construct agnostic learning algorithms. We then give bounds on the sample complexity for uniform convergence based on the covering number of the function classes.

In Section 3.2, we consider some special cases of agnostic learning. For function learning and learning with noise, we show that the sample complexity bounds can be improved. We then show that if the function class is closure-convex, better bounds can be achieved for proper agnostic learning. These results are summarised in Table 3.1.

Finally, in Section 3.3, we give bounds on the covering number based on the pseudo-dimension and fat-shattering function. The results are summarised in Table 3.2. We also give examples of function classes for which bounds on the pseudo-dimension and fat-shattering function are known.

Learning Problem	*Pseudo-dimension*	*Fat-shattering function*
Function learning	$O\left(\frac{1}{\epsilon}\left(\mathrm{Pdim}(\mathcal{F})\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$	$O\left(\frac{1}{\epsilon}\left(\mathrm{fat}_{\mathcal{F}}(\epsilon)\ln^2\frac{\mathrm{fat}_{\mathcal{F}}(\epsilon)}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$
Learning with noise	$O\left(\frac{1}{\epsilon}\left(\mathrm{Pdim}(\mathcal{F})\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$	$O\left(\frac{1}{\epsilon}\left(\mathrm{fat}_{\mathcal{F}}(\epsilon)\ln^2\frac{\mathrm{fat}_{\mathcal{F}}(\epsilon)}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$
Agnostic learning	$O\left(\frac{1}{\epsilon^2}\left(\mathrm{Pdim}(\mathcal{F})\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$	$O\left(\frac{1}{\epsilon^2}\left(\mathrm{fat}_{\mathcal{F}}(\epsilon)\ln^2\frac{\mathrm{fat}_{\mathcal{F}}(\epsilon)}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$
Agnostic learning ($\mathcal{F}$ closure-convex)	$O\left(\frac{1}{\epsilon}\left(\mathrm{Pdim}(\mathcal{F})\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$	$O\left(\frac{1}{\epsilon}\left(\mathrm{fat}_{\mathcal{F}}(\epsilon)\ln^2\frac{\mathrm{fat}_{\mathcal{F}}(\epsilon)}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$

Table 3.2: Sample complexity $m(1/\epsilon, 1/\delta)$ for learning with squared loss for function classes based on pseudo-dimension and fat-shattering function (assuming that $\mathrm{fat}_{\mathcal{F}}(\epsilon)$ grows polynomially with $1/\epsilon$).

3.1 *Uniform Convergence and Agnostic Learning*

Most of the results in this section are based on (Haussler 1992). The following lemma shows how a uniform convergence result can be used with an optimization algorithm to construct an agnostic learning algorithm.

Recall (from Chapter 2) that $\mathbf{E}(L_f)$ is the expected value of the loss of f, $\hat{\mathbf{E}}(L_f)$ is the empirical loss of f, $opt(\mathcal{F})$ is the smallest expected loss of functions in $\mathcal{F}$ and $\hat{opt}(\mathcal{F})$ is the smallest empirical loss of functions in $\mathcal{F}$.

Lemma 3.1 *Let $\mathcal{F}$ be a function class of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$. Let $\epsilon > 0$, $0 < \delta < 1$ and suppose that the sample size $m(\epsilon, \delta)$ is such that for any probability distribution on $\mathcal{X} \times \mathcal{Y}$,*

$$\Pr\left\{\exists f \in \mathcal{F}\colon \left|\hat{\mathbf{E}}(L_f) - \mathbf{E}(L_f)\right| \geq \epsilon/3\right\} \leq \delta. \tag{3.1}$$

Suppose further that we have an algorithm which, for any sample $S \in (X \times Y)^m$, produces $\hat{f} \in \mathcal{F}$ such that

$$\left|\hat{\mathbf{E}}(L_{\hat{f}}) - \hat{opt}_S(\mathcal{F})\right| \leq \epsilon/3.$$

Then

$$\Pr\left\{\left|\mathbf{E}(L_{\hat{f}}) - opt(\mathcal{F})\right| \geq \epsilon\right\} \leq \delta.$$

Proof. By the triangle inequality, if we have

$$\left|\hat{\mathbf{E}}(L_{\hat{f}}) - \mathbf{E}(L_{\hat{f}})\right| \leq \epsilon/3, \tag{3.2}$$

$$\left|\hat{\mathbf{E}}(L_{\hat{f}}) - \hat{opt}_S(\mathcal{F})\right| \leq \epsilon/3 \tag{3.3}$$

and

$$\left| \hat{opt}_S(\mathcal{F}) - opt(\mathcal{F}) \right| \leq \epsilon/3, \tag{3.4}$$

then $\left|\mathbf{E}(L_{\hat{f}}) - opt(\mathcal{F})\right| \leq \epsilon$. From the uniform convergence assumption (3.1), both (3.2) and (3.4) hold with probability at least $1-\delta$. (If (3.4) fails, there exists a $f \in \mathcal{F}$ such that $\left|\hat{\mathbf{E}}(L_f) - \mathbf{E}(L_f)\right| \geq \epsilon/3$.) Since (3.3) comes from the assumption, the result follows. □

It follows from Lemma 3.1, that if we have a uniform convergence result, all we need is an optimization algorithm which finds a function which gives small empirical loss. For the uniform convergence results, the following results from (Haussler 1992) can be used.

Let $d_\nu(r, s) = \frac{|r-s|}{r+s+\nu}$, for $r, s \geq 0$.

Theorem 3.2 ((Haussler 1992)) *Let $\mathcal{F}$ be a permissible class of functions from $\mathcal{Z}$ to $[0, M]$. Let P be any probability distribution on $\mathcal{Z}$. For $m \geq 1$, $\nu > 0$ and $0 < \alpha < 1$,*

$$\begin{aligned} &P^m \left\{ \boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, d_\nu \left(\hat{\mathbf{E}}(f), \mathbf{E}(f) \right) > \alpha \right\} \\ &\quad \leq 4 \max_{\boldsymbol{z}' \in \mathcal{Z}^{2m}} N\left(\frac{\alpha\nu}{8}, \mathcal{F}_{|\boldsymbol{z}'}, l_1 \right) e^{-\alpha^2 \nu m / 8M}. \end{aligned} \tag{3.5}$$

The d_ν metric (see (Haussler 1992) for properties of this metric) can be used to allow both additive and multiplicative deviations from optimality and allows a better sample complexity bound to be obtained for function learning (compared to the general agnostic case). The following corollary can be used to obtain results using additive deviations from optimality which is the form we are using for our definition of agnostic learning.

Corollary 3.3 ((Haussler 1992)) *Let $\mathcal{F}$ be a permissible class of functions from $\mathcal{Z}$ to $[0, M]$. Let P be any probability distribution on $\mathcal{Z}$. For $m \geq 1$ and $0 < \epsilon < M$,*

$$\begin{aligned} &P^m \left\{ \boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, \left| \hat{\mathbf{E}}(f) - \mathbf{E}(f) \right| > \epsilon \right\} \\ &\quad \leq 4 \max_{\boldsymbol{z}' \in \mathcal{Z}^{2m}} N\left(\frac{\epsilon}{16}, \mathcal{F}_{|\boldsymbol{z}'}, l_1 \right) e^{-\epsilon^2 m / 64M^2}. \end{aligned} \tag{3.6}$$

Since we are using the quadratic loss function, we are interested in uniform convergence of the quadratic loss function class

$$Q_{\mathcal{F}} = \{(x, y) \mapsto Q_f(x, y) : f \in \mathcal{F}\},$$

where $Q_f(x, y) = (f(x) - y)^2$. When the observation and function ranges are bounded, the covering number of the quadratic loss function classes can be bounded by the covering number of the function classes as shown by Bartlett et al. (1994).

Lemma 3.4 ((Bartlett et al. 1994)) *Let $\mathcal{F}$ be a class of functions from $\mathcal{X}$ to $\mathcal{Y}$. Suppose $\mathcal{Y} \subseteq [-T, T]$ and let $Q\colon \mathcal{Y} \times \mathcal{Y} \to [0, 4T^2]$, be the quadratic loss function, $Q(y', y) := (y' - y)^2$. Then*

$$\max_{\boldsymbol{z} \in (\mathcal{X} \times \mathcal{Y})^m} N(\epsilon, Q_{\mathcal{F}|\boldsymbol{z}}, l_1) \leq \max_{\boldsymbol{x} \in \mathcal{X}^m} N\left(\frac{\epsilon}{6T}, \mathcal{F}_{|\boldsymbol{x}}, l_1\right).$$

We can now bound the sample complexity for agnostic learning based on the covering number of the function class.

Corollary 3.5 *Let $\mathcal{F}$ be a permissible class of functions from $\mathcal{X}$ to $\mathcal{Y}$. Let $\mathcal{Y} \subseteq [-T, T]$ and let $Q\colon \mathcal{Y} \times \mathcal{Y} \to [0, 4T^2]$, be the quadratic loss function, $Q(y', y) := (y' - y)^2$. Then $\mathcal{F}$ is agnostically learnable from m observations, provided*

$$m \geq \frac{9216T^4}{\epsilon^2}\left(\ln\left(\max_{\boldsymbol{x} \in \mathcal{X}^{2m}} N\left(\frac{\epsilon}{288T}, \mathcal{F}_{|\boldsymbol{x}}, l_1\right)\right) + \ln\frac{4}{\delta}\right). \tag{3.7}$$

Proof. Assume that we have an optimization algorithm which can provide a hypothesis with empirical loss less than $\epsilon/3$. A suitable mapping always exists. From Lemma 3.1, uniform convergence to accuracy $\epsilon/3$ suffices for agnostic learning. Setting the right hand side of (3.6) to δ and using Lemma 3.4 we obtain (3.7). $\square$

3.2 Improving the Sample Complexity

In this section, we give some special cases of agnostic learning which allow improved bounds on the sample complexity. We also give improved bounds for closure convex function classes.

3.2.1 Function Learning

For function learning, Theorem 3.2 can be used to obtain a better bound for the sample complexity. Setting $\nu = \epsilon$ and $\alpha = 1/2$, we get $\mathbf{E}(Q_f) \leq 3\hat{\mathbf{E}}(Q_f) + \epsilon$. For function learning, it is possible to set the empirical loss to zero by choosing an appropriate function $\hat{f}$, hence giving $\mathbf{E}(Q_{\hat{f}}) \leq \epsilon$.

With these values of α and ν together with Theorem 3.2 and Lemma 3.4, a sample size of

$$m \geq \frac{128T^2}{\epsilon}\left(\ln\left(\max_{\boldsymbol{x}\in\mathcal{X}^{2m}} N\left(\frac{\epsilon}{96T}, \mathcal{F}_{|\boldsymbol{x}}, l_1\right)\right) + \ln\frac{4}{\delta}\right)$$

suffices for agnostically learning $\mathcal{F}$.

3.2.2 Learning with Noise

For learning with noise, Barron (1990) and McCaffrey & Gallant (1994) have shown that the sample complexity for functions with finite L_∞ covering number is $O\left(\frac{1}{\epsilon}\left(\ln N(\epsilon, \mathcal{F}, L_\infty) + \ln\frac{1}{\delta}\right)\right)$. The L_∞ covering number is always at least at large as the l_1 covering number but may be considerably larger. For example, the class of sigmoid functions without a bound on the input weight size has a finite l_1 covering number (see Section 3.3) but cannot have a finite L_∞ cover. (It is easy to see that for any finite set of functions, we can always find a sigmoid function, with distance close to $1/2$ from all the functions in the set by considering linear threshold functions which can be approximated arbitrarily closely by sigmoid functions.) We extend the result for learning with noise (Barron 1990, McCaffrey & Gallant 1994) to function classes with finite l_1 covering numbers by using the following theorem.

Theorem 3.6 *Let $\mathcal{F}$ be a permissible class of functions mapping from $\mathcal{X}$ to $\mathcal{Y} \subseteq [-T, T]$. Let P be an arbitrary probability distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $C = \max\{T, 1\}$. Assume $\nu, \nu_c > 0, 0 < \alpha \leq 1/2$. Let $f^* \in \mathcal{F}$ where $f^*(x) = \mathbf{E}[Y|X = x]$ and $g_f(x, y) = (y - f(x))^2 - (y - f^*(x))^2$. Then for $m \geq 1$,*

$$P^m\left\{\boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, \frac{\mathbf{E}(g_f) - \hat{\mathbf{E}}_{\boldsymbol{z}}(g_f)}{\nu + \nu_c + \mathbf{E}(g_f)} \geq \alpha\right\}$$
$$\leq \max_{\boldsymbol{x}\in\mathcal{X}^{2m}} 6N\left(\frac{\alpha\nu_c}{128C^3}, \mathcal{F}_{|\boldsymbol{x}}, l_1\right)\exp(-\alpha^2\nu m/(875C^4)). \quad (3.8)$$

The proof is included in Appendix A. The main idea (in addition to the ideas of the proof of Theorem 3.2) is to bound the variance of the random variable $g_f(X, Y)$ in terms of its expectation and to use Bernstein's inequality to take advantage of the variance bound.

To obtain a bound on the sample complexity, we first rescale the function class and target random variable by dividing by T to give $C = 1$ and consider the new learning problem. (This rescaling trick allows us to obtain a sample complexity which has a T^2 term instead of a T^4 term.) The ϵ covering number of the scaled function class is the same as the $T\epsilon$ covering number of the unscaled function class. To get the correct accuracy when the function class is scaled back to the original scale, we need to learn to accuracy ϵ/T^2. Assume the scaled function class is $\mathcal{F}$. Setting $\nu = \nu_c = \epsilon/(2T^2)$, $\alpha = 1/2$ and the right hand side of (3.8) to δ, we get with probability $1 - \delta$, $\mathbf{E}(g_f) \leq 2\hat{\mathbf{E}}_{\boldsymbol{z}}(g_f) + \epsilon/T^2$ for all $f \in \mathcal{F}$. From the definition of g_f, notice that it is possible to choose $\hat{f}$ such that $\hat{\mathbf{E}}_{\boldsymbol{z}}(g_{\hat{f}}) \leq 0$ (since it is possible to choose the function giving the best empirical loss which is no more than the empirical loss for f^*) giving $\mathbf{E}(g_{\hat{f}}) \leq \epsilon/T^2$. Setting the right hand side of (3.8) to δ and solving for m shows that

$$m \geq \frac{7000T^2}{\epsilon}\left(\ln\left(\max_{\boldsymbol{x}\in\mathcal{X}^{2m}} N\left(\frac{\epsilon}{512T}, \mathcal{F}_{|\boldsymbol{x}}, l_1\right) + \ln\frac{6}{\delta}\right)\right)$$

observations suffices for agnostically learning the function class.

3.2.3 Agnostic Learning of Closure-Convex Function Classes

Given that it is possible to obtain better sample complexity (with respect to ϵ) for the special cases of function learning and learning with noise, we would also like to investigate the possibilities for the more general agnostic case. However, better sample complexity is not possible without some conditions on the function class if we are restricted to hypotheses from the same class. For example, consider the class of functions which consists only of $f_1(x) \equiv 0$ and $f_2(x) \equiv 1$. Let the target be a $\{0, 1\}$ random variable which is 1 with probability p and 0 with probability $1 - p$. The sample complexity for properly learning this function class with this type of target is $\Omega\left(\frac{\ln 1/\delta}{\epsilon^2}\right)$ (see Lemma 4.5).

However, with closure-convex function classes, it is possible to obtain the same sample complexity bound as the case for learning with noise. This is done by using the following theorem. The proof of the theorem is given in the Appendix A. The convexity of the function class allows us to bound the variance of the random variable $g_f(X, Y)$ in terms of its expectation hence giving the better bound. The theorem is given in a more general form which is useful for learning the convex hull of function classes in Chapter 5.

Theorem 3.7 *Let $\mathcal{F} = \bigcup_{k=1}^{\infty} \mathcal{F}_k$ be a closure-convex class of functions mapping from $\mathcal{X}$ to $\mathcal{Y} \subseteq [-T, T]$ such that each $\mathcal{F}_k$ is permissible. Let P be an arbitrary probability distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $\bar{\mathcal{F}}$ be the closure of $\mathcal{F}$ in the space with inner product $\langle f, g\rangle = \int f(x)g(x)dP_{\mathcal{X}}(x)$. Let $C = \max\{T, 1\}$. Assume $\nu, \nu_c > 0, 0 < \alpha \leq 1/2$. Let $f^*(x) = \mathbf{E}[Y|X = x]$ and $g_f(x, y) = (y - f(x))^2 - (y - f_a(x))^2$ where $f_a \in \bar{\mathcal{F}}$ and $f_a = \mathrm{argmin}_{f\in\bar{\mathcal{F}}} \int (f(x) - f^*(x))^2 dP_{\mathcal{X}}(x)$. Then* *for $m \geq 1$ and each k,*

$$P^m \left\{ \boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}_k, \frac{\mathbf{E}(g_f) - \hat{\mathbf{E}}_{\boldsymbol{z}}(g_f)}{\nu + \nu_c + \mathbf{E}(g_f)} \geq \alpha \right\}$$
$$\leq \max_{\boldsymbol{x} \in \mathcal{X}^{2m}} 6N \left(\frac{\alpha \nu_c}{128 C^3}, \mathcal{F}_{k|\boldsymbol{x}}, l_1 \right) \exp(-\alpha^2 \nu m / (875 C^4)). \tag{3.9}$$

Let $\mathcal{F}_k = \mathcal{G}$ for $k = 1, \ldots, \infty$ where $\mathcal{G}$ is a closure-convex class of functions (hence $\mathcal{F} = \mathcal{G}$). As in the learning with noise case, we first rescale the function class and target random variable by dividing by T to give $C = 1$ and consider the new learning problem. The ϵ covering number of the scaled function class is the same as the $T\epsilon$ covering number of the unscaled function class. To get the correct accuracy when the function class is scaled back to the original scale, we need to learn to accuracy ϵ/T^2. Assume the scaled function class is $\mathcal{F}$. Set $\nu = \nu_c = \epsilon/2T^2$ and $\alpha = 1/2$, and set the right hand side of (3.9) to δ to get with probability $1 - \delta$, $\mathbf{E}(g_f) \leq 2\hat{\mathbf{E}}_{\boldsymbol{z}}(g_f) + \epsilon/T^2$ for all $f \in \mathcal{G}$. From the definition of g_f, again it is possible to choose $\hat{f}$ such that $\hat{\mathbf{E}}_{\boldsymbol{z}}(g_{\hat{f}}) \leq 0$ giving $\mathbf{E}(g_{\hat{f}}) \leq \epsilon/T^2$. Setting the right hand side of (3.8) to δ and solving for m shows that

$$m \geq \frac{7000 T^2}{\epsilon} \left(\ln \left(\max_{\boldsymbol{x} \in \mathcal{X}^{2m}} N \left(\frac{\epsilon}{512 T}, \mathcal{F}_{|\boldsymbol{x}}, l_1 \right) + \ln \frac{6}{\delta} \right) \right)$$

observations suffices for agnostically learning the function class.

3.3 Bounding the Covering Number

In order to use the results in the previous section, we need to bound the l_1 covering number of various function classes. For many classes of functions, this can be done using properties of the function classes called the the pseudo-dimension and the fat-shattering function (see Chapter 2 for the definitions of these properties). The following lemmas give bounds for the covering number of function classes in terms of these properties.

Lemma 3.8 ((Pollard 1984, Haussler 1992)) *Let $\mathcal{F}$ be a class of functions from a set $\mathcal{X}$ into $[-T, T]$ and suppose* $\mathrm{Pdim}(\mathcal{F}) = d$ *for some* $1 \leq d < \infty$*. Then for all* $0 < \epsilon \leq 2T$ *and any finite sequence $\boldsymbol{x}$ of points in $\mathcal{X}$,*

$$N(\epsilon, \mathcal{F}_{|\boldsymbol{x}}, l_1) < 2 \left(\frac{4eT}{\epsilon} \ln \frac{4eT}{\epsilon} \right)^d .$$

The following result follows from a result in (Alon, Ben-David, Cesa-Bianchi & Haussler 1993). The proof is given in the Appendix A. A slightly better result (without d but including m in the log term) is given in (Bartlett & Long 1995).

Lemma 3.9 *Let* $0 < \epsilon \leq 2T$ *and let* $\mathcal{F}$ *be a family of functions from a set* $\mathcal{X}$ *into* $[-T, T]$ *such that* $0 < d = \mathrm{fat}_{\mathcal{F}}(\epsilon/(8T)) \leq \infty$. *For any finite sequence* $\boldsymbol{x}$ *of points in* $\mathcal{X}$,

$$N(\epsilon, \mathcal{F}_{|\boldsymbol{x}}, l_1) < \exp\left(\frac{8d}{\ln 2}\ln^2\left(\frac{2048T^4 d}{\epsilon^4 \ln 2}\right)\right).$$

Corollary 3.10 *Let* $\mathcal{F}$ *be a permissible class of functions mapping from* $\mathcal{X}$ *to* $\mathcal{Y}$. *Let the observed range* $[-T, T] \supseteq \mathcal{Y}$ *where* $T \geq 1$, *and let* $d = \mathrm{Pdim}(\mathcal{F})$. *Then*

1. *The sample complexity for agnostic learning using* $\mathcal{F}$ *is bounded from above by*

$$\frac{9216T^4}{\epsilon^2}\left(d \ln\left(\frac{1152eT^2}{\epsilon}\ln\frac{1152eT^2}{\epsilon}\right) + \ln\frac{8}{\delta}\right).$$

2. *The sample complexity for function learning using* $\mathcal{F}$ *is bounded from above by*

$$\frac{128T^2}{\epsilon}\left(d \ln\left(\frac{384eT^2}{\epsilon}\ln\frac{384eT^2}{\epsilon}\right) + \ln\frac{8}{\delta}\right).$$

3. *The sample complexity for learning with noise using* $\mathcal{F}$ *or for agnostic learning if* $\mathcal{F}$ *is closure-convex is bounded from above by*

$$\frac{7000T^2}{\epsilon}\left(d \ln\left(\frac{2048eT^2}{\epsilon}\ln\frac{2048eT^2}{\epsilon}\right) + \ln\frac{12}{\delta}\right).$$

Corollary 3.11 *Let* $\mathcal{F}$ *be a permissible class of functions mapping from* $\mathcal{X}$ *to* $\mathcal{Y}$. *Let the observed range* $[-T, T] \supseteq \mathcal{Y}$ *where* $T \geq 1$, *and assume* $\mathrm{fat}_{\mathcal{F}}(\gamma) < \infty$ *for all* $\gamma > 0$. *Then*

1. *The sample complexity for agnostic learning using* $\mathcal{F}$ *is*

$$O\left(\frac{T^4}{\epsilon^2}\left(\mathrm{fat}_{\mathcal{F}}(\epsilon/(8T))\ln^2\left(\frac{T\mathrm{fat}_{\mathcal{F}}(\epsilon/(8T))}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right).$$

2. *The sample complexity for function learning using* $\mathcal{F}$ *is*

$$O\left(\frac{T^2}{\epsilon}\left(\mathrm{fat}_{\mathcal{F}}(\epsilon/(8T))\ln^2\left(\frac{T\mathrm{fat}_{\mathcal{F}}(\epsilon/(8T))}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right).$$

3. *The sample complexity for learning with noise using $\mathcal{F}$ or for agnostic learning if $\mathcal{F}$ is closure-convex is*

$$O\left(\frac{T^2}{\epsilon}\left(\text{fat}_{\mathcal{F}}(\epsilon/(8T))\ln^2\left(\frac{T\text{fat}_{\mathcal{F}}(\epsilon/(8T))(\mathcal{F})}{\epsilon}\right)+\ln\frac{1}{\delta}\right)\right).$$

3.3.1 Function classes with known dimension bounds

In this section we give examples of function classes with known bounds on the pseudo-dimension or the fat-shattering function.

Dudley (1978) has shown that the pseudo-dimension of a d-dimensional vector space of functions from a set $\mathcal{X}$ to $\mathbb{R}$ is d. This gives the pseudo-dimension of linear functions as well as linear combinations of fixed basis functions such as polynomials. Pollard (1990) gives useful invariance properties of the pseudo-dimension including the fact that if each function in a function class is composed with the same non-decreasing function, the resulting function class cannot have a larger pseudo-dimension than the original function class. This gives a bound of $d+1$ for the pseudo-dimension of the sigmoid function in d dimensions. (A sigmoid function is just a linear function with a bias composed with an increasing function.) Goldberg & Jerrum (1993) and Maass (1995) have shown that multilayer neural networks with piecewise polynomial functions have pseudo-dimension bounded by $O(W^2)$ where W is the number of adjustable parameters. Karpinski & Macintyre (1995) have shown that the pseudo-dimension of multilayer neural networks with standard sigmoid activation functions is bounded by $O(W^4)$.

It is easy to see that the fat-shattering function of the class of non-decreasing functions with a bounded range is $O(1/\gamma)$. Gurvits & Koiran (1995) have shown that single hidden layer neural networks with linear threshold hidden units and bounded sum of absolute values of output weights have fat-shattering function bounded by $O\left(\frac{n^2}{\gamma^2}\ln\frac{n^2}{\gamma^2}\right)$, where n is the input dimension.

From these examples, we see that for many commonly used function classes, the sample complexity grows slowly (polynomially) with many of the complexity parameters of interest such as the input dimension and the number of parameters parametrizing the class.

Some for renown, on scraps of learning dote,

And think they grow immortal as they quote.

— Edward Young

Love of Fame, Satire i, 89.

Chapter 4

Lower Bounds for Sample Complexity

In this chapter, we give lower bounds on the sample complexity for agnostic learning.

In Section 4.1, we first review a result of Bartlett et al. (1994) which gives a lower bound on the sample complexity for agnostic learning (with the absolute loss function) based on the fat-shattering function. This shows that if the fat-shattering function $\text{fat}_{\mathcal{F}}(\gamma)$ is infinite for some γ, then the function class is not agnostically learnable (with the absolute loss function). We then observe that this is true for the squared loss function as well. It is easy to see that the class of single hidden layer neural networks with an unbounded number of hidden units and no bound on the magnitude of the output weights (with sigmoid or linear threshold hidden units) has infinite fat-shattering function. Hence, for single hidden layer neural networks with an unbounded number of hidden units to be agnostically learnable, some constraints on the output weights are necessary. We will show that a bound on the sum of the absolute values of the output weights is sufficient (and give the sample complexity) in Chapter 5.

In Chapter 3, we showed that if the function class is closure-convex, then the sample complexity for agnostic learning is $O\left(\frac{1}{\epsilon}\left(\ln\left(\max_{\boldsymbol{x}\in\mathcal{X}^{2m}} N\left(\epsilon, \mathcal{F}_{|\boldsymbol{x}}, l_1\right)\right) + \ln\frac{1}{\delta}\right)\right)$. In Section 4.2, we give a partial converse for this result by showing that if the function class is not closure-convex, then the sample complexity for *proper* agnostic learning is $\Omega(\ln(1/\delta)/\epsilon^2)$. This shows that if the logarithm of the covering number of a function class grows slower than $1/\epsilon$, then the sample complexity for learning with noise is better than the sample complexity for proper agnostic learning. Lemma 3.8 shows that function classes with finite pseudo-dimension satisfy this growth condition.

4.1 A Lower Bound based on the Fat-Shattering Function

In (Bartlett et al. 1994), it was shown that *efficient* agnostic learning of a function class with the absolute loss function is possible only if the fat-shattering function of the function class grows at most polynomially with $1/\epsilon$ and the relevant complexity parameters. With minor modifications to the proof in (Bartlett et al. 1994), it is possible to show that this is also true for the quadratic loss function.

Definition 4.1 *For $\alpha \in \mathbb{R}^+$, define the quantization function*

$$\Delta_\alpha(y) = \alpha \left\lceil \frac{y - \alpha/2}{\alpha} \right\rceil .$$

For a set $S \subset \mathbb{R}$, let $\Delta_\alpha(S) = \{\Delta_\alpha(y) \colon y \in S\}$. For a function class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$, let $\Delta_\alpha(\mathcal{F})$ be the set $\{\Delta_\alpha \circ f \colon f \in \mathcal{F}\}$ of $\Delta_\alpha([0,1])$-valued functions defined on $\mathcal{X}$.

Lemma 4.2 ((Bartlett et al. 1994)) *Let $\alpha \in \mathbb{R}^+$. Choose a set $\mathcal{F}$ of functions from $\mathcal{X}$ to $\Delta_\alpha([0,1])$, $d > 400$ and $\gamma > 0$ such that $\mathrm{fat}_{\mathcal{F}}(\gamma) \geq d$. With fewer than*

$$\frac{d - 400}{4 + 192 \ln\lceil 1/\alpha \rceil}$$

examples, there is no algorithm which can produce a hypothesis with expected absolute loss less than $\gamma/32$ with probability at least $1/16$.

The following is essentially from (Bartlett et al. 1994) with minor modifications so that the squared loss is used in place of the absolute loss function.

Theorem 4.3 *Let $\mathcal{F}$ be a class of $[0,1]$-valued functions defined on $\mathcal{X}$. Suppose $0 < \gamma < 1$, $0 < \epsilon \leq \gamma/65$, $0 \leq \delta \leq 1/16$ and $d \in \mathbb{N}$. If $\mathrm{fat}_{\mathcal{F}}(\gamma) \geq d > 800$, then with the quadratic loss function, no algorithm can agnostically learn $\mathcal{F}$ to accuracy $3\epsilon^2$ with probability $1 - \delta$ with fewer than*

$$m > \frac{d}{400 \log \frac{40}{\gamma}}$$

observations.

Proof. Set $\epsilon = \gamma/65$, $\delta = 1/16$. Consider the class of distributions on $\mathcal{X} \times [0,1]$ for which there exists an $f \in \mathcal{F}$ such that, for all $x \in \mathcal{X}$,

$$P(Y|x) = \begin{cases} 1 & \text{if } Y = \Delta_{2\epsilon}(f(x)) \\ 0 & \text{otherwise.} \end{cases}$$

Fix a distribution P in this class. Let h be the hypothesis produced by an algorithm that can agnostically learn $\mathcal{F}$ to accuracy $3\epsilon^2$ with probability $1-\delta$. Then

$$\Pr(\mathbf{E}[Q_h] \geq \inf_{f\in\mathcal{F}} \mathbf{E}[Q_f] + 3\epsilon^2) < \delta$$

where Q is the quadratic loss. By definition $\inf_{f\in\mathcal{F}} \mathbf{E}[Q_f] \leq \epsilon^2$. So

$$\Pr(\mathbf{E}[Q_h] < 4\epsilon^2) > 1 - \delta.$$

By the Cauchy-Schwartz inequality,

$$\mathbf{E}[Q_h] < 4\epsilon^2 \Rightarrow E[\Lambda_h] < 2\epsilon$$

where Λ is the absolute loss function. Hence the algorithm can learn the quantized function class $\Delta_{2\epsilon}(\mathcal{F})$ to accuracy 2ϵ with probability $1-\delta$. By hypothesis, $\text{fat}_{\mathcal{F}}(\gamma) \geq d$, so $\text{fat}_{\Delta_{2\epsilon}(\mathcal{F})}(\gamma-\epsilon) \geq d$. Since $\epsilon \leq \gamma/65$, $2\epsilon \leq (\gamma-\epsilon)/32$. Also, $\delta = 1/16$, so Lemma 4.2 implies

$$\begin{aligned} m &> \frac{d - 400}{4 + 192\ln\lceil 1/(2\epsilon)\rceil} \\ &> \frac{d}{8 + 384\log(65/(2\gamma))} \\ &> \frac{d}{400\log(40/\gamma)}. \end{aligned}$$

$\square$

Note that a $[-T, T]$-valued function class can be transformed into a $[0,1]$-valued function class by adding T to the function class then dividing by $2T$. With ϵ and γ similarly transformed into $2\epsilon T$ and $2\gamma T$, the lower bound holds for $[-T, T]$-valued functions.

The lower bound also implies that if a function class has infinite fat-shattering function, then it is not agnostically learnable.

4.2 Lower Bounds for Classes which are not Closure-Convex

In this section we give a lower bound on the sample complexity for agnostic learning with squared loss.

Theorem 4.4 *Let $\mathcal{F}$ be a class of functions mapping from $\mathcal{X}$ to $\mathcal{Y}$. If $\mathcal{F}$ is not closure-convex, then the sample complexity for agnostically learning $\mathcal{F}$ with squared loss is $\Omega\left(\frac{\ln(1/\delta)}{\epsilon^2}\right)$.*

The idea behind the proof is to show that if the closure of $\mathcal{F}$ is not convex, an agnostic algorithm for learning $\mathcal{F}$ to accuracy ϵ can be used to estimate the expected value of a Bernoulli random variable to accuracy $k\epsilon$ for some constant k using the same number of observations. Since, as we now show, estimating the expected value of a Bernoulli random variable requires $\Omega(\ln(1/\delta)/\epsilon^2)$ observations, the agnostic learning algorithm also requires $\Omega(\ln(1/\delta)/\epsilon^2)$ observations.

Lemma 4.5 *Let $\xi_1, \ldots, \xi_m$ be a sequence of i.i.d. $\{0,1\}$-valued random variables where $\Pr(\xi_i = 1) = \alpha$ where α can take the value $\alpha_1 = 1/2 + \epsilon/2$ with probability $1/2$ and $\alpha_2 = 1/2 - \epsilon/2$ with probability $1/2$. Deciding the value of α correctly with probability at least $1-\delta$ requires a sample of size $m = \Omega\left(\frac{\ln 1/\delta}{\epsilon^2}\right)$.*

Proof. The decision rule which minimizes the probability of choosing the wrong α is to choose α_1 when half or more of the sample is 1 and α_2 otherwise (Fukunaga 1972) (assuming ties are broken in favour of α_1). We show that with such a rule, if m is less than $\Omega(\ln(\delta)/\epsilon^2)$, then for small enough ϵ and δ, the probability of choosing the wrong α is greater than δ.

We require m such that

$$\frac{1}{2}\sum_{i<m/2}^{m}\binom{m}{i}\left(\frac{1}{2}+\frac{\epsilon}{2}\right)^{i}\left(\frac{1}{2}-\frac{\epsilon}{2}\right)^{m-i}+\frac{1}{2}\sum_{i\geq m/2}^{m}\binom{m}{i}\left(\frac{1}{2}-\frac{\epsilon}{2}\right)^{i}\left(\frac{1}{2}+\frac{\epsilon}{2}\right)^{m-i}>\delta.$$

This will be satisfied if one term (choose $\frac{1}{2}\binom{m}{m/2}\left(\frac{1}{4}-\frac{1}{4}\epsilon^2\right)^{m/2}$ and assume m even for convenience) is larger than δ. Using Stirling's approximation, $\binom{m}{m/2}$ is approximately $2^{m+1}/\sqrt{m}$ for m large enough. Hence $\frac{1}{2}\binom{m}{m/2}\left(\frac{1}{4}-\frac{1}{4}\epsilon^2\right)^{m/2} \sim (1-\epsilon^2)^{m/2}/\sqrt{m}$. Hence for small enough ϵ and δ, there will be some k_1 and k_2 (both positive) such that

$$\ln\left(\frac{1}{2}\binom{m}{m/2}\left(\frac{1}{4}-\frac{1}{4}\epsilon^2\right)^{m/2}\right) > k_1 m\ln(1-\epsilon^2) > -k_2 m\epsilon^2.$$

Hence for $m < k_2\ln(1/\delta)/\epsilon^2$, the probability is greater than δ. □

We now give some results on function classes which are not closure-convex. They are used in the proof of Theorem 4.4.

The following lemma shows that if $\bar{\mathcal{F}}$ is not convex, there is a ball touching at least two points in $\bar{\mathcal{F}}$ whose interior does not intersect $\bar{\mathcal{F}}$.

Lemma 4.6 *Let $\mathcal{F}$ be a subset of a Hilbert space H. If $\bar{\mathcal{F}}$ is not convex then there exist $f_1, f_2 \in \bar{\mathcal{F}}$, $c \in H$ such that $f_1 \neq f_2$, $\|c - f_1\| = \|c - f_2\| > 0$ and $\{f \in \bar{\mathcal{F}} : \|c - f\| < \|c - f_1\|\} = \emptyset$.*

Proof. Since $\bar{\mathcal{F}}$ is closed and not convex, there exists $g, h \in \bar{\mathcal{F}}$, $\alpha \in (0, 1)$ and $\delta > 0$ such that $f_c = \alpha g + (1 - \alpha)h$ is not in $\bar{\mathcal{F}}$ and $\{f \in \bar{\mathcal{F}} : \|f - f_c\| < \delta\} = \emptyset$. Let $\delta' = \min\{\delta : \{f \in \bar{\mathcal{F}} : \|f - f_c\| \leq \delta\} \neq \emptyset\}$. If the set $G = \{f \in \bar{\mathcal{F}} : \|f - f_c\| = \delta'\}$ contains more than one function, we are done. If G contains only one function f_1, setting $c = tf_c + (1 - t)f_1$ with the smallest $t > 1$ such that $\{f \in \bar{\mathcal{F}} : f \neq f_1, \|f - c\| = \|f_1 - c\|\} \neq \emptyset$ gives the required result, provided such a t exists. We show that such a t must exist. Now $\|f_1 - c\|^2 = t^2\|f_1 - f_c\|^2$ and

$$\begin{aligned}
\|h - c\|^2 &= \|h - f_1\|^2 + t^2\|f_1 - f_c\|^2 + 2t\langle h - f_1, f_1 - f_c\rangle \\
&= \|h - f_1\|^2 + t^2\|f_1 - f_c\|^2 + 2t\langle h - f_c + f_c - f_1, f_1 - f_c\rangle \\
&= \|h - f_1\|^2 + t^2\|f_1 - f_c\|^2 - 2t\|f_1 - f_c\|^2 + 2t\langle h - f_c, f_1 - f_c\rangle \\
&= \|h - f_1\|^2 + \|f_1 - c\|^2 - 2t\|f_1 - f_c\|^2 + 2t\langle h - f_c, f_1 - f_c\rangle.
\end{aligned}$$

Similarly, $\|g - c\|^2 = \|g - f_1\|^2 + \|f_1 - c\|^2 - 2t\|f_1 - f_c\|^2 + 2t\langle g - f_c, f_1 - f_c\rangle$. Now $\langle h - f_c, f_1 - f_c\rangle$ and $\langle g - f_c, f_1 - f_c\rangle$ must have opposite sign unless they are both zero. In any case, for t large enough, either $\|f_1 - c\| \geq \|h - c\|$ or $\|f_1 - c\| \geq \|g - c\|$ or both. Since g and h belong to $\bar{\mathcal{F}}$ this completes the proof. $\square$

Again suppose $\mathcal{F} \subseteq H$ such that $\bar{\mathcal{F}}$ is not convex. Let $f_1, f_2 \in \bar{\mathcal{F}}$ and $c \in H$ be as in Lemma 4.6. (Figure 4.1 shows f_1, f_2, c in a two-dimensional slice through H.) Define the sets $\mathcal{B} := \{f \in H : \|f - c\| = \|f_1 - c\|\}$ and $\mathcal{B}_F := \bar{\mathcal{F}} \cap \mathcal{B}$. ($\mathcal{B}$ contains the circle in Figure 4.1, and $\mathcal{B}_{\mathcal{F}}$ contains f_1 and f_2.) Pick $f_m \in \mathcal{B}$ in the hyperplane $\{f \in H : \langle f_1 - c, f - c\rangle = \langle f_2 - c, f - c\rangle\}$ such that $\langle f_1 - c, f_m - c\rangle$ is maximized. Choose a two dimensional plane P through c, f_1, f_2 and f_m. P is illustrated in Figure 4.1. Let $f_{d1}, f_{d2} \in P$ be such that $f_{d1} - c$ and $f_{d2} - c$ are orthogonal to $f_m - c$, and $\|f_{d1} - f_1\| < \|f_{d_1} - f_2\|$.

For $0 < p < 1$ define $f_1^* = pf_1 + (1 - p)c$ and $f_2^* = pf_2 + (1 - p)c$. Define γ by $f_1^* = (f_1^* + f_2^*)/2 + \gamma(f_{d1} - c)$. It is easy to show that $f_2^* = (f_1^* + f_2^*)/2 - \gamma(f_{d1} - c)$. The

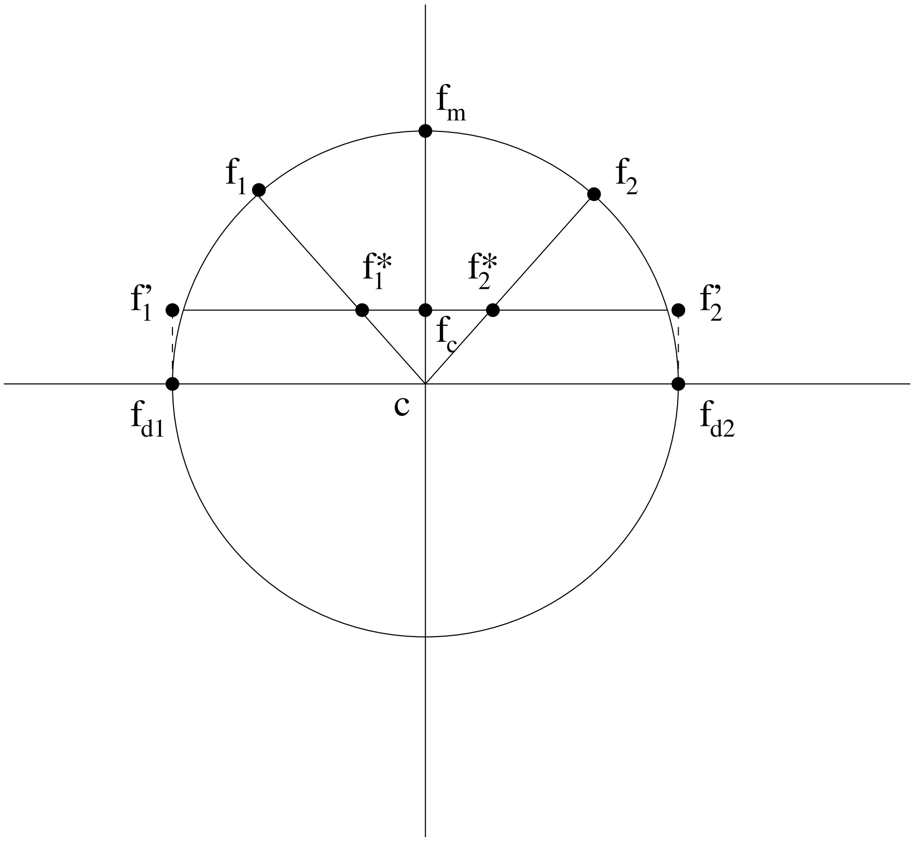

Figure 4.1: Function class with labelled functions schematically represented in two dimensions.

following claim relates γ to p, and to $\epsilon := \|f_m - f_1^*\|^2 - \|f_1 - f_1^*\|^2$.

Claim 4.7

$$\gamma = \frac{p\langle f_1 - c, f_{d1} - c\rangle}{\|f_{d1} - c\|^2} = \frac{\epsilon\langle f_1 - c, f_{d1} - c\rangle}{2\left(1 - \frac{\langle f_1 - c, f_m - c\rangle}{\|f_m - c\|^2}\right)\|f_m - c\|^4}.$$

Proof. Let $f_c = (f_1^* + f_2^*)/2$. Then

$$\begin{aligned} f_1^* - f_c &= \gamma(f_{d1} - c) \\ &= \frac{\langle pf_1 + (1-p)c - c, f_{d1} - c\rangle}{\|f_{d1} - c\|\|f_{d1} - c\|}(f_{d1} - c) \\ &= \frac{p\langle f_1 - c, f_{d1} - c\rangle}{\|f_{d1} - c\|^2}(f_{d1} - c) \end{aligned} \tag{4.1}$$

which gives the first equality. To prove the second, first notice that

$$\begin{aligned} f_c - c &= \frac{\langle f_1^* - c, f_m - c\rangle}{\|f_m - c\|^2}(f_m - c) \\ &= \frac{\langle pf_1 + (1-p)c - c, f_m - c\rangle}{\|f_m - c\|^2}(f_m - c) \end{aligned}$$

$$= \frac{p\langle f_1 - c, f_m - c\rangle}{||f_m - c||^2}(f_m - c) \tag{4.2}$$

and

$$||f_1 - f_1^*||^2 = ||f_1 - pf_1 - (1-p)c||^2 = (1-p)^2||f_1 - c||^2 = (1 - 2p + p^2)||f_1 - c||^2. \tag{4.3}$$

With that, by Pythagoras Theorem

$$\begin{aligned} ||f_m - f_1^*||^2 &= ||f_1^* - f_c||^2 + ||f_m - f_c||^2 \\ &= ||f_1^* - f_c||^2 + ||f_m - c + c - f_c||^2 \\ &= \frac{p^2\langle f_1 - c, f_{d1} - c\rangle^2}{||f_{d1} - c||^2} + ||f_m - c||^2 + \frac{p^2\langle f_1 - c, f_m - c\rangle^2}{||f_m - c||^2} + 2\langle f_m - c, c - f_c\rangle \\ &= p^2||f_1 - c||^2 + ||f_m - c||^2 - \frac{2p\langle f_1 - c, f_m - c\rangle}{||f_m - c||^2}||f_m - c||^2, \end{aligned} \tag{4.4}$$

where the third equality follows from (4.1) and (4.2).

Note that from the construction, $||f_1 - c|| = ||f_m - c|| = ||f_{d1} - c||$. From the definition of ϵ, we see that taking (4.3) from (4.4), we get

$$2p\left(1 - \frac{\langle f_1 - c, f_m - c\rangle}{||f_m - c||^2}\right)||f_m - c||^2 = \epsilon$$

$$p = \frac{\epsilon}{2\left(1 - \frac{\langle f_1 - c, f_m - c\rangle}{||f_m - c||^2}\right)||f_m - c||^2}. \tag{4.5}$$

From (4.1) and (4.5),

$$\begin{aligned} \gamma &= \frac{p\langle f_1 - c, f_{d1} - c\rangle}{||f_{d1} - c||^2} \\ &= \frac{\epsilon}{2\left(1 - \frac{\langle f_1 - c, f_m - c\rangle}{||f_m - c||^2}\right)||f_m - c||^2}\frac{\langle f_1 - c, f_{d1} - c\rangle}{||f_{d1} - c||^2} \\ &= \frac{\epsilon\langle f_1 - c, f_{d1} - c\rangle}{2\left(1 - \frac{\langle f_1 - c, f_m - c\rangle}{||f_m - c||^2}\right)||f_m - c||^4}. \end{aligned} \tag{4.6}$$

□

The following lemma can be used to show how an agnostic learning algorithm can be used for selecting between f_1^* and f_2^* when either function can be the target conditional expectation.

Lemma 4.8 *Suppose f_1, f_2, f_1^*, f_2^*, ϵ and the function class $\mathcal{F}$ are as defined above. Then for any $\hat{f} \in \mathcal{F}$ and $\epsilon' \leq \epsilon$,*

$$||\hat{f} - f_1^*||^2 - ||f_1 - f_1^*||^2 \leq \epsilon' \Rightarrow ||\hat{f} - f_1|| \leq ||\hat{f} - f_2|| \tag{4.7}$$

and

$$||\hat{f} - f_2^*||^2 - ||f_1 - f_2^*||^2 \leq \epsilon' \Rightarrow ||\hat{f} - f_2|| \leq ||\hat{f} - f_1||. \tag{4.8}$$

Proof. Recall that $||f_m - f_1^*||^2 - ||f_1 - f_1^*||^2 = \epsilon$. We show that

$$||\hat{f} - f_2|| < ||\hat{f} - f_1|| \Rightarrow ||\hat{f} - f_1^*|| > ||f_m - f_1^*||$$

which implies

$$||\hat{f} - f_1^*||^2 - ||f_1 - f_1^*||^2 > \epsilon'.$$

We have

$$\begin{aligned}
|| \hat{f} - f_1^*||^2 &= ||\hat{f} - f_c + f_c - f_1^*||^2 \\
&= ||\hat{f} - f_c||^2 + ||f_c - f_1^*||^2 + 2\langle \hat{f} - f_c, f_c - f_1^* \rangle \\
&\geq ||f_m - f_c||^2 + ||f_c - f_1^*||^2 + 2\langle \hat{f} - f_c, f_c - f_1^* \rangle \\
&= ||f_m - f_1^*||^2 + 2\langle \hat{f} - f_c, f_c - f_1^* \rangle,
\end{aligned}$$

where the inequality follows from the fact that $\hat{f}$ is in $\mathcal{F}$. Thus we need only show that the second term is greater than zero when $||\hat{f} - f_2|| < ||\hat{f} - f_1||$.

We have

$$\begin{aligned}
& ||\hat{f} - f_2||^2 < ||\hat{f} - f_1||^2 \\
\Leftrightarrow\ & ||\hat{f} - f_c||^2 + ||f_c - f_2||^2 + 2\langle \hat{f} - f_c, f_c - f_2 \rangle \\
& \qquad\qquad < ||\hat{f} - f_c||^2 + ||f_c - f_1||^2 + 2\langle \hat{f} - f_c, f_c - f_1 \rangle \\
\Leftrightarrow\ & \langle \hat{f} - f_c, f_c - f_2 \rangle < \langle \hat{f} - f_c, f_c - f_1 \rangle \\
\Leftrightarrow\ & \langle \hat{f} - f_c, f_2 - f_1 \rangle > 0 \\
\Leftrightarrow\ & \langle \hat{f} - f_c, f_c - f_1^* \rangle > 0.
\end{aligned}$$

since $f_2 - f_1$ and $f_c - f_1^*$ are in the same direction.

By symmetry, the second statement of the lemma is also true. □

Assuming the agnostic learning algorithm is successful, we can choose the correct target conditional expectation by choosing f_1^* if $\|\hat{f} - f_1\| < \|\hat{f} - f_2\|$ and f_2^* if $\|\hat{f} - f_2\| < \|\hat{f} - f_1\|$. The case $\|\hat{f} - f_1\| = \|\hat{f} - f_2\|$ cannot happen if we choose $\epsilon' < \epsilon$. (For convenience we will use $\epsilon' = \epsilon/2$.)

Proof (Theorem 4.4). Assume an algorithm A exists such that for any probability distribution on $\mathcal{X} \times \mathcal{Y}$, the algorithm draws m examples and with probability at least $1 - \delta$, it produces $\hat{f}$ such that $\|\hat{f} - f^*\|^2 - \|f_a - f^*\|^2 \leq \epsilon$, where $f^*(x) = \mathbf{E}[Y|X = x]$ and $\|f_a - f^*\| = \inf_{f \in \mathcal{F}} \|f - f^*\|$. The function $f_a \in \bar{\mathcal{F}}$ is the best approximation to f^* in $\bar{\mathcal{F}}$. Algorithm A is an agnostic learning algorithm for $\bar{\mathcal{F}}$.

If the sample complexity of Algorithm A (to accuracy $\epsilon/2$ and confidence $1 - \delta$) is m, then there exists an algorithm, Algorithm B (which depends on $P_{\mathcal{X}}$ and the non-convex $\bar{\mathcal{F}}$) which with probability $1-\delta$ solves the problem in Lemma 4.5 (for γ which depends on ϵ according to Claim 4.7) with sample complexity m. Let $f_1' = (f_1^* + f_2^*)/2 + (f_{d1} - c)$ and $f_2' = (f_1^* + f_2^*)/2 + (f_{d2} - c)$. Algorithm B generates a sequence $(x_1, \ldots, x_m) \in \mathcal{X}^m$ independently from $P_{\mathcal{X}}$. If $\xi_i = 1$, Algorithm B gives $(x_i, f_1'(x_i))$ to Algorithm A; otherwise it gives $(x_i, f_2'(x_i))$ to Algorithm A. The target conditional expectation is f_1^* if $\alpha = 1/2 + \gamma/2$ and f_2^* if $\alpha = 1/2 - \gamma/2$. Algorithm B receives $\hat{f}$ from Algorithm A. If $\|\hat{f} - f_1\| \leq \|\hat{f} - f_2\|$ the Algorithm B chooses $\alpha = 1/2 + \gamma/2$; otherwise it chooses $\alpha = 1/2 - \gamma/2$. Lemma 4.8 shows that it is the correct choice (with probability at least $1 - \delta$). From Lemma 4.5 obtaining the correct α with probability $1 - \delta$ requires $\Omega(\ln(1/\delta)/\gamma^2)$ observations. Algorithm B uses the same number of examples as Algorithm A, so Algorithm A also requires at least $\Omega(\ln(1/\delta)/\gamma^2) = \Omega(\ln(1/\delta)/\epsilon^2)$ observations.

To satisfy the definition of agnostic learning, we require that the range of the random variables be bounded. This can be done if the appropriate functions f_1, f_2 and c are chosen according to the construction in Lemma 4.6. Assume that f_1 and f_2 are bounded (if they are not, arbitrarily close bounded functions can be chosen since they are in the closure of $\mathcal{F}$ which contains uniformly bounded functions). From the construction, c is also a bounded function. This means that f_1^* and f_2^* are always bounded since they are convex combinations of c with f_1 and f_2. Finally, $f_{d1} - c = \frac{1}{\gamma}\left(f_1^* - \frac{f_1^* + f_2^*}{2}\right)$ for all γ in the appropriate range. We can fix a value for γ to bound $f_{d1} - c$. Hence f_1' and f_2' can be constructed to have ranges which are always bounded by a quantity which depend only the function class and not on γ. $\square$

4.3 Discussion

A lower bound on the sample complexity for *proper* agnostic learning is not as strong as a lower bound for the sample complexity for agnostic learning. (We show in Chapter 5 that this lower bound is valid in general only for *proper* agnostic learning by giving examples where learning the function class using the convex hull of the original function class gives better sample complexity.) However, proper learning is important in many cases where the form of the representation is important. This is often the case when it is desired to be able to interpret the results. The parameters in the representation may have some physical significance or it may be necessary to have a particular representation which is easy to understand.

While positive results for agnostic learning are very useful, negative results have to be interpreted with care. The results in this chapter show that for small enough accuracy and high enough confidence, there will be probability distributions for which we cannot achieve the desired performance unless we have more than the number of observations stated in the bounds. However, this may not necessarily be the case for any particular learning problem we are interested in. For example, we have shown in Chapter 3, restrictions on the probability distributions, as in the case of function learning and learning with noise, can give smaller sample complexity than the lower bound for proper agnostic learning given in this chapter.

The brain is composed of about 10^{11} neurons of many different types.

— Hertz, Krogh and Palmer,
Introduction to the Theory of Neural Computation.

Chapter 5

Learning Single Hidden Layer Neural Networks

In Chapter 4, we showed that if a function class is not closure-convex, then the sample complexity for properly agnostically learning the function class can be worse than the sample complexity for learning with noise. In view of this, we now consider agnostically learning the convex hull of the function class (which is closure-convex). Besides being closure convex, the convex hull will usually give a better approximation to the target function if the target function is not in the function class. This makes learning the convex hull a fairly natural way of using a different hypothesis class to learn a function class agnostically.

However, the convex hull of the function class may have a larger covering number than the function class. In this chapter, we study the sample complexity of learning the convex hulls of function classes. We obtain bounds for learning the convex hull of a function class in terms of the covering number of the original function class. For function classes with finite pseudo-dimension which are not closure-convex, we find that the sample complexity for agnostically learning the convex hull is not significantly worse (within constant and logarithmic factors) of the sample complexity for properly agnostically learning the function class.

The class of single hidden layer neural networks with an unbounded number of hidden units and a bound on the sum of absolute value of output weights is the convex hull of a symmetric class of hidden units (see Chapter 2 for definitions). Many classes of hidden units such as classes of sigmoid or linear threshold hidden units are not closure-convex making the results in this chapter particularly applicable. In fact, it is easy to see that the class of single hidden layer neural networks with any fixed number of sigmoid or linear threshold hidden units is not closure-convex.

With an unbounded number of hidden units, the class of single hidden layer neural networks can also be used to approximate many nonparametric function classes. We start off in Section 5.1 by reviewing work by Barron (1994) on learning classes of functions with finite first absolute moment of the Fourier transform using single hidden layer neural networks. We then extend this result to agnostic learning of single hidden layer neural networks with linear threshold hidden units and other more general hidden units in Section 5.2 by giving sample complexity bounds. We end this chapter with a discussion on the optimality of the results.

5.1 Function Classes with Finite First Absolute Moment of Fourier Transform

The approximation properties of single hidden layer neural networks with sigmoid hidden units was studied in (Barron 1993). There it was shown that for functions with a finite first absolute moment of Fourier transform, a single hidden layer neural network can achieve integrated squared error of $O(1/k)$ where k is the number of hidden units.

Let Γ_1 the the class of functions with the first absolute moment of the Fourier transform bounded by C, that is $\int_{\mathbb{R}^n} \sum_{j=1}^{n} |2\pi u_j||F(u)|du \leq C$ where $F(u) = \int_{\mathbb{R}^n} f(x)e^{-i2\pi u\cdot x}dx$ is the Fourier transform of f. Recall that $\mathcal{N}_{K,k}^{\mathcal{G}}$ is a single hidden layer neural networks with k hidden units.

Theorem 5.1 *Suppose that $\mathcal{G}$ is either the class of linear threshold functions or the class of sigmoid functions. For every function in Γ_1 and every probability measure P, there exists a function $f_k \in \mathcal{N}_{K,k}^{\mathcal{G}}$ such that*

$$\int (f(x) - f_k(x))^2 dP(x) \leq \frac{(2C)^2}{k}.$$

Here $K = \sum_{i=0}^{k} |w_i| \leq 2C$ and $w_0 = f(0)$.

Barron (1993) has also provided examples for which the constant in the bound grows only moderately with dimension including positive definite functions that are continuously differentiable at the origin. Various closure properties for sums, products and certain compositions of functions where the constants grow polynomially are also given.

Using the approximation result and the sample complexity result from (Barron 1990), it was shown in (Barron 1994) that the sample complexity for learning with noise from functions in Γ_1 is $O\left(\frac{1}{\epsilon}\left(\frac{1}{\epsilon}\ln\frac{1}{\epsilon}+\log\frac{1}{\delta}\right)\right)$.

5.2 Learning Convex Combinations of Basis Functions

In this section, we give bounds on the sample complexity for agnostic learning in terms of the l_1 covering number. This extends the result of (Barron 1994) in several ways. First it extends the result for learning functions with bounded first moment of the Fourier transform from the case of learning with noise to agnostic learning. By using the l_1 covering number we are also able to learn a larger class of functions than the class of functions with a bound on the first moment of the Fourier transform . In (Barron 1994), the function class is learned by discretizing the weights of the neural networks (finding an L_∞ cover) and optimising over the discrete set of weights. By using bounds involving the l_1 covering number, it is possible to learn some classes of functions which are not continuous and have no finite L_∞ cover such as single hidden layer neural networks with linear threshold hidden units. Using the bounds involving the pseudo-dimension and fat-shattering function in Chapter 3, we also give upper bounds on the sample complexity for learning classes of single hidden layer neural networks with other basis functions as hidden units, in terms of the pseudo-dimension or fat-shattering function of the basis function classes. The results are stated in the following theorem and corollaries.

Theorem 5.2 *Let $\mathcal{G}$ be an admissible class of basis functions mapping from $\mathcal{X}$ into $\mathcal{Y}$ with $|g(x)| \leq b$ for all $g \in \mathcal{G}$. The sample complexity for agnostically learning $\mathcal{N}_K^{\mathcal{G}}$ is no more than*

$$\frac{14000C^2}{\epsilon}\left(\frac{4C^2}{\epsilon}\left(\ln\max_{\boldsymbol{x}\in\mathcal{X}^{2m}}\left(N\left(\frac{\epsilon}{1024CK},\mathcal{G}_{|\boldsymbol{x}},l_1\right)+1\right)+\ln 2\right)+\ln\frac{6}{\delta}\right)$$

where $C = \max\{Kb, T, 1\}$. **See Errata for correction**

Using bounds from Lemma 3.8 and Lemma 3.9, we obtain the following two corollaries.

Corollary 5.3 *Let $\mathcal{G}$ be an admissible class of basis functions mapping from $\mathcal{X}$ into $\mathcal{Y}$ with $|g(x)| \leq b$ for all $g \in \mathcal{G}$. Suppose the pseudo-dimension of $\mathcal{G}$ is d. The sample complexity for* *agnostically learning $\mathcal{N}_K^{\mathcal{G}}$ is no more than*

$$\frac{14000C^2}{\epsilon}\left(\frac{4C^2 d}{\epsilon}\ln\left(\frac{4096eC^2}{\epsilon}\ln\frac{4096C^2}{\epsilon}+1\right)+\frac{8C^2}{\epsilon}\ln 2+\ln\frac{6}{\delta}\right)$$

where $C = \max\{Kb, T, 1\}$.

Corollary 5.4 *Let $\mathcal{G}$ be an admissible class of basis functions mapping from $\mathcal{X}$ into $\mathcal{Y}$ with $|g(x)| \leq b$ for all $g \in \mathcal{G}$. Let $d = \mathrm{fat}_{\mathcal{G}}(\epsilon/8192C^4)$. The sample complexity for agnostically learning $\mathcal{N}_K^{\mathcal{G}}$ is no more than*

$$\frac{14000C^2}{\epsilon}\left(\frac{4C^2}{\epsilon}\left(\frac{8d}{\ln 2}\ln^2\left(\frac{2048(1024C)^4 d}{\epsilon^4\ln 2}\right)+1\right)+\frac{4C^2}{\epsilon}\ln 2+\ln\frac{6}{\delta}\right)$$

where $C = \max\{Kb, T, 1\}$.

For the proof of Theorem 5.2, we will need to bound the number of terms in the convex combination needed to achieve a desired accuracy. For that we use the following result attributed to Maurey in (Barron 1993).

Lemma 5.5 *If $\bar{f}$ is in the closure of the convex hull of a set $\mathcal{G}$ in a Hilbert space, with $||g|| \leq b$ for each $g \in \mathcal{G}$, then for every $k \geq 1$, and every $c > b^2 - ||\bar{f}||^2$, there is an f_k in the convex hull of k points of $\mathcal{G}$ such that*

$$||\bar{f} - f_k||^2 \leq \frac{c}{k}.$$

Furthermore f_k can be chosen to be $f_k = \frac{1}{k}\sum_{i=1}^{k} g_i$ where $g_i \in \mathcal{G}$ for $i = 1, \ldots, k$.

Observe that for functions with range in $[-B, B]$, $||g||^2$ is always bounded by B^2.

Recall (from Chapter 2) that $\mathcal{A}_{K,k}^{\mathcal{G}} = \left\{x \mapsto \frac{1}{k}\sum_{i=1}^{k} g_i(x) : g_i \in \mathcal{G}_K^1\right\}$ where $\mathcal{G}_K^1 = \{x \mapsto Kg(x), x \mapsto -Kg(x), x \mapsto K, x \mapsto -K : g \in \mathcal{G}\}$. Note that $\mathcal{N}_K^{\mathcal{G}}$ is the convex hull of $\mathcal{G}_K^1$. We will use functions from $\mathcal{A}_{K,k}^{\mathcal{G}}$ to approximate functions from $\mathcal{N}_K^{\mathcal{G}}$. We bound the covering number of these function classes in terms of the covering number of $\mathcal{G}$.

Lemma 5.6 *Let $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) \in \mathcal{X}^m$. Then*

$$N(\epsilon, \mathcal{G}_{K|\boldsymbol{x}}^1, l_1) \leq 2\left(N(\epsilon/K, \mathcal{G}_{|\boldsymbol{x}}, l_1) + 1\right).$$

Proof. Let U be an ϵ/K-cover for $\mathcal{G}_{|\boldsymbol{x}}$ with $|U| = N(\epsilon/K, \mathcal{G}_{|\boldsymbol{x}}, l_1)$. Pick an arbitrary function $Kf \in \mathcal{G}_K^1$ where $f \in \mathcal{G}$. Pick $v \in U$ such that $d_{l_1}(f_{|\boldsymbol{x}}, v) \leq \epsilon/K$. Then

$$\frac{1}{m}\sum_{i=1}^{m}|Kf(\boldsymbol{x}_i) - Kv_i| = \frac{K}{m}\sum_{i=1}^{m}|f(\boldsymbol{x}_i) - v_i| \leq \epsilon.$$

Obviously, $\{Kv, -Kv : v \in U\} \bigcup \{(K, \ldots, K), (-K, \ldots, -K)\}$ is a an ϵ-cover $\mathcal{G}_K^1$. □

Lemma 5.7 *Let $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m) \in \mathcal{X}^m$. Then*

$$N(\epsilon, \mathcal{A}^{\mathcal{G}^1}_{K,k|\boldsymbol{x}}, l_1) \leq \left(N(\epsilon, \mathcal{G}^1_{K|\boldsymbol{x}}, l_1)\right)^k \leq 2^k \left(N(\epsilon/K, \mathcal{G}_{|\boldsymbol{x}}, l_1) + 1\right)^k. \tag{5.1}$$

Proof. Let U be an ϵ-cover for $\mathcal{G}_K^1$ with $|U| = N(\epsilon, \mathcal{G}^1_{K|\boldsymbol{x}}, l_1)$. Let $f = \frac{1}{n}\sum_{i=1}^{k} f_i$ ($f_i \in \mathcal{G}_K^1$, $i = 1, \ldots, k$) be a function in $\mathcal{N}^{\mathcal{G}^1}_{K,k}$. For each f_i, pick a member u_i of U such that $l_1(f_{i|\boldsymbol{x}}, u_i) < \epsilon$. Let $h = \frac{1}{k}\sum_{i=1}^{k} u_i$. Then

$$\begin{aligned}
\frac{1}{m}\sum_{j=1}^{m}|f(\boldsymbol{x}_j) - h_j| &= \frac{1}{m}\sum_{j=1}^{m}\left|\frac{1}{k}\sum_{i=1}^{k}(f_i(\boldsymbol{x}_j) - u_{ij})\right| \\
&\leq \frac{1}{m}\frac{1}{k}\sum_{j=1}^{m}\sum_{i=1}^{k}|f_i(\boldsymbol{x}_j) - u_{ij}| \\
&= \frac{1}{k}\sum_{i=1}^{k}\frac{1}{m}\sum_{j=1}^{m}|f_i(\boldsymbol{x}_j) - u_{ij}| \\
&\leq \frac{1}{k}\sum_{i=1}^{k}\epsilon = \epsilon.
\end{aligned}$$

So for any $f_{|\boldsymbol{x}} \in \mathcal{A}^{\mathcal{G}^1}_{K,k}$, there is a vector in the set $\{\frac{1}{k}\sum_{i=1}^{k} u_i \colon u_i \in U\}$ with distance less than ϵ from it. Since $|\{\frac{1}{k}\sum_{i=1}^{k} u_i \colon u_i \in U\}| < |U|^k$ the first inequality in (5.1) follows. The second inequality in (5.1) follows from Lemma 5.6. □

We are now give the proof of Theorem 5.2.

Proof. (**Theorem 5.2**) First note that $\mathcal{N}_K^{\mathcal{G}}$ is convex and hence closure-convex. We also have $\mathcal{A}^{\mathcal{G}}_{K,k} \subset \mathcal{N}_K^{\mathcal{G}}$ permissible for each k. Scale the function class and target random variable by dividing by C. The covering number of the scaled function class is the same as the $C\epsilon$ covering number of the unscaled class. By learning to accuracy ϵ/C^2 and rescaling back, we obtain the desired bound. Assume the scaled function class is $\mathcal{F}$. In Theorem 3.7, set $\alpha = 1/2$ and use

Theorem 3.7 and Lemma 5.7 with $\nu = \nu_c = \epsilon/4C^2$ to get

$$\begin{aligned}
P^m \Big\{ \boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, \mathbf{E}\left[(y - f(x))^2 - (y - f_a(x))^2\right] & \\
\geq 2\hat{\mathbf{E}}_{\boldsymbol{z}}\left[(y - f(x))^2 - (y - f_a(x))^2\right] + \epsilon/(2C^2)\Big\} & \\
\leq 6 \max_{\boldsymbol{x} \in \mathcal{X}^{2m}} N\left(\frac{\epsilon}{1024C}, \mathcal{A}^{\mathcal{G}}_{K,k|\boldsymbol{x}}, l_1\right) \exp(-\epsilon m/14000C^2) & \\
\leq 6 \times 2^k \max_{\boldsymbol{x} \in \mathcal{X}^{2m}} \left(N\left(\frac{\epsilon}{1024CK}, \mathcal{G}_{|\boldsymbol{x}}, l_1\right) + 1\right)^k \exp(-\epsilon m/14000C^2). & \qquad (5.2)
\end{aligned}$$

Suppose $f'(x) = \hat{\mathbf{E}}_{\boldsymbol{z}}[Y|X = x]$. Let $\hat{f}_k$ be the estimated function and let $\hat{f}_a$ be the function in the convex closure which minimizes the empirical error. Then $\hat{\mathbf{E}}_{\boldsymbol{z}}\left[(y - \hat{f}_k(x))^2 - (y - f_a(x))^2\right] = \hat{\mathbf{E}}_{\boldsymbol{z}}\left[(f'(x) - \hat{f}_k(x))^2 - (f'(x) - f_a(x))^2\right]$. Note that $\hat{\mathbf{E}}_{\boldsymbol{z}}\left[(f'(x) - \hat{f}_k(x))^2 - (f'(x) - f_a(x))^2\right] \leq \hat{\mathbf{E}}_{\boldsymbol{z}}\left[(f'(x) - \hat{f}_k(x))^2 - (f'(x) - \hat{f}_a(x))^2\right] = \hat{\mathbf{E}}_{\boldsymbol{z}}\left[(\hat{f}_k(x) - \hat{f}_a(x))^2\right]$.

In Lemma 5.5, set $c = 1$. To get approximation within $\epsilon/4C^2$ (with respect to the empirical mean squared error), we require $k \geq 4C^2/\epsilon$. Setting the right hand side of (5.2) to be δ and $k = 4C^2/\epsilon$, we see that

$$m = \frac{14000C^2}{\epsilon}\left(\frac{4C^2}{\epsilon} \ln \max_{\boldsymbol{x} \in \mathcal{X}^{2m}} \left(N\left(\frac{\epsilon}{1024CK}, \mathcal{G}_{|\boldsymbol{x}}, l_1\right) + 1\right) + \frac{4C^2}{\epsilon} \ln 2 + \ln \frac{6}{\delta}\right)$$

will suffice for agnostic learning. □

5.3 *Discussion*

Corollary 5.3 shows that for function classes with finite pseudo-dimension which are not closure convex, the sample complexity for agnostically learning the convex hull of the function class is at worst within a logarithmic factor of the sample complexity for properly agnostically learning the function class itself. The convex hull can be learned by increasing the number of hidden units as a function of the required accuracy. Learning the convex hull gives better approximation capabilities and hence may be preferable to properly agnostically learning the function class in view of the sample complexity bounds.

The function class Γ_1 is in the closure of the convex hull of single hidden layer neural networks with linear threshold hidden units. Since the pseudo-dimension of linear threshold units is $n + 1$, the class Γ_1 is agnostically learnable with sample complexity $O\left(\frac{1}{\epsilon}\left(\frac{n}{\epsilon}\ln\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$. Barron

(1992) has also shown that the sample complexity for learning Γ_1 using an arbitrary estimator cannot be better than $\Omega(1/\epsilon^{(2n+2)/(n+2)})$. Hence the sample complexity bound is close to optimal for learning Γ_1. The bound is also close to optimal for learning the class of single hidden layer neural networks with linear threshold hidden units since functions in Γ_1 can be approximated arbitrarily closely single hidden layer neural networks.

There are also function classes for which using the convex hull as the hypothesis class (instead of doing proper agnostic learning) results in a much better sample complexity. For example if $\mathcal{G}$ has a finite number of functions, then the pseudo-dimension of the convex hull of $\mathcal{G}$ is bounded by $|\mathcal{G}|$ (the convex hull is a subset of a $|\mathcal{G}|$-dimensional vector space of functions, hence as mentioned in Section 3.3.1, the pseudo-dimension is bounded by $|\mathcal{G}|$ (Dudley 1978)). Since the pseudo-dimension is finite, Corollary 3.10 shows that the sample complexity is $O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$. In contrast, since $\mathcal{G}$ is not closure-convex, the sample complexity for properly agnostically learning $\mathcal{G}$ is $\Omega(\ln(1/\delta)/\epsilon^2)$. This shows that for such classes, by learning the convex hull of the function class, not only do we get better approximation, we also get a better sample complexity. Since the upper bound is smaller than $\Omega(\ln(1/\delta)/\epsilon^2)$ (for small enough ϵ) this also shows that the lower bound for the sample complexity for learning function classes which are not closure convex only holds for proper agnostic learning and not agnostic learning in general.

'... a computer whose merest operational parameters I am not worthy to calculate and yet I will design it for you. A computer which can calculate the Question to the Ultimate Answer. '

— Deep Thought,

in *The Hitchhiker's Guide to the Galaxy.*

Chapter 6

Computational Complexity

Another important component of a learning problem is its computational complexity. As computing capabilities increase, we expect to be able to solve harder learning problems. However, if the computational requirements grow too quickly with the size of the problem, the size of solvable problems will remain quite restricted. One of the main aims of computational learning theory is the study of the maximum size of learning problems which can be solved using a reasonable amount of computation (Valiant 1984). We delineate the boundary of what is feasibly learnable by requiring the computational requirements of learnable problems to be polynomial in $1/\epsilon$, $1/\delta$ and the relevant complexity parameters. Classes of functions for which this can be done are said to be *efficiently* learnable.

In this chapter, we study the computational requirements of agnostically learning single hidden layer neural networks. We first relate the computational complexity of agnostically learning the basis function class to the computational complexity of learning the class of single hidden layer neural networks. We do this via an iterative approximation result which shows that by iteratively adding a function to the convex combination of a function class such that the distance to the target function is minimised, good convergence to the best approximation in the convex hull of the function class can be achieved even when the target function is not in the convex hull. The iterative approximation result is described in Section 6.1. We then show in Section 6.2 how the iterative approximation result can be used to show that if a basis function class is efficiently agnostically learnable, then the convex hull of the function class is also efficiently agnostically learnable. Since a basis function class is contained in the convex hull, this means that the convex hull of a function class (and the class of single hidden layer neural networks with those basis functions) is efficiently agnostically learnable if and only if the basis function class is efficiently agnostically learnable.

Learning $\{0, 1\}$-valued functions with $\{0, 1\}$ valued targets is widely studied in computational learning theory. We call the proper agnostic version of this problem proper agnostic PAC learning (Kearns et al. 1994). In Section 6.3, we show how a proper efficient agnostic PAC learning algorithm for a basis function class $\mathcal{G}$ can be used to efficiently agnostically learn single hidden layer neural networks (with real-valued outputs) with hidden units from $\mathcal{G}$.

In Section 6.4, we show that the problem of agnostically learning of some classes of single hidden layer neural networks (including networks with linear threshold hidden units) is likely to be difficult computationally. We do this by showing that an algorithm for agnostically learning the network can be used for PAC learning polynomial sized DNF formulae. Whether the class of polynomial sized DNF formulae is PAC learnable has been an open problem in computational learning theory since it was first posed by Valiant (1984). It is generally believed that polynomial sized DNF is not efficiently learnable (Jerrum 1994).

In view of this, we consider learning subclasses of single hidden layer neural networks. In Chapter 7, we will study functions with finite q-th absolute moment of the Fourier transform. In Section 6.5, we show that the class of single hidden layer neural networks with linear threshold hidden units and bounded fan-in is efficiently agnostically learnable. We end the chapter with a discussion of the results in Section 6.6.

6.1 Iterative Approximation

The iterative approximation result in this section is an extension of the results of Jones (1992) and Barron (1993). They showed that if a function is in the closure of the convex hull of a bounded set of functions in a Hilbert space, then it can be approximated by iteratively adding functions from the set such that the squared distance to the target function is of order $O(1/k)$, where k is the number of functions added. We extend the result in order to allow agnostic learning. We show that even if the target function is not in closure of the convex hull, the iterative approximation scheme will converge to the best possible approximation such that the squared distance to the target will approach the optimal squared distance at a rate of $O(1/k)$.

We now give the iterative approximation result which is the key to showing the equivalence between efficient agnostic learning of a function class and efficient agnostic learning of its convex hull.

Theorem 6.1 *Let $\mathcal{H}$ be a Hilbert space with norm $\| \cdot \|$. Let $\mathcal{G}$ be a subset of $\mathcal{H}$ with $\| g \| \leq b$ for each $g \in \mathcal{G}$. Let $co(\mathcal{G})$ be the convex hull of $\mathcal{G}$. For any $f \in \mathcal{H}$, let $d_f = \inf_{g' \in co(\mathcal{G})} \| g' - f \|$. Suppose that f_1 is chosen to satisfy*

$$\| f_1 - f \|^2 \leq \inf_{g \in \mathcal{G}} \| g - f \|^2 + \epsilon_1$$

and iteratively, f_k is chosen to satisfy

$$\| f_k - f \|^2 \leq \inf_{g \in \mathcal{G}} \| \alpha f_{k-1} + \overline{\alpha} g - f \|^2 + \epsilon_k$$

where $\alpha = 1 - 2/(k+1)$, $\overline{\alpha} = 1 - \alpha$, $c \geq b^2$, and $\epsilon_k \leq \frac{4(c-b^2)}{(k+1)^2}$. Then for every $k \geq 1$,

$$\| f - f_k \|^2 - d_f^2 \leq \frac{4c}{k}. \tag{6.1}$$

Proof. Given $\delta > 0$, let f^* be a point in the convex hull of $\mathcal{G}$ with $\| f^* - f \| \leq d_f + \delta$. Thus $f^* = \sum_{i=1}^{N} \gamma_i g_i$ with $g_i \in \mathcal{G}$, $\gamma_i \geq 0$ and $\sum_{i=1}^{N} \gamma_i = 1$ for some sufficiently large N. Then for all $\alpha \in [0, 1]$,

$$\begin{aligned} & \| \alpha f_{k-1} + \overline{\alpha} g - f \|^2 \\ & \quad = \| \alpha f_{k-1} + \overline{\alpha} g - f^* + f^* - f \|^2 \\ & \quad = \| \alpha f_{k-1} + \overline{\alpha} g - f^* \|^2 + \| f^* - f \|^2 + 2\langle \alpha f_{k-1} + \overline{\alpha} g - f^*, f^* - f \rangle, \end{aligned}$$

where $(\cdot, \cdot)$ is the inner product in the Hilbert space $\mathcal{H}$. Thus,

$$\begin{aligned} & \| \alpha f_{k-1} + \overline{\alpha} g - f \|^2 - \| f^* - f \|^2 \\ & \quad = \| \alpha f_{k-1} + \overline{\alpha} g - f^* \|^2 + 2\langle \alpha f_{k-1} + \overline{\alpha} g - f^*, f^* - f \rangle \\ & \quad = \| \alpha (f_{k-1} - f^*) + \overline{\alpha}(g - f^*) \|^2 + 2\langle \alpha f_{k-1} + \overline{\alpha} g - f^*, f^* - f \rangle \\ & \quad = \alpha^2 \| f_{k-1} - f^* \|^2 + \overline{\alpha}^2 \| g - f^* \|^2 + \end{aligned}$$

$$2\alpha\overline{\alpha}\langle f_{k-1} - f^*, g - f^*\rangle + 2\langle \alpha f_{k-1} + \overline{\alpha} g - f^*, f^* - f\rangle.$$

Let g be independently drawn from the set $\{g_1, \ldots, g_N\}$ with $P\{g = g_i\} = \gamma_i$. The average value of $\| \alpha f_{k-1} + \overline{\alpha} g - f \|^2 - \| f^* - f \|^2$ is

$$\begin{aligned}
&\sum_{i=1}^{N} \gamma_i \Big[\alpha^2 \| f_{k-1} - f^* \|^2 + \overline{\alpha}^2 \| g_i - f^* \|^2 + 2\alpha\overline{\alpha}\langle f_{k-1} - f^*, g_i - f^*\rangle \\
&\quad + 2\langle \alpha f_{k-1} + \overline{\alpha} g_i - f^*, f^* - f\rangle\Big] \\
&\quad = \alpha^2 \| f_{k-1} - f^* \|^2 + \overline{\alpha}^2 \sum_{i=1}^{N} \gamma_i \| g_i - f^* \|^2 + 2\alpha\overline{\alpha} \sum_{i=1}^{N} \gamma_i \langle f_{k-1} - f^*, g_i - f^*\rangle \\
&\qquad + 2\sum_{i=1}^{N} \gamma_i \langle \alpha f_{k-1} + \overline{\alpha} g_i - f^*, f^* - f\rangle \\
&\quad = \alpha^2 \| f_{k-1} - f^* \|^2 + \overline{\alpha}^2 \left(\sum_{i=1}^{N} \gamma_i \left(\| g_i \|^2 - 2\langle g_i, f^*\rangle + \| f^* \|^2\right)\right) + 0 \\
&\qquad + 2\sum_{i=1}^{N} \gamma_i \langle \alpha f_{k-1} + g_i - \alpha g_i - f^*, f^* - f\rangle \\
&\quad = \alpha^2 \| f_{k-1} - f^* \|^2 + \overline{\alpha}^2 \left(\sum_{i=1}^{N} \gamma_i \| g_i \|^2 - \| f^* \|^2\right) + 2\alpha\langle f_{k-1} - f^*, f^* - f\rangle \\
&\quad \le \alpha^2 \| f_{k-1} - f^* \|^2 + \overline{\alpha}^2 b^2 + 2\alpha\langle f_{k-1} - f^*, f^* - f\rangle.
\end{aligned}$$

Since the average is bounded in this way, there must be a $g \in \{g_1, \ldots, g_N\}$ such that

$$\begin{aligned}
&\| \alpha f_{k-1} + \overline{\alpha} g - f \|^2 - \| f^* - f \|^2 \\
&\quad \le \alpha^2 \| f_{k-1} - f^* \|^2 + \overline{\alpha}^2 b^2 + 2\alpha\langle f_{k-1} - f^*, f^* - f\rangle \\
&\quad = \alpha\left[\alpha \| f_{k-1} - f^* \|^2 + 2\langle f_{k-1} - f^*, f^* - f\rangle\right] + \overline{\alpha}^2 b^2 \\
&\quad \le \alpha\left[\| f_{k-1} - f^* \|^2 + 2\langle f_{k-1} - f^*, f^* - f\rangle\right] + \overline{\alpha}^2 b^2 \qquad (6.2)
\end{aligned}$$

since $\alpha \in [0, 1]$. Noting that

$$\begin{aligned}
\| f_{k-1} - f \|^2 &= \| f_{k-1} - f^* + f^* - f \|^2 \\
&= \| f_{k-1} - f^* \|^2 + \| f^* - f \|^2 + 2\langle f_{k-1} - f^*, f^* - f\rangle,
\end{aligned}$$

we get

$$\| f_{k-1} - f \|^2 - \| f^* - f \|^2 = \| f_{k-1} - f^* \|^2 + 2\langle f_{k-1} - f^*, f^* - f \rangle.$$

Substituting into (6.2) and letting δ go to 0, we get

$$\inf_{g \in \mathcal{G}} \| \alpha f_{k-1} + \overline{\alpha} g - f \|^2 - d_f^2 \leq \alpha \left[\| f_{k-1} - f \|^2 - d_f^2 \right] + \overline{\alpha}^2 b^2.$$

Setting $k = 1$, $\alpha = 0$ and $f_0 = 0$, we see that

$$\inf_{g \in \mathcal{G}} \| g - f \|^2 - d_f^2 \leq b^2.$$

Hence the theorem is true for $k = 1$. Assume as an inductive hypothesis that

$$\| f_{k-1} - f \|^2 - d_f^2 \leq \frac{4c}{k-1}.$$

Then

$$\inf_{g \in \mathcal{G}} \| \alpha f_{k-1} + \overline{\alpha} g - f \|^2 - d_f^2 + \epsilon_k \leq \frac{4c}{k-1} + \overline{\alpha}^2 b^2 + \frac{4(c - b^2)}{(k+1)^2}.$$

Letting $\alpha = 1 - 2/(k+1) = \frac{k-1}{k+1}$, $\overline{\alpha} = 2/(k+1)$,

$$\begin{aligned} \inf_{g \in \mathcal{G}} \| \alpha f_{k-1} + \overline{\alpha} g - f \|^2 - d_f^2 + \epsilon_k &\leq \frac{4c}{k+1} + \frac{4b^2}{(k+1)^2} + \frac{4(c - b^2)}{(k+1)^2} \\ &= \frac{4c}{k+1} + \frac{4c}{(k+1)^2} \\ &\leq \frac{4c}{k}. \end{aligned}$$

Then (6.1) follows directly. □

Recently, Koiran (1994) has independently obtained a similar result for iterative approximation when the target function is not in the convex closure of the set of functions. In (Koiran 1994), he obtained bounds of the form $\| f - f_k \|^2 - d_f^2 \leq \frac{2Cd_f}{\sqrt{k}} + \frac{C^2}{k}$ where $C > \sqrt{b^2 + d_f^2}$. In comparison, our bound of $O\left(\frac{1}{k}\right)$ is asymptotically better than Koiran's bound of $O\left(\frac{1}{\sqrt{k}}\right)$. The constant in our bound is also independent of the target function, unlike the constant in his bound.

6.2 Equivalence in Efficient Learning

In this section we show that the class of single hidden layer neural networks with hidden units from an admissible class of basis functions is efficiently agnostically learnable if and only if the class of basis functions is efficiently agnostically learnable. This is done by using the iterative approximation result together with the agnostic learning algorithm of the basis function class as a subroutine to learn one hidden unit at a time.

We will need the following uniform convergence result which follows from Hoeffding's inequality (Hoeffding 1963) and the union bound.

Theorem 6.2 *Let $\mathcal{F}$ be a finite set of functions on some set $\mathcal{Z}$ with $0 \leq f(z) \leq C$ for all $f \in \mathcal{F}$ and $z \in \mathcal{Z}$. Let S be a sequence of m points drawn independently from $\mathcal{Z}$ according to an arbitrary distribution P on $\mathcal{Z}$, and let $\epsilon > 0$. Then*

$$P^m(\exists f \in \mathcal{F} : |\hat{\mathbf{E}}_S(f) - \mathbf{E}(f)| > \epsilon) \leq 2|\mathcal{F}|e^{-2\epsilon^2 m/C}.$$

For $0 < \delta \leq 1$ and sample size

$$m \geq \frac{C^2}{2\epsilon^2}\left(\ln|\mathcal{F}| + \ln\frac{2}{\delta}\right)$$

this probability is at most δ.

We now show how an agnostic learning algorithm for $\mathcal{G}$ (using any hypothesis class $\mathcal{H}$) can be used as a subroutine to construct an algorithm for agnostically learn $\mathcal{N}_K^{\mathcal{G}}$.

Theorem 6.3 *Let $\mathcal{G}$ be an admissible class of basis functions with $|g(x)| \leq b$ for all $g \in \mathcal{G}$. Then, with the quadratic loss function and $K > 0$, $\mathcal{N}_K^{\mathcal{G}}$ is efficiently agnostically learnable if and only if $\mathcal{G}$ is efficiently agnostically learnable.*

Proof. The *only if* part is trivial because K and ϵ can be rescaled such that $\mathcal{G}$ is a subset of $\mathcal{N}_K^{\mathcal{G}}$.

The function class $\mathcal{N}_K^{\mathcal{G}}$ is the convex hull of $\mathcal{G}^1 = \{wg\colon |w| = K, g \in \mathcal{G}\} \bigcup \{x \mapsto K, x \mapsto -K\}$. Theorem 6.1 shows that to get within ϵ of the best expected loss, a number of iterations equal to $k = \frac{4c}{\epsilon}$ will do. Set $c = 2K^2b^2$ and $\epsilon_i = \frac{4K^2b^2}{(i+1)^2}$ for $1 \leq i \leq k$. Assume that the agnostic algorithm for learning $\mathcal{G}$ produces an hypothesis from $\mathcal{H}$. Since we are not making any

SUBLEARN(i,K,b,T,ϵ,δ,k,f_{i-1})

1. *Set confidence to $\frac{\delta}{2k}$ and accuracy to $\frac{(i+1)^2\epsilon_i}{8K^2}$ where $\epsilon_i = \frac{4K^2b^2}{(i+1)^2}$.*
2. *Pass this confidence and accuracy to algorithm A and, for each observation (X, Y) from the original probability distribution, pass $(X, \frac{i+1}{2K}((1 - 2/(i+1))f_{i-1}(X) - Y)$ to A. The bound on the magnitude of the new target random variable, $(i+1)T/K$, is also passed to A. Assume the hypothesis produced is h_1.*
3. *Repeat Step 2, except that each observation (X, Y) is replaced by $(X, \frac{-(i+1)}{2K}((1 - 2/(i + 1))f_{i-1}(X) - Y)$. Assume the hypothesis produced is h_2.*
4. *Draw $\frac{8(2T)^4}{\epsilon_i^2}(\ln 4 + \ln\frac{4k}{\delta})$ additional observations from the original distribution. Test the four hypotheses $clip_T \circ ((1-2/(i+1))f_{i-1} + 2Kh_1/(i+1))$, $clip_T \circ ((1-2/(i+1))f_{i-1} - 2Kh_2/(i+1))$, $(1-2/(i+1))f_{i-1} + 2K/(i+1)$ and $(1-2/(i+1))f_{i-1} - 2K/(i+1)$ against the observations and select the one which gives the minimum error as f_i.*

Figure 6.1: Pseudo code for each iteration of the agnostic learning algorithm for $\mathcal{N}_K^G$ (algorithm A is an agnostic learning algorithm for learning $\mathcal{G}$).

assumptions about $\mathcal{H}$ we do not know that it is bounded. We introduce the function $clip_T$ where

$$clip_T(x) = \begin{cases} -T & \text{if } x \leq T \\ x & \text{if } -T < x < T \\ T & \text{if } x \geq T. \end{cases}$$

After each iteration, we compose the resulting function with $clip_T$. This can only improve the performance of the function since the observation range is a subset of $[-T, T]$. It also allows us to bound the range of the resulting hypothesis at each iteration. Assume the agnostic learning algorithm for learning $\mathcal{G}$ is A. The pseudo-code for each iteration of the algorithm (*SUBLEARN*) is shown in Figure 6.1.

Let f be the target function (conditional expectation of target Y given input X). To satisfy Theorem 6.1, at the ith iteration, we must find function h such that

$$\int_{X\times Y}(2h(x)/(i+1) + (1-2/(i+1))f_{i-1}(x) - y)^2 dP(x,y)$$
$$\leq \inf_{g\in G^1}\int_{X\times Y}(2g(x)/(i+1) + (1-2/(i+1))f_{i-1}(x) - y)^2 dP(x,y) + \epsilon_i.$$

where f_{i-1} is the linear combination which has been found so far (possibly composed with $clip_T$ after each iteration).

Now members of $\mathcal{G}^1$ consist of wg where $w = \pm K$ and $g \in \mathcal{G}\bigcup\{x \mapsto 1, x \mapsto -1\}$.

Furthermore

$$\int_{X\times Y}(2wg(x)/(i+1)+(1-2/(i+1))f_{i-1}(x)-y)^2dP(x,y)$$
$$= \left(\frac{2w}{i+1}\right)^2\int_{X\times Y}\left(g(x)+\frac{i+1}{2w}\left((1-2/(i+1))f_{i-1}(x)-y\right)\right)^2dP(x,y).$$

We now use the agnostic learning algorithm for $\mathcal{G}$ with respect to the new target random variable which has magnitude bounded by $(i+1)T/K$ (Step 2 and 3 in *SUBLEARN*). Set confidence to $\delta/2k$ and accuracy to $(i+1)^2\epsilon_i/8K^2$. Then with probability at least $1-\delta/2k$, the hypothesis h_i produced is such that

$$\int_{X\times Y}(2wh_i(x)/(i+1)+(1-2/(i+1))f_{i-1}(x)-y)^2dP(x,y)$$
$$\leq \left(\frac{2w}{(i+1)}\right)^2\left[\inf_{g\in\mathcal{G}}\int_{X\times Y}\left(g(x)+\frac{i+1}{2w}((1-2/(i+1))f_{i-1}(x)-y)\right)^2dP(x,y)\right.$$
$$\left.+(i+1)^2\epsilon_i/8K^2\right]$$
$$= \inf_{g\in\mathcal{G}}\int_{X\times Y}(wg(x)/i+(1-1/i)f_{i-1}(x)-y)^2dP(x,y)+\epsilon_i/2.$$

This has to be done for both $w=K$ and $w=-K$. We also have to compare the performances of the functions $x\mapsto K$ and $x\mapsto -K$. Hence at each iteration, we produce four hypotheses from which we have to choose one. If we have no other way of choosing between the four hypotheses, we have to do hypothesis testing (Step 4 in *SUBLEARN*). From Theorem 6.2, a sample size of $\frac{8(2T)^4}{\epsilon_i^2}(\ln 4+\ln\frac{4k}{\delta})$ is large enough so that the empirical quadratic loss is no more that $\epsilon_i/4$ from the expected quadratic loss for all functions with probability at least $1-\delta/2k$. If we choose the hypothesis which has the smallest empirical loss, the expected loss will be no more that $\epsilon_i/2$ away from the expected loss of the best hypothesis with probability $1-\delta/2k$.

So at each iteration, given an efficient agnostic learning algorithm for learning $\mathcal{G}$, we can produce an hypothesis which satisfies the requirements of Theorem 6.1 with probability at least $1-\delta/k$. Since the probability of failure at any of the k iterations is no more than δ, we have produced a learning algorithm for $\mathcal{N}_K^{\mathcal{G}}$. It is easy to see that if the time complexity of the algorithm for learning $\mathcal{G}$ is polynomial in the relevant parameters, the time complexity of the resulting algorithm for learning $\mathcal{N}_K^{\mathcal{G}}$ will be polynomial in the desired parameters. □

In Theorem 6.3, we make no assumptions about the hypothesis class used by the agnostic learning algorithm for learning $\mathcal{G}$. If we have a proper agnostic learning algorithm (or we know the hypothesis class), we can use a different algorithm which minimizes the empirical error at each stage instead of the expected error. With a proper agnostic learning algorithm for $\mathcal{G}$ this algorithm gives a better bound on the sample complexity.

We need to bound the covering number of the network constructed using Theorem 6.1. Let $\mathcal{C}^{\mathcal{G}}_{K,k} = \left\{x \mapsto \sum_{i=1}^{k} a_i g_i(x) \colon g_i \in \mathcal{G}^1_K\right\}$ where $\mathcal{G}^1_K = \{x \mapsto Kg(x), x \mapsto -Kg(x), x \mapsto K, x \mapsto -K \colon g \in \mathcal{G}\}$ and $a_1, \ldots, a_k$ is the fixed sequence of numbers constructed according to Theorem 6.1 with $a_i \geq 0$, $i = 1, \ldots, k$ and $\sum_{i=1}^{k} a_i = 1$.

Lemma 6.4 *Let $\boldsymbol{x} = (\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m)$ where $\boldsymbol{x}_i \in \mathcal{X}$ for $i = 1, \ldots, m$.*

$$N(\epsilon, \mathcal{C}^{\mathcal{G}^1}_{K,k|\boldsymbol{x}}, l_1) \leq \left(N(\epsilon, \mathcal{G}^1_{K|\boldsymbol{x}}, l_1)\right)^k \leq 2^k \left(N(\epsilon/K, \mathcal{G}_{|\boldsymbol{x}}, l_1) + 1\right)^k. \tag{6.3}$$

The proof is essentially the same as the proof for Lemma 5.7.

Theorem 6.5 *Let $\mathcal{G}$ be an admissible function class. Then $\mathcal{N}^{\mathcal{G}}_K$ is properly efficiently agnostically learnable if $\mathcal{G}$ is properly efficiently agnostically learnable. Furthermore the sample complexity for properly efficiently learning $\mathcal{N}^{\mathcal{G}}_K$ is at most*

$$\frac{14000C^4}{\epsilon}\left(\frac{32C^2}{\epsilon}\ln \max_{\boldsymbol{x}\in\mathcal{X}^{2m}}\left(N\left(\frac{\epsilon}{1024CK}, \mathcal{G}_{|\boldsymbol{x}}, l_1\right) + 1\right) + \frac{32C^2}{\epsilon}\ln 2 + \ln\frac{12}{\delta}\right)$$

where $C = \max\{Kb, T, 1\}$.

Proof. As in the proof of Theorem 5.2, we can scale the function class and target random variable by dividing by C.

The covering number of the scaled function class is the same as the $C\epsilon$ covering number of the unscaled class. By calculating the bounds for accuracy ϵ/C^2 and rescaling back, we obtain the desired bound. Assume the scaled function class is $\mathcal{F}$. Let $\alpha = 1/2$ in Theorem 3.7 and use Lemma 6.4 with $\nu = \nu_c = \epsilon/4C^2$ to get

$$\begin{aligned}
P^m \Big\{\boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, \mathbf{E}\left[(y - f(x))^2 - (y - f_a(x))^2\right] & \\
\geq 2\hat{\mathbf{E}}_{\boldsymbol{z}}\left[(y - f(x))^2 - (y - f_a(x))^2\right] + \epsilon/(2C^2)\Big\} & \\
\leq 6 \max_{\boldsymbol{x}\in\mathcal{X}^{2m}} N\left(\frac{\epsilon}{1024C}, \mathcal{C}^{\mathcal{G}}_{K,k|\boldsymbol{x}}, l_1\right)\exp(-\epsilon m/14000C^2) & \\
\leq 6 \times 2^k \max_{\boldsymbol{x}\in\mathcal{X}^{2m}}\left(N\left(\frac{\epsilon}{1024CK}, \mathcal{G}_{|\boldsymbol{x}}, l_1\right) + 1\right)^k \exp(-\epsilon m/14000C^2). &
\end{aligned}$$

Let $f'(x) = \hat{\mathbf{E}}_{\boldsymbol{z}}[Y|X = x]$. Let $\hat{f}_k$ be the estimated function and let $\hat{f}_a$ be the function in the convex closure which minimizes the empirical error. Then $\hat{\mathbf{E}}_{\boldsymbol{z}}\left[(y - \hat{f}_k(x))^2 - (y - f_a(x))^2\right] = \hat{\mathbf{E}}_{\boldsymbol{z}}\left[(f'(x) - \hat{f}_k(x))^2 - (f'(x) - f_a(x))^2\right]$. Note that $\hat{\mathbf{E}}_{\boldsymbol{z}}\left[(f'(x) - \hat{f}_k(x))^2 - (f'(x) - f_a(x))^2\right] \leq \hat{\mathbf{E}}_{\boldsymbol{z}}\left[(f'(x) - \hat{f}_k(x))^2 - (f'(x) - \hat{f}_a(x))^2\right]$.

In Theorem 6.1, set $c = 2C^2$. To get approximation within $\epsilon/4$ (with respect to the empirical mean squared error), we require $k \geq 32C^2/\epsilon$. Setting the right hand side of (5.2) to be $\delta/2$ and $k = 32C^2/\epsilon$ and solving for m, we get a sample size bound of

$$\frac{14000C^4}{\epsilon}\left(\frac{32C^2}{\epsilon}\ln \max_{\boldsymbol{x}\in X^{2m}}\left(N\left(\frac{\epsilon}{1024CK}, \mathcal{G}_{|\boldsymbol{x}}, l_1\right) + 1\right) + \frac{32C^2}{\epsilon}\ln 2 + \ln\frac{12}{\delta}\right).$$

Having selected a sample of size m, we now need an algorithm to find $\hat{f}_k$. We show that a proper efficient agnostic learning algorithm for $\mathcal{G}$ can be used as an efficient randomized algorithm for optimizing the error on the sample using $\mathcal{C}^{\mathcal{G}}_{K,k}$. Learning algorithms are often formed from optimization algorithms, and in such cases, the algorithms can be used directly to minimize the error on the sample. The idea is to use the learning algorithm to sample and learn from the empirical distribution so that at each stage i of the iterative approximation, the error relative to the optimum is less than ϵ_i (from Theorem 6.1) with probability greater than $1 - \delta/2k$. This can be done in a way similar to the proof of Theorem 6.3 except that we can test the hypotheses directly using the same sample (and we do not have to compose the resulting function with $clip_T$ at each iteration). Knowing the covering number of $\mathcal{G}$ enables us to bound the size of the sample required to be sampled according to the empirical distribution (Corollary 3.5). Note that since we are sampling from the empirical distribution, no new observations need to be drawn from the original distribution. Theorem 6.1 assures us that if we are successful at each iteration, we will be within the desired error on the empirical distribution which gives us the desired error on the sample. □

6.3 Relationship with Agnostic PAC learning

Let $\mathcal{G}$ be a class of $\{0, 1\}$-valued functions. Let the observed range be $\{0, 1\}$. We call proper agnostic learning with discrete loss under these assumptions *proper agnostic PAC learning*. In this section, we show that if $\mathcal{G}$ is properly efficiently agnostically PAC learnable, then $\mathcal{N}^{\mathcal{G}}_K$ is properly efficiently agnostically learnable (with the squared loss function). Note that $\mathcal{N}^{\mathcal{G}}_K$ has real-valued output with real-valued targets while the algorithm for agnostic PAC learning only handles $\{0, 1\}$-valued function classes with $\{0, 1\}$-valued targets.

As shown by Jones (1992), the iterative approximation result holds even if the inner product of the basis function with $f_k - f$ (where f, the target function is in the closure of the convex hull and f_k is the current network) is minimized instead of the empirical quadratic error. This is also true for the proof given by Koiran (1994) for the case where the target function is not in the closure of the convex hull of the function class. We use this property and transform the problem of minimizing the inner product on a finite set of observations into the problem of agnostic PAC learning.

The following theorem follows from the proof of Theorem 1 given in (Koiran 1994) with minor changes. For completeness we include the proof here.

Theorem 6.6 *Let $\mathcal{G}$ be a subset of a Hilbert space $\mathcal{H}$ with $\| g \| \leq b$ for each $g \in \mathcal{G}$. Let $co(\mathcal{G})$ be the convex hull of $\mathcal{G}$. For any $f \in \mathcal{H}$, let $d_f = \inf_{g' \in co(\mathcal{G})} \| g' - f \|$. Let $f_0 = 0$, $c > 2b + d_f$ and iteratively for $k \geq 1$, suppose f_k is chosen to be $f_k = (1 - 1/k)f_{k-1} + g'/k - f$, where $g' \in \mathcal{G}$ is chosen to satisfy*

$$\langle f_{k-1} - f, g' \rangle \leq \inf_{g \in G} \langle f_{k-1} - f, g \rangle + \epsilon_k$$

and $\epsilon_k \leq \frac{c^2 - (2b+d_f)^2}{k^2}$. Then

$$\| f - f_k \|^2 - d_f^2 \leq \frac{2cd_f}{\sqrt{k}} + \frac{c^2}{k}.$$

Proof. We will show that for any function h in $\mathcal{H}$ and any $\alpha \in [0, 1]$ and $\bar{\alpha} = 1 - \alpha$,

$$\|\alpha h + \bar{\alpha} g - f\|^2 \leq \alpha^2 \|h - f\|^2 + 2\alpha\bar{\alpha} d_f \|h - f\| + \bar{\alpha}^2 (2b + d_f)^2 + \epsilon_k \tag{6.4}$$

where g is chosen to satisfy $\langle h - f, g \rangle \leq \inf_{g' \in G} \langle h - f, g' \rangle + \epsilon_k$. Setting $\alpha = 0$ shows that the result holds for $k = 1$. Assume the desired inequality holds for f_{k-1}. The result then follows by induction. From (6.4), with $\alpha = 1 - 1/k$ and $\bar{\alpha} = 1/k$, we get

$$\|f_n - f\|^2 \leq \frac{(k-1)^2 \|f_{k-1} - f\|^2}{k^2} + \frac{2(k-1)d_f \|f_{k-1} - f\|}{k^2} + \frac{(2b + d_f)^2}{k^2} + \frac{(c^2 - (2b + d_f)^2)}{k^2}.$$

By the induction hypothesis,

$$\|f_k - f\|^2 \quad \leq \quad \frac{(k-1)^2}{k^2} \left[d_f^2 + \frac{2d_f c}{\sqrt{k-1}} + \frac{c^2}{(k-1)} \right]$$

$$+\frac{2d_f(k-1)(d_f+c/\sqrt{k-1})}{k^2}+\frac{c^2}{k^2}$$
$$= \frac{d_f^2(k^2-1)}{k^2}+\frac{2d_fc\sqrt{k-1}}{k}+\frac{c^2}{k}.$$

It follows that

$$\|f_k-f\|^2 \le d_f^2+\frac{2cd_f}{\sqrt{k}}+\frac{c^2}{k}$$

as required. We now verify (6.4).

For any $g \in \mathcal{G}$,

$$\|\alpha h+\bar{\alpha}g-f\|^2 = \alpha^2\|h-f\|^2+\bar{\alpha}^2\|g-f\|^2+2\alpha\bar{\alpha}\langle h-f,g-f\rangle.$$

Given $\delta > 0$, let $f^* \in co(\mathcal{G})$ be such that $\|f^*-f\| \le d_f+\delta$. For some sufficiently large p, f^* is of the form $\sum_{i=1}^p \gamma_i g_i$ with $\gamma_i \ge 0$, $\sum_{i=1}^p = 1$ and $g_i \in \mathcal{G}$. The average value of the inner product $\langle h-f, f-f\rangle$ for $g \in \{g_1,\dots,g_p\}$ is

$$\sum_{i=1}^p \gamma_i\langle h-f,g_i-f\rangle = \langle h-f,f^*-f\rangle \le (d_f+\delta)\|h-f\|. \tag{6.5}$$

Furthermore, for any $g \in \mathcal{G}$,

$$\|g-f\|^2 = \|g-f^*+f^*-f\|^2 \le (\|g\|+\|f^*\|+\|f^*-f\|)^2 \le (2b+d_f+\delta)^2.$$

Hence, with the average of the inner product bounded as in (6.5), if the g chosen as described,

$$\|\alpha h+\bar{\alpha}g-f\|^2 \le \alpha^2\|h-f\|^2+\bar{\alpha}^2[(2b+d_f+\delta)^2]+2\alpha\bar{\alpha}\|h-f\|(d_f+\delta)+2\alpha\bar{\alpha}\epsilon_k.$$

Letting δ go to 0 and noting that $2\alpha\bar{\alpha} \le 1$ completes the proof. □

Theorem 6.7 *Let $\mathcal{G}$ be a class of admissible $\{0,1\}$-valued basis functions. Then $\mathcal{N}_K^{\mathcal{G}}$ is properly efficiently agnostically learnable with the quadratic loss if $\mathcal{G}$ is properly efficiently agnostically PAC learnable.*

Proof. Since the target range is bounded we can easily find a bound for d_f. Using Theorem 6.6, pick the number of basis functions k in the linear combination to obtain approximation $\epsilon/4$ for approximation under the empirical distribution. Then, as in the proof of Theorem 5.2, find the sample complexity so that $\mathbf{E}[(y-f(x))^2-(y-f_a(x))^2] < 2\hat{\mathbf{E}}_{\boldsymbol{z}}[(y-f(x))^2-(y-f_a(x))^2]+\epsilon/2$. Since $\mathcal{G}$ is agnostically PAC learnable, the VC-dimension (hence pseudo-dimension) is polynomial in the complexity parameters (Blumer et al. 1989). Hence, the sample size needs to grow no faster than polynomially in all the desired parameters. Finally we need to show how an algorithm for agnostically PAC learning $\mathcal{G}$ can be used to obtain an efficient randomized optimization algorithm for the selected sample. With the empirical distribution, Theorem 6.6 shows that this can be done by approximating to accuracy ϵ_i at each iteration.

For each iteration i, $1 \leq i \leq k$, we want to find $g \in \mathcal{G}$ to minimize $\langle f_{k-1} - f, wg \rangle = \frac{w}{m}\sum_{i=1}^{m}(f_{k-1}(x_i) - f(x_i))g(x_i)$ for both $w = K$ and $w = -K$ where f is the conditional expectation on the sample under the empirical distribution. For $w = K$ define a function h on the sample such that

$$\begin{array}{ll} h(x_i) = 0 & \text{if } f_{k-1}(x_i) - f(x_i) > 0, \\ h(x_i) = 1 & \text{if } f_{k-1}(x_i) - f(x_i) \leq 0 \end{array}.$$

Thus h is the $\{0,1\}$-valued function which minimizes the inner product. A similar $\{0,1\}$-valued function can be defined for $w = -K$. We will use the agnostic PAC learning algorithm to learn h under a modified distribution.

Let $s = \sum_{i=1}^{m} |f_{k-1}(x_i) - f(x_i)|$. Set up a distribution P on $x_1, \ldots, x_m$ such that

$$P(x_i) = \frac{|f_{k-1}(x_i) - f(x_i)|}{s}.$$

Let $g^* \in \mathcal{G}$ minimize the error for target h and distribution P. Let g' be produced by the agnostic algorithm such that

$$\begin{aligned}
& \Pr(h \neq g') \leq \Pr(h \neq g^*) + \frac{m\epsilon_i}{Ks} \\
\Leftrightarrow \quad & \sum_{h(x_i)=1} \frac{|f_{k-1}(x_i) - f(x_i)|}{s}(h(x_i) - g'(x_i)) + \\
& \sum_{h(x_i)=0} \frac{|f_{k-1}(x_i) - f(x_i)|}{s}(g'(x_i) - h(x_i)) \\
& \qquad \leq \sum_{h(x_i)=1} \frac{|f_{k-1}(x_i) - f(x_i)|}{s}(h(x_i) - g^*(x_i)) + \\
& \qquad \quad \sum_{h(x_i)=0} \frac{|f_{k-1}(x_i) - f(x_i)|}{s}(g^*(x_i) - h(x_i)) + \frac{m\epsilon_i}{Ks} \\
\Leftrightarrow \quad & \sum_{h(x_i)=1} \frac{|f_{k-1}(x_i) - f(x_i)|}{s}(g^*(x_i) - g'(x_i)) -
\end{aligned}$$

$$
\begin{aligned}
& \sum_{h(x_i)=0} \frac{|f_{k-1}(x_i) - f(x_i)|}{s}(g^*(x_i) - g'(x_i)) \leq \frac{m\epsilon_i}{Ks} \\
\Leftrightarrow \quad & \frac{K}{m}\sum_{i=1}^{m}(f_{k-1}(x_i) - f(x_i))(g'(x_i) - g^*(x_i)) \leq \epsilon_i \\
\Leftrightarrow \quad & \langle f_{k-1} - f, Kg' \rangle - \langle f_{k-1} - f, Kg^* \rangle \leq \epsilon_i.
\end{aligned}
$$

Similarly we can show that Kg^* minimizes the inner product $\langle f_{k-1} - f, Kg \rangle$. A similar argument can also be used for $w = -K$. Finally note that Ks/m is bounded by $K(T + K)$. The rest of the proof follows in a manner similar to Theorem 6.5. □

6.4 Hardness Results

While Theorem 6.7 is interesting in relating agnostic PAC learning to learning a single hidden layer neural network, there do not appear to be many basis function classes which are properly efficiently agnostically PAC learnable. Available results show the hardness of properly agnostically PAC learning monomials and halfspaces under the assumption $RP \neq NP$ (Kearns et al. 1994, Höffgen & Simon 1992). This implies that for networks of functions from these classes, it is unlikely that an efficient algorithm can be obtained from the approach given here. Since the quadratic loss is equivalent to the discrete loss when the function class as well as the target functions are $\{0, 1\}$-valued, the approach given in Section 6.2 for properly learning networks by properly learning the basis functions with the quadratic loss is also unlikely to produce an efficient agnostic learning algorithm for these function classes. However, these results do not rule out efficient agnostic learning using other methods or other hypothesis classes. To do that requires representation independent hardness results.

In (Kearns et al. 1994), it was shown that if the class of monomials is efficiently agnostically learnable (with any hypothesis class) with respect to the discrete loss function, then the class of polynomial-size DNF is efficiently learnable in the PAC learning model. (It is generally believed that polynomial-sized DNF is not likely to be efficiently learnable (Jerrum 1994).) Using techniques similar to that in (Kearns et al. 1994), it is possible to show that if a class of $\{0, 1\}$-valued basis functions include monomials, then an efficient agnostic learning algorithm for the class using the quadratic loss function can be used to efficiently find a *randomized hypothesis* for polynomial-sized DNF. (We say a hypothesis h is randomized if there exists a probabilistic polynomial time algorithm that, given h and an instance v, computes h's prediction on v.) If we assume that it is hard to find a learning algorithm for DNF, then agnostically learning such basis function classes as well as the network of the basis functions is hard.

The idea behind the proof is to show that the network can be used as a weak PAC learning algorithm for learning $p(n)$-term DNF. The result then follows from the fact that a $\{0,1\}$-valued function class is efficiently PAC learnable if and only if it is efficiently weakly PAC learnable (Schapire 1990).

Definition 6.8 *The class of monomials over n Boolean variables $x_1, \ldots, x_n$ consist of all conjunctions of literals over the variables. For any k, the class of k-term DNF consist of all disjunctions of the form $M_1 \vee \ldots \vee M_k$ where each M is a monomial.*

Definition 6.9 *Let $\mathcal{G}$ be a class of functions mapping from $\mathcal{X}$ to $\{0,1\}$. Suppose $\mathcal{G}$ is parametrized by complexity parameter n. Then $\mathcal{G}$ is efficiently weakly PAC learnable if there exists a polynomial p and an algorithm A such that for all $n \geq 1$, for all target functions $g \in \mathcal{G}$, for any probability distribution D on $\mathcal{X}$, and for all $0 < \delta \leq 1$, algorithm A, given the parameters n and δ, draws instances from D labelled by g, runs in time polynomial in n and $1/\delta$, and outputs a hypothesis h that with probability at least $1-\delta$ has expected error no more than $1/2 - 1/p(n)$.*

Theorem 6.10 ((Schapire 1990)) *Let $\mathcal{G}$ be a function class mapping from $\mathcal{X}$ to $\{0,1\}$. $\mathcal{G}$ is efficiently weakly PAC learnable if and only if it is efficiently PAC learnable.*

For any function $h : \mathcal{X} \to [0,1]$, define $\$_h(x)$ to be a boolean random variable that is 1 with probability $h(x)$ and 0 with probability $1 - h(x)$. We will need the following result.

Lemma 6.11 ((Kearns et al. 1994)) *Let $f : \mathcal{X} \to \{0,1\}$ be any boolean function, and let $h : \mathcal{X} \to [0,1]$ be a real-valued function. Then for any distribution D on $\mathcal{X}$*

$$\Pr(f(x) \neq \$_h(x)) \leq \mathbf{E}[(f(x) - h(x))^2] + 1/4.$$

Theorem 6.12 *Let $\mathcal{G} = \bigcup_{n=1}^{\infty}$ where each $\mathcal{G}_n$ is a permissible class of $\{0,1\}$-valued functions on $\mathbb{R}^n$ such that the class of monomials is a subset of $\mathcal{G}_{|\{0,1\}^n}$ and let $p(n)$ be any polynomial in n. If $\mathcal{G}$ is efficiently agnostically learnable with respect to the quadratic loss function, then there exists an efficient algorithm (which produces randomized hypotheses) for learning $p(n)$-term DNF.*

Proof. We will show that there exists a weak learning algorithm (which produces randomized hypotheses) for $p(n)$-term DNF. The result then follows from Theorem 6.10.

For any target $p(n)$-term DNF formula, there exists a monomial that never makes an error on a negative example and gets at least $1/p(n)$ of the positive examples right (because the $p(n)$ terms cover all the positive examples). Let $\omega \in \mathcal{G}$ be equivalent to this monomial when restricted to $\{0,1\}^n$. Then $\omega' = \frac{1}{2}(\omega + 1) \in \mathcal{N}_1^{\mathcal{G}}$ will have quadratic error $1/4$ on the negative examples. On the positive examples the quadratic error of ω' will be zero when the monomial ω gives the correct classification and $1/4$ when it gives the wrong classification.

The algorithm for producing the randomized hypothesis goes as follows. (The constants are chosen for convenience.) Assume that the probability that an instance is labelled 1 is α. Draw a large enough sample (using e.g. Theorem 6.2) so that with probability at least $1 - \delta/2$, the empirical average $\hat{\alpha}$ is within $\epsilon/2$ of α for some small $\epsilon < 7/(32p(n))$. If the empirical average is less than $1/4 + \epsilon/2$, choose the all zero monomial. If the empirical average is more than $3/4 - \epsilon/2$ choose the the all one monomial. Either of these hypotheses will then have error no more than $1/4 + \epsilon$. Otherwise the probability of a positive example is between $1/4$ and $3/4$. We then use the agnostic learning algorithm to learn the function using $\mathcal{N}_1^{\mathcal{G}}$ with quadratic loss. From Section 6.2, $\mathcal{N}_1^{\mathcal{G}}$ is efficiently agnostically learnable if $\mathcal{G}$ is efficiently agnostically learnable. The above argument shows that there exists a function in $\mathcal{N}_1^{\mathcal{G}}$ with expected quadratic error less than $\frac{1}{16}\left(1 - \frac{1}{p(n)}\right) + \frac{1}{4}\left(1 - \frac{1}{4}\right) = \frac{1}{4} - \frac{1}{16p(n)}$. Let f be our target DNF. Use the agnostic algorithm to produce a hypothesis h which is no more that $1/(32p(n))$ away from the optimum with probability at least $1 - \delta/2$. Then from Lemma 6.11, we have $\Pr[f(x) \neq \$_h(x)] \leq \mathbf{E}[(f(x) - h(x))^2] + 1/4 < 1/2 - 1/(32p(n))$, where $\$_h(x)$ is a boolean random variable that is 1 with probability $h(x)$ and zero with probability $1 - h(x)$. The probability of the algorithm failing to produce a hypothesis with error less than $1/2 - 1/(32p(n))$ is no more than δ. Hence the algorithm is a weak learning algorithm which produces randomized hypotheses for learning $p(n)$-term DNF. □

It is easy to see that the result also holds in the logarithmic cost model of computation (Aho et al. 1974) because the second layer weights of the neural network are fixed at $1/k$, where k is the number of hidden units.

6.5 Learning Bounded Fan-in Neural Networks

From the result in the previous section, it would appear that agnostically learning a single hidden layer neural network with linear threshold hidden units is likely to be computationally difficult. In this section, we consider learning a computationally tractable subclass: the class of single hidden layer neural networks with bounded fan-in.

6.5.1 Sample Complexity

We first bound the sample complexity for learning networks with bounded fan-in. Let the basis function class be

$$\mathcal{G}_\tau := \{x \mapsto h(v_i \cdot x + v_{i0}) : \text{at most } \tau \text{ of the coordinates } v_{ij} \text{ of } v_i \text{ are nonzero}\}$$

where h is the step function, $h(z) = 1$ for $z > 0$ and $h(z) = 0$ otherwise, $x \in \mathbb{R}^n$, and $v_i \cdot x = \sum_{j=1}^{n} v_{ij} x_j$.

Lemma 6.13 *Let $\boldsymbol{x}$ be a sequence of points from $\mathbb{R}^n$. Then for $0 < \epsilon \leq 1$,*

$$N(\epsilon, \mathcal{G}_{\tau|\boldsymbol{x}}, l_1) \leq 2n^\tau \left(\frac{2e}{\epsilon} \ln \frac{2e}{\epsilon}\right)^\tau.$$

Proof. The class $\mathcal{L}$ of linear threshold functions in τ dimensions has pseudo-dimension $\tau + 1$. From Lemma 3.8, for any sequence of points $\boldsymbol{x}$ the covering number $N(\epsilon, \mathcal{L}_{|\boldsymbol{x}}, l_1) \leq 2\left(\frac{2e}{\epsilon} \ln \frac{2e}{\epsilon}\right)^\tau$. A cover for $\mathcal{G}_\tau$ can be formed from the union of $\binom{n}{\tau} \leq n^\tau$ such covers. □

Corollary 6.14 *The sample complexity for efficiently agnostically learning $\mathcal{N}_K^{\mathcal{G}_\tau}$ is bounded by*

$$\frac{14000C^2}{3\epsilon}\left(\frac{32C^2\tau}{\epsilon}\left(\ln\left(\frac{2048eC^2n}{\epsilon}\ln\frac{2048eC^2}{\epsilon}+1\right)+\right)+\frac{64C^2}{\epsilon}\ln 2+\ln\frac{12}{\delta}\right)$$

where $C = \max\{K, T, 1\}$.

Proof. The proof follows from Theorem 6.5 and Lemma 6.13. □

6.5.2 Loading Algorithm

Having bounded the sample complexity, we still need to find an algorithm which will produce a network which will give the required approximation. In this section we describe an algorithm *CONSTRUCT* which produces a network by iteratively adding hidden units as suggested by the approximation result in Section 6.1. Note that in Theorem 6.1, we have a fixed set of output layer weights for each iteration. The algorithm receives as inputs the number of iterations

SPLITTING(A)

$\mathcal{W} := \emptyset$;
$P := \emptyset$;
for all τ-tuples $(t_1, \ldots, t_\tau)$ from $\{1, \ldots, n\}$ with $t_1 < \ldots < t_\tau$
  for all $(\tau+1)_{\leq}$-tuples $(r_1, \ldots, r_l)$ from $\{1, \ldots, m\}$ with $r_1 < \ldots < r_l$
    for all l-tuples $(\alpha_1, \ldots, \alpha_l) \in \{-1, 1\}^l$
      if a solution
        $(v_0, v_{0_t}) \in \{(v, v_t) \in \mathbb{R}^{n+1}$: only $v_{t_1}, \ldots, v_{t_\tau}, v_t$ are nonzero$\}$
        to the system of linear equations

$$x_{r_v} \cdot v + v_t = \alpha_\nu \qquad \nu = 1, \ldots, l$$

      exists **then**
        Split A into two subsets A' and A'' by the hyperplane given
          by $v_0 \cdot x = -v_{0_t}$;
        if $\{A', A''\} \notin P$ **then**
          $\mathcal{W} := \mathcal{W} \cup \{(v_0, v_{0_t})\}$;
          $P := P \cup \{\{A', A''\}\}$;
        endif;
      endif;
    endfor;
  endfor;
endfor;
return $\mathcal{W}$;

Figure 6.2: Subroutine *SPLITTING*

k, the bound on the sum of magnitudes of output weights K, the fan-in τ and a sequence $S := \{(x_1, y_1), \ldots, (x_m, y_m)\}$ from $\mathcal{X} \times \mathcal{Y}$. At each iteration, the algorithm generates all possible dichotomies of the sample with a bounded fan-in hidden unit and then adds the hidden unit which minimizes the empirical loss at each stage.

We first describe a subroutine *SPLITTING* for generating all possible dichotomies (Figure 6.2). The subroutine and proof of correctness are adapted from (Farago & Lugosi 1993). The input to the subroutine is a set $A := \{x_1, \ldots, x_m\}$. It returns a set $\mathcal{W}$ of weight (and bias) vectors which correspond to all possible dichotomies on A. For notational convenience, an $m_{\leq}$-tuple means an l-tuple for some $1 \leq l \leq m$.

Lemma 6.15 *The subroutine SPLITTING generates all possible dichotomies of m points using a linear threshold unit with τ or fewer nonzero weights in $\mathbb{R}^n$ in $O(2^\tau \tau^3 n^{2\tau} m^{2(\tau+1)})$ steps.*

Proof. The algorithm goes through $\binom{n}{\tau} \sum_{i=1}^{\tau+1} 2^i \binom{m}{i} \leq n^\tau 2^{\tau+1} m^{\tau+1}$ iterations of the innermost loop where it does comparisons and solves linear equations. Each set of linear equations has no more than $\tau + 1$ variables and no more than $\tau + 1$ equations. By Gaussian elimination solving

CONSTRUCT(k,K,τ,S)

> $f := 0$;
> $f' := 0$;
> $\mathcal{W}$:= *SPLITTING*(S_A);
> $\mathcal{G} := \{x \mapsto \omega\sigma(v \cdot x - v_0)\colon v \in \mathcal{W}, |\omega| = K\} \cup \{K, -K\}$;
> **for** $j := 1$ **to** k
>   **for each** $g \in \mathcal{G}$
>   **if** *COST*$((1 - 2/(j+1))f + 2g/(j+1), S)$ < *COST*(f', S) **then**
> $f' := (1 - 2/(j+1))f + 2g/(j+1)$;
>   **endif**;
> $f := f'$;
>   **endfor**;
> **endfor**;
> **return** f;

Figure 6.3: Algorithm *CONSTRUCT*

each system takes $O(\tau^3)$ operations. Each comparison against P takes $O(n^\tau m^{\tau+1})$. So the total number of operations is $O(2^\tau \tau^3 n^{2\tau} m^{2(\tau+1)})$.

We now show that all dichotomies are generated. Note that a dichotomy generated by a unit with fewer than τ weights can also be generated with a unit with τ weights. All possible combinations of τ inputs are generated. So we only need to ensure that all possible dichotomies of a unit with a set of τ weights are generated. This is the same as considering a linear threshold unit in τ dimensions.

First note that any dichotomy in $\mathbb{R}^\tau$ can be implemented by a hyperplane of the form $b \cdot x + b_t = 0$ with some $(b, b_t) \in \mathbb{R}^{\tau+1}$. For any $x_i \in A$, either $b \cdot x_i + b_t > 0$ or $b \cdot x_i + b_t < 0$. A suitable b can be found by replacing > 0 and < 0 by ≥ 1 and ≤ -1 and solving the arising system of inequalities[1] for (b, b_t). Let H_1 and H_2 be the two open halfspaces generated by the hyperplane. Then given a partition $A \cap H_1$, $A \cap H_2$ of A, an appropriate (b, b_t) can be found by taking any solution to the linear system of inequalities

$$\begin{aligned} x_i \cdot v + v_t &\geq 1, \qquad x_i \in A \cap H_1 \\ x_j \cdot v + v_t &\leq -1, \qquad x_j \in A \cap H_2 \end{aligned}$$

This system of inequalities defines a polyhedron in weight space. It is possible to select l

[1]Farago & Lugosi (1993) set b_t to -2 and then replace the inequalities with $b \cdot x \geq 3$ and $b \cdot x \leq 1$ but that gives an incorrect result. For example, let $x_1 = (1, 0)$ be labelled 1, $x_2 = (1, 1)$ be labelled 0 and $x_3 = (0, 1)$ be labelled 1. Although the dichotomy can be implemented by a b such that $b \cdot x = 2$, there is no b that satisfies $b \cdot x_1 \geq 3$, $b \cdot x_2 \leq 1$ and $b \cdot x_3 \geq 3$.

inequalities for some $l \leq \tau + 1$ such that they can be satisfied if and only if the polyhedron is nonempty and every solution of the arising system of linear equation also satisfies the whole system of inequalities (see (Farago & Lugosi 1993)). Since we are considering all possible systems of $\tau + 1$ or fewer equations, all possible dichotomies are generated by the algorithm. $\Box$

The pseudo-code for the algorithm *CONSTRUCT* is given in Figure 6.3. It receives as inputs the number of iterations k, the bound on the sum of magnitudes of output weights K, the fan-in τ and a sequence $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$. At each iteration j, we multiply the previous network f by $1 - 2/(j+1)$ and add the function $2g/(j+1)$ which minimizes *COST*$((1 - 2/(j+1))f + 2g/(j+1), S)$ over all g given by the subroutine *SPLITTING* (here *COST*$((1 - 2/(j+1))f + 2g/(j+1), S)$ is the sum of squared errors on the sequence S). The time complexity of the algorithm is $O(2^\tau \tau^3 k n^{2\tau} m^{2(\tau+1)})$.

Since the algorithm minimises the error at each iteration, the algorithm is an agnostic learning algorithm if the sample size is chosen according to Corollary 6.14. Since the computation time is polynomial in all the parameters (with the fan-in fixed) the algorithm is efficient.

If the hidden unit used is not the linear threshold unit but has a Lipschitz bound, the parameters can be discretized with an appropriate grid size (instead of obtaining all dichotomies, see (Barron 1994)) to get a similar result.

6.6 Discussion

We have shown that if a basis function class is efficiently agnostically learnable, then a single hidden layer neural networks with hidden units from the basis function class is efficiently agnostically learnable. This is done by iteratively learning one function of the linear combination at a time and using the agnostic learning algorithm for the basis function class as a subroutine. For many function classes, for example classes of functions with finite pseudo-dimension, the sample complexity for agnostically learning the class of single hidden layer neural networks is not much worse than the sample complexity for properly agnostically learning the basis function class. Another advantage of the iterative approximation approach is the ease of learning linear combinations of functions from more than one basis function class. In the fixed network approach common in the neural network literature, the number of basis functions from each basis function class has to be specified in advance. For linear combinations of k basis functions from s basis function classes, this leads to $(k + s - 1)!/k!(s - 1)! = \Omega(s^k)$ (when k is fixed) possible combinations to choose from.

In contrast, the time complexity of the greedy iterative approximation approach is approximately s times the time taken to learn the linear combinations of the most difficult basis function class (ignoring increased sample complexity). Since the covering number of the union of s basis function classes is bounded by s times the largest covering number of function classes in the union, the sample size needs to grow only logarithmically with s, assuming a similar covering number for each basis function class.

Unfortunately, our results also indicate that agnostic learning of single hidden layer neural networks is likely to be computationally difficult for many interesting basis function classes. However, we should note that the agnostic learning framework is a very demanding framework and it is possible that learning many of these function classes may not be difficult under other realistic assumptions.

Within the agnostic learning framework, we have shown that the class of single hidden layer neural networks with bounded fan-in (with linear threshold hidden units) is efficiently learnable. In fixed dimension or with bounded fan-in, the class of linear combinations of axis parallel rectangles with bounded sum of magnitudes of weights is also efficiently agnostically learnable. These results are generalized in the following corollary which is particularly useful for $\{0, 1\}$-valued basis function classes.

Corollary 6.16 *Let $\mathcal{G}$ be an admissible basis function class. Let $x = (x_1, x_2, \ldots, x_m)$ be an arbitrary sequence of points from $\mathcal{X}$. If $\mathcal{G}_{|x}$ can be enumerated in time polynomial in m and the complexity parameters, then $\mathcal{N}_K^{\mathcal{G}}$ is properly efficiently agnostically learnable.*

Proof. The covering number is bounded by the number of functions in $\mathcal{G}_{|x}$ which is polynomial in m and the complexity parameter. Since the functions can be efficiently enumerated, choosing the function which minimizes the loss on a large enough (but polynomial) sample size will result in an efficient learning algorithm for $\mathcal{G}$. □

By having bounded fan-in, we lose some of the approximation capabilities of these networks. For boolean functions, Minsky & Papert (1969) have shown that some functions such as parity cannot be learned (or even well approximated) by two layer networks with bounded fan-in. However, the class of functions that can be approximated by such networks is still interesting and useful. For example, Boser et al. (1992) have shown that even with networks with bounded fan-in (low degree polynomials), good results can be achieved for the task of handwritten digit recognition. Hence, we think that the study of such subclasses of efficiently learnable functions is worthwhile.

[W]e have seen a rabble of functions arise whose only job, it seems, is to look as little as possible like decent and useful functions. No more continuity, or perhaps continuity but no derivatives. Moreover, from the point of view of logic, it is these strange functions which are most general; whilst those one meets unsearched for and which follow simple laws are seen just as very special cases to be given their own tiny corner.

— Henri Poincaré

Collected Works, Vol. 11, p. 130.

Chapter 7

Learning Smooth Functions

In Chapter 6, we showed that networks which can approximate monomials are unlikely to be efficiently agnostically learnable. This means that fairly severe conditions have to be imposed on function classes in order for them to be efficiently agnostically learnable. In this chapter, we study how smoothness of the function class affects efficient agnostic learning.

Barron (1993) has shown that the class of functions with bounded first absolute moment of the Fourier transform can be learned with sample complexity $O\left(\frac{1}{\epsilon}\left(\frac{1}{\epsilon}\ln\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$. However, it is not known if this function class is efficiently learnable (computationally). In this chapter, we put more restrictions on the function class by demanding that the q-th absolute moment of the Fourier transform of the functions (where q depends linearly on the input dimension n) and the L_1 norm of the functions in the class be uniformly bounded. We show that such a function class is efficiently agnostically learnable.

Previous work in nonparametric statistics on learning functions in high input dimensions usually does not concentrate on the computational complexity but rather on the sample size

Basis Functions	*Agnostic learning*	*Learning with noise*
Sinusoidal	$O\left(\epsilon^{-\left(2+\frac{2n}{q}\right)}\left(\ln\left(\frac{1}{\epsilon}\right)+\ln\frac{1}{\delta}\right)\right)$	$O\left(\epsilon^{-\left(1+\frac{n}{q}\right)}\left(\ln\left(\frac{1}{\epsilon}\right)+\ln\frac{1}{\delta}\right)\right)$
Linear threshold	$O\left(\epsilon^{-\left(2+\frac{2n+2}{q}\right)}\left(\frac{1}{q}\ln\left(\frac{1}{\epsilon}\right)+\ln\frac{1}{\delta}\right)\right)$	$O\left(\epsilon^{-\left(1+\frac{n+1}{q}\right)}\left(\frac{1}{q}\ln\left(\frac{1}{\epsilon}\right)+\ln\frac{1}{\delta}\right)\right)$
Sigmoid	$O\left(\epsilon^{-\left(2+\frac{2n+2}{q}\right)}\left(\frac{1}{q}\ln\left(\frac{1}{\epsilon}\right)+\ln\frac{1}{\delta}\right)\right)$	$O\left(\epsilon^{-\left(1+\frac{n+1}{q}\right)}\left(\frac{1}{q}\ln\left(\frac{1}{\epsilon}\right)+\ln\frac{1}{\delta}\right)\right)$

Table 7.1: Number of basis functions for used for efficiently learning the class of functions with bounded q-th absolute moment of the Fourier transform (q and n fixed).

required (stated in terms of the rate of convergence of the risk of the estimator as a function of the sample size). Kernel methods are computationally efficient in high dimensions and give good rates of convergence for certain classes such as the class of s-times differentiable functions when the s-th derivative is Hölder continuous and s is proportional to n (see (Hardle 1990)). However, unlike our framework, which requires bounds to hold for arbitrary input distributions, the bounds for kernel methods depend on the input distribution. For functions where all partial derivatives of order s are square-integrable, the asymptotic minimax rate of convergence of the mean integrated squared error is $O\left((1/m)^{2s/(2s+n)}\right)$ (Ibragimov & Hasminskii 1980, Pinsker 1980, Stone 1982, Nussbaum 1986), where m is the sample size and n is the input dimension, and this rate can be achieved by using a linear combination of fixed basis functions. With s of order n, learning can be done with a reasonable sample size. However, to achieve this rate, an exponential number (with respect to the input dimension) of basis functions is used.

To obtain our results for functions with bounded Fourier transform moments, we use a Monte Carlo method to evaluate the function via the inverse Fourier transform. For computational efficiency, we multiply the Fourier transform with an appropriate sized uniform window and evaluate the resulting inverse Fourier transform (integral) by sampling uniformly over the appropriate subset of the parameter space. Because we are using the Fourier transform, our hypothesis class consists of linear combinations of sinusoidal basis functions. Similar results can be achieved using linear threshold basis functions and sigmoid basis functions. The sample complexity is bounded by $O\left(\frac{1}{\epsilon}\left(k\ln\left(\frac{1}{\epsilon}\right)+\ln\frac{1}{\delta}\right)\right)$ where k, the number number of basis functions used, is shown in Table 7.1.

The results are obtained in the agnostic learning framework. However, as shown in Table 7.1, we are able to obtain a better bound for the case of learning with noise, where the target conditional expectation satisfies the assumptions we are using.

For the class of functions with a uniform bound on the q-th absolute moment of the Fourier series and a uniform bound on the L_1 norms of the functions (with q growing linearly with n), we are also able to show that for a desired accuracy of approximation, the size of a fixed set of basis functions which will provide the required approximation to all the functions in the class grows only polynomially (instead of exponentially) with the input dimension. The fact that the set of basis functions is fixed means that for multi-output networks, the number of hidden units does not need to grow as the number of outputs grow. This is interesting because in most neural network applications, all the different outputs of the network share the same hidden units.

In Section 7.1, we describe the class of functions with bounded q-th moment of the Fourier transform. We state the results and describe the algorithm used in Section 7.2. We discuss the results on learning smooth functions in Section 7.3.

In Section 7.4, we show the existence of small (polynomial size) sets of fixed basis functions that can be used to uniformly approximate all the functions with uniform bounds on q-th absolute moment of the Fourier series and the L_1 norm (q growing linearly with n).

7.1 Functions with Bounded q-th Absolute Moment of the Fourier Transform

We will restrict the domain to $[-\pi, \pi]^n$. Any bounded subset of $\mathbb{R}^n$ can be rescaled to be within this domain. For $T, M, C \in \mathbb{R}^+$, let Γ_q be the class of functions satisfying the following conditions:

1. $|f(x)| \leq T$ for all $x \in [-\pi, \pi]^n$
2. $\int_{\mathbb{R}^n} |f(x)| dx \leq M$
3. $\int_{\mathbb{R}^n} \sum_{j=1}^{n} |2\pi u_j|^q |F(u)| du \leq C$ where $F(u) = \int_{\mathbb{R}^n} f(x) e^{-i2\pi u \cdot x} dx$ is the Fourier transform of f and u_j is the jth component of u.

Functions on a bounded domain can be represented as a Fourier series by having a periodic extension outside the domain. However, for a condition similar to (3) on the Fourier series to be satisfied, the functions and their derivatives have to be continuous on the boundary of the domain. By having a Fourier transform representation, the functions do not have to satisfy the boundary conditions.

Using techniques from (Barron 1993), it is possible to show that if all partial derivatives of order less than or equal to $s = \lfloor n/2 \rfloor + q + 1$ of a function f are square-integrable, then it satisfies the moment condition for Γ_q. Write $|u_j|^q|F(u)| = a(u)b(u)$ with $a(u) = (1+|u|^{2t})^{-1/2}$ and $b(u) = |u_j|^q|F(u)|(1+|u|^{2t})^{1/2}$. By the Cauchy-Schwarz inequality, $\int a(u)b(u)du \leq (\int a^2(u)du)^{1/2}(\int b^2(u)du)^{1/2}$. The integral $\int a^2(u)du = \int(1+|u|^{2t})^{-1}du$ is finite for $2t > n$. By Parseval's theorem the integral $\int b^2(u)du = \int |F(u)|^2(|u_j|^{2q} + |u_j|^{2q}|u|^{2t})du$ is finite when the partial derivatives of f of order $t+q$ and of order q are square-integrable on $\mathbb{R}^n$. This relates the class Γ_q to more traditional smoothness classes considered in nonparametric statistics.

7.2 *Results and Algorithms*

The results are stated in the following theorems.

Theorem 7.1 *Let $\epsilon \leq T$ and $B = M + C + T$. The function class Γ_q is efficiently agnostically learnable using single hidden layer neural networks with sample complexity (for fixed n and q)*

$$O\left(\frac{B^2}{\epsilon}\left(k \ln\left(\frac{B^2}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right)$$

where

$$k = O\left(\frac{(TC^2)^{\frac{2n}{q}} M^2(T+M+C)}{\epsilon^{2+\frac{2n}{q}}}\left(\frac{n^2}{q} ln\left(\frac{TCM}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right)$$

if sinusoidal basis functions are used as hidden units and

$$k = O\left(\frac{(TC^2)^{\frac{2n+2}{q}} M^2(T+M+C)}{\epsilon^{2+\frac{2n+2}{q}}}\left(\frac{n^2}{q} \ln\left(\frac{TCM}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right)$$

if either linear threshold functions or sigmoid functions are used as hidden units.

Theorem 7.2 *Let $\epsilon \leq T$ and $B = M + C + T$. The function class Γ_q is efficiently learnable with noise using single hidden layer neural networks with sample complexity (for fixed n and q)*

$$O\left(\frac{B^2}{\epsilon}\left(k \ln\left(\frac{B^2}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right)$$

where

$$k = O\left(\frac{C^{\frac{2n}{q}} M^2}{\epsilon^{1+\frac{n}{q}}}\left(\frac{n^2}{q} \ln\left(\frac{CM}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right)$$

if sinusoidal basis functions are used as hidden units and

$$k = O\left(\frac{C^{\frac{2n+2}{q}} n^2 M^2}{\epsilon^{1+\frac{n+1}{q}}}\left(\frac{n^2}{q}\ln\left(\frac{CM}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right)$$

if either linear threshold functions or sigmoid functions are used as hidden units.

The algorithms for achieving these results are given in the next subsection. The proofs that the algorithms give the required bounds are given in Appendix B.

7.2.1 The Algorithms

The pseudo-code for the basic algorithm is shown in Figure 7.1. The algorithm needs to use M, C, T and q. Explicit bounds on the number of observations, the size of $\mathcal{W}$ and the number of basis functions k can be calculated (see Appendix B). To be able to do the Monte Carlo integral, we first approximate the Fourier transform by multiplying it with an appropriate uniform window. The set $\mathcal{W}$ is a subset of the parameters of the basis function class which results from the windowing procedure. It depends on the basis function class, M, C, T, q and is described below. The number k of basis functions required is given in Theorem 7.1 and Theorem 7.2. Note that constrained mean squared optimization of a linear combination of fixed basis function can be done efficiently (Nestorov & Nemirovskii 1994).

Sinusoidal basis functions

Each parameter U_i drawn from $\mathcal{W}$ is used to parametrize two basis functions $x \mapsto \cos(2\pi U_i \cdot x)$ and $x \mapsto \sin(2\pi U_i \cdot x)$. Here the set $\mathcal{W} := \{u \in \mathbb{R}^n : |u_i| \leq r\}$ where $r = O\left(\frac{(TC^2)^{1/q}}{\epsilon^{1/q}}\right)$ for agnostic learning and $r = O\left(\frac{C^{1/q}}{\epsilon^{1/2q}}\right)$ for learning with noise.

SMOOTHLEARN(M,C,T,k,$\mathcal{W}$)

1. *Select a set of k parameters by uniformly sampling from $\mathcal{W}$. The set of k parameters is used to parametrize a set of basis functions $\mathcal{G}$.*
2. *Let $B = M + C + T$. Draw $O\left(\frac{B^2}{\epsilon}\left(k\ln\left(\frac{B^2}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right)$ observations.*
3. *Do constrained mean squared optimization over the empirical loss on the observations in Step 2 to accuracy of order ϵ using the linear combinations of all the basis functions from $\mathcal{G}$ as the hypothesis class. The constraints to be satisfied are $f(X_i) \leq B$ for all the observations, where f is the hypothesis produced.*

Figure 7.1: Pseudo-code for the algorithms for learning Γ_q.

Linear threshold basis functions

Each parameter (U_i, τ_i) drawn from $\mathcal{W}$ is used to parametrize a basis function $x \mapsto h(2\pi U_i \cdot x - \tau_i)$ where h is the threshold function. Here $\mathcal{W} := \{(u,t) \in \mathbb{R}^{n+1} \colon |u_i| \leq r, |t| \leq 2\pi^2 nr\}$ where $r = O\left(\frac{(TC^2)^{1/q}}{\epsilon^{1/q}}\right)$ for agnostic learning and $r = O\left(\frac{C^{1/q}}{\epsilon^{1/2q}}\right)$ for learning with noise. An additional basis function $g(x) \equiv 1$ has to be added to the set of basis functions.

Sigmoid basis functions

Each parameter (U_i, τ_i) drawn from $\mathcal{W}$ is used to parametrize a basis function $x \mapsto \sigma(\alpha(2\pi U_i \cdot x - \tau_i))$ where σ is the sigmoid function. Here $\mathcal{W} := \{(u,t) \in \mathbb{R}^{n+1} \colon |u_i| \leq r, |t| \leq 2\pi^2 nr\}$ where $r = O\left(\frac{(TC^2)^{1/q}}{\epsilon^{1/q}}\right)$ for agnostic learning and $r = O\left(\frac{C^{1/q}}{2\pi\epsilon^{1/2q}}\right)$ for learning with noise. An additional basis function $g(x) \equiv 1$ has to be added to the set of basis functions. The value of α depends on ϵ and is of order $(T + M + C)(M + C)/\epsilon$.

The basic idea behind the proof (given in Appendix B) is to separate the error into three components and to bound them separately. The three components are the component caused by windowing the Fourier transform, the component caused by approximating the windowed Fourier transform by the sampling procedure and the component caused by estimating the parameters of the linear combinations of basis functions from the sample.

7.3 Discussion on Learning Smooth Functions

The L_1 norm of the function (over the domain $\mathbb{R}^n$) is used to bound the magnitude of the Fourier transform. This in turn bounds the variance of the random variable for the Monte Carlo approximation of the Fourier transform. The L_1 norm of the function appears to be an important parameter for *efficient* agnostic learning. For example, function classes which appear to be difficult to learn, such as polynomial sized DNF, do not have a uniform polynomial bound on the L_1 norm (over the appropriate domain).

From Chapter 5, we see that the sample complexity for agnostically learning Γ_1 is $O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon} + \log\frac{1}{\delta}\right)\right)$. The functions Γ_q are subsets of Γ_1 and hence can be learned with the same sample complexity. However the order of the sample complexity for our algorithm (the agnostic version) is $1/\epsilon^{3+\frac{2n}{q}}$ (ignoring log factors). This shows that we may be trading off some of the sample complexity to be able to learn efficiently.

The algorithms used for learning Γ_q choose the basis functions randomly from a uniform distribution over an appropriate sized subset in the parameter space. This is similar to the way that the initial set of parameters are chosen for the basis functions for some gradient descent learning algorithms such as backpropagation (Rumelhart, Hinton & Williams 1986) which optimize over all parameters including those which parametrize the basis functions. This suggests that for smooth enough target functions with small L_1 norm, if the number of the hidden units is large enough, gradient descent algorithms which randomly choose the initial values of the parameters are likely to find local minimums which will perform well. Even if the target function is not smooth, the fact that the algorithms perform well in the agnostic model suggests that if a good smooth approximation to the target exists, the gradient descent algorithm will still perform well. (These suggestions are strictly true only if the output weights are learned first with the hidden units fixed before all the weights are optimized together and if the hidden units are chosen as described earlier in this chapter.)

Similar work on learning by randomly drawing basis functions has been done by (Delyon, Juditsky & Benveniste 1995). They used the wavelet transform and learned the transform from the data before drawing the basis functions from the distribution induced by the learned transform. However, they did not consider computational complexity issues. We have not considered using the data to learn the transform. It would be interesting to know what advantage such a method will offer.

Igelnik & Pao (1995) have examined a similar scheme to that described in this chapter where the basis functions are drawn uniformly from a suitable subset of the parameter space. However, they did not consider any specific function class and did not study the rate at which the size of the subset must increase to give bounds on the performance of the algorithm. They claimed that the method worked well for several practical applications.

7.4 Small Set of Fixed Basis Functions

In this section, we give upper bounds on the size of a fixed set of basis functions, linear combinations of which can approximate smooth functions with a given accuracy. By a set of fixed basis functions, we mean that the same set of basis functions can be used to give the required approximation to all the functions in the class i.e. we do not require different sets of basis function for different functions in the class.

Let Γ_q^s be the class of functions on $[-\pi, \pi]^n$ which satisfy the following conditions:

1. $\int_{[-\pi,\pi]^n} |f(x)|dx \leq M$

2. $\sum_{u\in\mathbb{Z}^n}\sum_{j=1}^{n} |2\pi u_j|^q|F(u)| \leq C$ where $F(u) = \int_{[-\pi,\pi]^n} f(x)e^{-i2\pi u\cdot x}dx$ is the Fourier coefficient of f at u, u_j is the jth component of u and $\mathbb{Z}$ is the set of integers.

Since functions in Γ_q^s have Fourier series representations, they have periodic extensions outside the domain $[-\pi,\pi]^n$. We do not consider the class Γ_q defined in Section 7.1 because it is technically more difficult.

The approximation provided by our fixed basis function is in the sup-norm sense i.e. $|f_k(x) - f^*(x)| \leq \epsilon$ for all $x \in [-\pi,\pi]^n$, where f_k is the approximating function and $f^* \in \Gamma_q^s$ is the target function. By truncating the Fourier series appropriately (by multiplying it with a window with sides of length $2r$), it is possible to show that the rate of convergence for the approximation error is $O(1/r^q)$ for $r = 1, 2, \ldots, \infty$. Unfortunately, even with q growing linearly with n, we still need to use an exponential number of basis functions to achieve any specified accuracy (because the number of basis functions in the window is $(2r+1)^n$).

We would like to be able to choose an appropriate polynomial sized subset of the exponential number of basis functions for any specified accuracy. Theorem 7.3 shows this can be done with a number of basis function k of size $O\left(\frac{n^3\ln^2(1/\epsilon)}{q^2\epsilon^{4+4n/q}}\right)$ for fixed n and q. This shows that if q grows linearly with the dimension of the input space, the number of fixed basis functions required for any accuracy grows only polynomially (instead of exponentially) with the input dimension. However, the bound given in Theorem 7.3 is worse than the bound on the number of basis functions required for learning such functions as shown in Section 7.2. This is because instead of approximating one function (the function that is being learned), we require the basis functions to be good for all functions in the class Γ_q^s.

We will only consider approximation using sinusoidal basis functions. Similar results can be obtained for linear threshold and sigmoid basis functions. We do not give an explicit method for constructing the set of basis functions. However, the method in the proof can be used for constructing a probabilistic algorithm for finding the set of basis functions.

The main result of this section is stated in the following theorem.

Theorem 7.3 *There is a fixed set of sinusoidal basis functions (for fixed n and q) of size*

$$O\left(\frac{n^3M^4C^{4n/q}}{q^2\epsilon^{4+\frac{4n}{q}}}\left(\ln^2\frac{CM}{\epsilon}\right)\right)$$

which can be used to approximate any function in Γ_q^s to accuracy ϵ.

The proof is given in Appendix B. The idea behind the proof is similar to that used for the proof of the results in Section 7.2. We show that if the Fourier series is appropriately truncated and the basis functions of the truncated series are sampled uniformly, an appropriate set of basis functions is likely to be found.

7.5 Discussion on Approximation with Fixed Basis Functions

Work on L_∞ approximation using methods similar to those used here was first done by Barron (1992) who showed that a sup-norm approximation of order $O(1/k)$, where k is the number of hidden units, exists for single hidden layer neural networks with linear threshold hidden units. Barron (1992) used basis functions which are adapted for the particular function that is being approximated. Related work on sup-norm approximation has also been done by (Gurvits & Koiran 1995, Darken, Donahue, Gurvits & Sontag 1993, Yukich, Stinchcombe & White 1995).

For the class Γ_1^s, Barron (1993) has also shown that for L_2 approximation (which is implied by sup-norm approximation), the number of fixed basis functions required is $\Omega(1/\epsilon^n)$. With adaptable basis functions, the number of basis functions required is $O(1/\epsilon^2)$. This shows that for approximating a single function, having adaptable basis functions may give considerable advantage over having fixed basis functions. It also shows that the existence of a polynomial-sized set of fixed basis functions requires stronger conditions (such as those imposed in this chapter) than those satisfied by functions in Γ_1^s.

Paul, thou art beside thyself;

much learning doth make thee mad.

— Acts, xxvi, 24.

Chapter 8

Discussion and Conclusions

In this thesis, we have studied agnostic learning with the squared loss function. We showed that if the closure of a function class is not convex, the sample complexity for agnostic learning can be worse than the sample complexity for learning with noise if we are restricted to hypotheses from the same class. Furthermore, for some function classes, the order of the sample complexity for learning the convex hull of the function class (a single hidden layer neural networks) is comparable with the order of the sample complexity for learning the function class itself. Since the convex hull usually gives better approximation than the original function class, it may be advantageous to use the convex hull for agnostic learning.

We have also found that agnostic learning of the class of single hidden layer neural networks can be done in a computationally efficient manner if agnostic learning of the basis function class can be done in a computationally efficient manner. Unfortunately, we have also shown that agnostic learning of many classes of single hidden layer neural networks is likely to be computationally difficult. In view of this, we studied some natural but fairly restricted classes of functions which can be approximated by single hidden layer neural networks. We found that properties of function classes which make learning computationally tractable includes being smooth, having small L_1 norm and having low order.

The aim of this thesis (and most of the research in computational learning theory) is to use rigorous mathematical analysis to gain insights into various aspects of learning. To be tractable, the formal model of learning which we use must be simple. We have examined one fairly extreme model of learning and a reasonably flexible class of functions. Although the results obtained are not refined enough for practical use, we believe they give useful insights on the amount of information and computation needed for learning and on how the difficulty of learning scales as the complexity of the learning problems grows. Much remains to be done. We list a few interesting questions related to the work done in this thesis.

8.1 Sample Complexity

Agnostic Learning with other Loss Functions. We have studied the sample complexity of agnostic learning using the squared loss function. For a function class with finite pseudo dimension, the sample complexity is $O(\ln(1/\epsilon)/\epsilon)$ if the closure of the function class is convex and $\Omega(1/\epsilon^2)$ if the closure of the function class is not convex and we are restricted to hypotheses from the same class. For general loss functions, by considering $\{0, 1\}$-valued functions with $\{0, 1\}$-valued targets, we see that a lower bound of $\Omega(1/\epsilon^2)$ holds for agnostic learning if we are restricted to hypotheses from the same class. Finding conditions on the function classes which will allow a better sample complexity than $\Omega(1/\epsilon^2)$ for other loss functions, such as $(y, y') \mapsto |y - y'|^p$, $p \neq 2$, would give interesting generalizations of our results.

8.2 Computational Complexity

L_1 Norm. We think that identifying the properties of functions which makes learning computationally tractable is important because knowing these properties makes intelligent preprocessing of data and partitioning of the learning task into smaller tasks possible. In Chapter 7, we have shown that it is useful for the L_1 norm of the functions to be small when considering efficient agnostic learning. While it is well known that having low order can make function classes easier to learn, we have not seen any work on how the difficulty of learning scales with the L_1 norm, a parameter which we think is both natural and interesting. It would also be worthwhile to identify other natural parameters which affect the difficulty of learning.

Input Distribution. In studying agnostic learning, we have neglected the effects of the probability distribution on learnability in favour of properties of the function class. To gain insights into the effect of input distributions on the learnability of function classes, it would be worthwhile to study the learnability of function classes under input distributions with well understood properties. One well understood input distribution is the uniform distribution.

For efficient agnostic learning, we have shown that it is useful for the L_1 norm of the functions to be small. It is interesting to note that if the domain is large enough, functions with small L_1 norm are essentially negligible under the uniform distribution. This suggests that less restricted function classes may be learnable under the uniform distribution. Another interesting input distribution with well understood properties is a mixture of Gaussians where the inputs are clustered around few centres. Finding out what other function classes are efficiently learnable under these distributions will help us understand better the effect of input distributions on the computational complexity of learning.

8.3 Smooth Functions

Other Transforms. Functions with small L_1 norms are well localised in space (otherwise they are small everywhere) but have a highly distributed network representation in the sense that the weights of the hidden units are close to uniformly distributed. This corresponds to using the Fourier transform for defining the output weights. (The L_1 norm bounds the magnitude of the Fourier transform preventing the hidden units from being concentrated in any one region. This also bounds the variance of the random variable in the Monte Carlo procedure that is used for learning the smooth function class in Chapter 7.) Other transforms such as wavelet transforms may have other properties (such as locality) which are not present in the Fourier transform. It may be possible to exploit these properties for efficient learning.

Comparison between Algorithms. The classes of efficiently learnable functions which we have studied have fairly intuitive properties (small L_1 norm and fast decay of the Fourier transform). We have studied the performance of a neural network algorithm which uses randomly chosen basis functions on these function classes. It would be interesting to compare the performance of different algorithms on these function classes. For example, we may want to consider the performance of the k-nearest neighbour algorithm for learning the class Γ_q. Given that we know the properties of these function classes, comparisons of the sample and computational complexity of different algorithms on these function classes may reveal some useful insights into the properties of the different algorithms.

Trade-off between Computation and Information. The class of functions with bounded q-th absolute moment of Fourier transform can be learned with sample complexity

$O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon}+\log\frac{1}{\delta}\right)\right)$ for all $q \geq 1$ if we are not concerned with computational complexity. The order of the sample complexity for our algorithm, which is efficient when q grows linearly with n, is $1/\epsilon^{3+\frac{2n}{q}}$ (ignoring log factors). We do not know whether a computationally efficient algorithm with sample complexity $O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon}+\log\frac{1}{\delta}\right)\right)$ exists or if the increase in the the sample complexity is essential in order to have an efficient algorithm. More generally, the trade-off between computation and sample size for learning is not well understood and is a worthwhile but presumably deep research problem.

Appendix A

Proofs of Results from Chapter 3

In Section A.1, we give the proofs of Theorem 3.6 and 3.7. In Section A.2, we give the proof of Lemma 3.9.

A.1 Proof of Theorems 3.6 and 3.7

We restate the two theorems into a single theorem and then give the proof.

Theorem A.1 *Let $\mathcal{F} = \bigcup_{k=1}^{\infty} \mathcal{F}_k$ be a class of functions mapping from $\mathcal{X}$ to $\mathcal{Y} \subseteq [-T, T]$ such that each $\mathcal{F}_k$ is permissible. Let P be an arbitrary probability distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}$. Let $C = \max\{T, 1\}$. Assume $\nu, \nu_c > 0, 0 < \alpha \leq 1/2$. Let $\bar{\mathcal{F}}$ be the closure of $\mathcal{F}$ in the space with inner product $\langle f, g \rangle = \int f(x)g(x)dP_{\mathcal{X}}(x)$. Let $f^*(x) = \mathbf{E}[Y|X = x]$ and $g_f(x, y) = (y - f(x))^2 - (y - f_a(x))^2$ where $f_a = \operatorname{argmin}_{f \in \bar{\mathcal{F}}} \int (f(x) - f^*(x))^2 dP_{\mathcal{X}}(x)$. Assume either $f^* = f^a \in \mathcal{F}$ or $\mathcal{F}$ is a closure-convex class of functions. Then for $m \geq 1$ and each $k = 1, \ldots, \infty$,*

$$P^m \left\{ \boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}_k, \frac{\mathbf{E}(g_f) - \hat{\mathbf{E}}_{\boldsymbol{z}}(g_f)}{\nu + \nu_c + \mathbf{E}(g_f)} \geq \alpha \right\}$$
$$\leq \sup_{\boldsymbol{x} \in \mathcal{X}^{2m}} 6N \left(\frac{\alpha \nu_c}{128C^3}, \mathcal{F}_{k|\boldsymbol{x}}, l_1 \right) \exp(-\alpha^2 \nu m/(875C^4)).$$

Theorem 3.6 follows by letting $\mathcal{F}_k = \mathcal{G}$ for $k = 1, \ldots, \infty$.

The proof is similar to that used by Haussler (1992) and Pollard (1995). Theorem A.1 is a uniform convergence result for the empirical average of i.i.d. random variables $g_f(X_i, Y_i) = (Y_i - f(X_i))^2 - (Y_i - f_a(X_i))^2$ indexed by $f \in \mathcal{F}$. Haussler's result applies to more general random variables but only when they are nonnegative while Pollard's result provides bounds in terms of the magnitudes of the random variables instead of the random variables themselves. We use a Bernstein-type inequality in place of Hoeffding's inequality used by Haussler and Pollard and require that the second moment of each random variable be bounded by a linear function of the expectation of the random variable. The convexity condition on $\mathcal{F}$ is used to satisfy this condition. The condition is also satisfied if the conditional expectation is a member of $\mathcal{F}$. First we introduce the following functions for notational convenience. For $r, s \in \mathbb{R}, \nu, \nu_c \in \mathbb{R}^+$, let the functions d_{ν,ν_c} and d^1_{ν,ν_c} be defined by

$$d_{\nu,\nu_c}(r,s) = \frac{|r-s|}{\nu+\nu_c+r+s} \quad d^1_{\nu,\nu_c}(r,s) = \frac{r-s}{\nu+\nu_c+r}.$$

The function $d_{\nu,\nu_c}(r,s)$ is a variant of the function d_ν introduced by Haussler (1992).

We will bound the probability of the event with the probability of the union of two events.

$$\begin{aligned} &P^m\{\boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, d^1_{\nu,\nu_c}(\mathbf{E}(g_f), \hat{\mathbf{E}}_{\boldsymbol{z}}(g_f)) \geq \alpha\} \qquad\qquad \text{(A.1)} \\ &\leq P^m\{\boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, d^1_{\nu,\nu_c}(\mathbf{E}(g_f), \hat{\mathbf{E}}_{\boldsymbol{z}}(g_f)) \geq \alpha \text{ and } d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(g_f^2), \mathbf{E}(g_f^2)) \leq \alpha\} + \\ &\qquad\qquad P^m\{\boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(g_f^2)), \mathbf{E}(g_f^2)) > \alpha\}. \end{aligned}$$

The two probabilities will be bounded separately. The random variables in the second term on the right hand side of inequality A.1 are nonnegative; hence a result similar to Haussler's can be used. With minor modification to the proof, Theorem 3 in (Haussler 1992) becomes

Theorem A.2 ((Haussler 1992)) *Let $\mathcal{F}$ be a permissible set of functions on $\mathcal{Z}$ with $0 \leq f(z) \leq M$ for all $f \in \mathcal{F}$ and $z \in \mathcal{Z}$. Assume $\nu, \nu_c > 0$ and $0 < \alpha < 1$. Suppose that $\boldsymbol{z}$ is generated by m independent random draws according to any probability measure P on $\mathcal{Z}$. Then*

$$\begin{aligned} &P^m\{\boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(f), \mathbf{E}(f)) > \alpha\} \\ &\qquad\qquad \leq \sup_{\boldsymbol{z} \in \mathcal{Z}^{2m}} 4N(\alpha\nu_c/2, \mathcal{F}_{|\boldsymbol{z}}, l_1) \exp(-\alpha^2\nu m/2M). \end{aligned}$$

We will now bound the first term on the right hand side of equation (A.1). First we will turn the problem of bounding the probability involving the difference between the empirical average and expectation into a problem of bounding a probability involving the difference between the empirical averages of two independently sampled sequences of the same length. This is done in Lemma A.6. Then we make use of the independence property of the random variables to bound the probability involving the difference between the two empirical averages by the probability involving the difference between the empirical averages of two fixed sequences, when each member of the first sequence is equally likely to be interchanged with the member of the other sequence in the same position. This last probability depends only on the sequences involved, and thus allows the use of l_1 covering numbers instead of the L_∞ covering numbers. This is done in Lemma A.9. The following three results will be useful for the proof.

We will use the following result derived by Haussler (1992) from Chebyshev's inequality.

Lemma A.3 ((Haussler 1992)) *Let $V_1, \ldots, V_m$ be independent identically distributed random variables with range $0 \leq V_i \leq M$ and $\mathbf{E}(V_i) = \mu, 1 \leq i \leq m$. Assume $\nu + \nu_c > 0$ and $0 < \alpha < 1$. Then*

$$\Pr\left(d_{\nu,\nu_c}\left(\frac{1}{m}\sum_{i=1}^{m} V_i, \mu\right) > \alpha\right) < \frac{M}{4\alpha^2(\nu+\nu_c)m}.$$

As in (Barron 1990), we use the following inequality developed by Craig (1933) in his proof of Bernstein's inequality.

Lemma A.4 ((Craig 1933)) *Let $V_i, \ldots, V_m$ be independent identically distributed random variables which satisfy $|V_i - \mathbf{E}V_i| \leq 3h$ for $i = 1, \ldots, m$. Then*

$$\Pr\left(\frac{1}{n}\sum_{i=1}^{m} V_i - \mathbf{E}\left[\frac{1}{m}\sum_{i=1}^{m} V_i\right] \geq \frac{\tau}{m\xi} + \frac{m\xi \mathbf{Var}\left(\frac{1}{m}\sum_{i=1}^{m} V_i\right)}{2(1-c)}\right) \leq \exp(-\tau),$$

where $\tau > 0$ and $0 < \xi h \leq c < 1$.

Lemma A.5 *Let $V_1, \ldots, V_m$ be independent identically distributed random variables with $|V_i| < K_1$, $\mathbf{E}V_i \geq 0$ and $\mathbf{E}(V_i^2) < K_2\mathbf{E}V_i$, $K_2 \geq 1$ for $i = 1, \ldots, m$. Then for $0 < \alpha < 1$,*

$$\Pr\left(d^1_{\nu,\nu_c}\left(\mathbf{E}\left[\frac{1}{m}\sum_{i=1}^{m} V_i\right], \frac{1}{m}\sum_{i=1}^{m} V_i\right) \geq \alpha\right) \leq \exp\left(-\frac{3\alpha^2(\nu+\nu_c)m}{2(K_1+K_2)}\right).$$

Proof. Let $S_V = \mathbf{E}\left[\frac{1}{m}\sum_{i=1}^{m} V_i\right]$ and $\hat{S}_V = \frac{1}{m}\sum_{i=1}^{m} V_i$. Use the random variables $-V_i$ in Lemma A.4 to interchange the position of the empirical average and the expectation. Note that $\mathbf{Var}(-V_i) = \mathbf{Var}(V_i)$ and $m\mathbf{Var}(\hat{S}_V) = \mathbf{Var}(V_i) \leq \mathbf{E}(V_i^2)$. We get

$$\Pr\left(S_V - \hat{S}_V \geq \frac{\tau}{m\xi} + \frac{K_2\xi S_V}{2(1-c)}\right) \leq \exp(-\tau). \tag{A.2}$$

Now $|V_i - \mathbf{E}V_i| \leq 2K_1$ so $h = \frac{2K_1}{3}$ satisfies the required condition in Lemma A.4. Let $c = \xi h = \frac{2K_1\xi}{3}$. Set $\xi = \frac{6\alpha}{3K_2+4K_1}$ to get $\frac{K_2\xi}{2(1-2K_1\xi/3)} = \alpha$. Next set $\tau = \frac{6\alpha^2(\nu+\nu_c)m}{3K_2+4K_1}$ which gives $\frac{\tau}{\xi m} = \alpha(\nu+\nu_c)$. Note that $\tau \geq \frac{3\alpha^2(\nu+\nu_c)m}{2(K_1+K_2)}$. Substituting into (A.2) gives the required inequality. □

Lemma A.6 *Let $\mathcal{F}$ be a permissible class of functions, with $|f(z)| < K_1$ for all $f \in \mathcal{F}$ and $z \in \mathcal{Z}$. Suppose the distribution P of z is such that* $\mathbf{E}(f) \geq 0$ *and* $\mathbf{E}(f^2) \leq K_2\mathbf{E}(f)$, $K_2 \geq 1$ *for all $f \in \mathcal{F}$. Assume $(\nu+\nu_c) > 0, 0 < \alpha < 1$. Then for* $m \geq \max\left\{\frac{4(K_1+K_2)}{\alpha^2(\nu+\nu_c)}, \frac{K_1^2}{\alpha^2(\nu+\nu_c)}\right\}$,

$$\begin{aligned}
&P^m\{\boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, d^1_{\nu,\nu_c}(\mathbf{E}(f), \hat{\mathbf{E}}_{\boldsymbol{z}}(f)) \geq \alpha \text{ and } d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2), \mathbf{E}(f^2)) \leq \alpha\} \\
&\leq 2P^{2m}\left\{\boldsymbol{z}\boldsymbol{z}' \in \mathcal{Z}^{2m} : \exists f \in \mathcal{F}, \frac{\hat{\mathbf{E}}_{\boldsymbol{z}'}(f) - \hat{\mathbf{E}}_{\boldsymbol{z}}(f)}{(\nu+\nu_c)+\mathbf{E}(f)} \geq \frac{\alpha}{2} \text{ and}\right. \\
&\qquad \left. d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2), \mathbf{E}(f^2)) \leq \alpha \text{ and } d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2), \mathbf{E}(f^2)) \leq \alpha\right\}.
\end{aligned}$$

Proof. Consider any f and a sample $\boldsymbol{z} \in \mathcal{Z}^m$ such that

$$\mathbf{E}(f) - \hat{\mathbf{E}}_{\boldsymbol{z}}(f) \geq \alpha(\nu+\nu_c) + \alpha\mathbf{E}(f) \tag{A.3}$$

and $d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2), \mathbf{E}(f^2)) \leq \alpha$. Draw another independent random sample $\boldsymbol{z}'$ of length m. From Lemma A.5, for $m \geq \frac{4(K_1+K_2)}{\alpha^2(\nu+\nu_c)} > \frac{8\ln 4(K_1+K_2)}{3\alpha^2(\nu+\nu_c)}$, the probability that $\mathbf{E}(f) - \hat{\mathbf{E}}_{\boldsymbol{z}'}(f) \geq \alpha(\nu+\nu_c)/2 + \alpha\mathbf{E}(f)/2$ is less than $1/4$. Since $|f(z)^2| \leq K_1^2$, from Lemma A.3 we find that for $m \geq \frac{K_1^2}{\alpha^2(\nu+\nu_c)}$, the probability that $d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2), \mathbf{E}(f^2)) > \alpha$ is less than $1/4$. So for $m \geq \max\left\{\frac{4(K_1+K_2)}{\alpha^2(\nu+\nu_c)}, \frac{K_1^2}{\alpha^2(\nu+\nu_c)}\right\}$, with probability at least $1/2$, both

$$\mathbf{E}(f) - \hat{\mathbf{E}}_{\boldsymbol{z}'}(f)) < \alpha(\nu+\nu_c)/2 + \alpha\mathbf{E}(f)/2 \tag{A.4}$$

and $d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2), \mathbf{E}(f^2)) \leq \alpha$. Subtracting (A.4) from (A.3), and using the independence of the samples, we have

$$\begin{aligned}
&P^{2m}\left\{\boldsymbol{z}\boldsymbol{z}' \in \mathcal{Z}^{2m} : \exists f \in \mathcal{F}, \frac{\hat{\mathbf{E}}_{\boldsymbol{z}'}(f) - \hat{\mathbf{E}}_{\boldsymbol{z}}(f)}{(\nu+\nu_c)+\mathbf{E}(f)} \geq \frac{\alpha}{2} \text{ and}\right. \\
&\qquad \left. d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2), \mathbf{E}(f^2)) \leq \alpha \text{ and } d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2), \mathbf{E}(f^2)) \leq \alpha\right\} \\
&\geq P^{2m}\left\{\boldsymbol{z}\boldsymbol{z}' \in \mathcal{Z}^{2m} : \exists f \in \mathcal{F}, d^1_{\nu,\nu_c}(\mathbf{E}(f) - \hat{\mathbf{E}}_{\boldsymbol{z}}(f)) \geq \alpha \text{ and } d^1_{\nu,\nu_c}(\mathbf{E}(f) - \hat{\mathbf{E}}_{\boldsymbol{z}'}(f)) \geq \alpha/2\right. \\
&\qquad \left. \text{and } d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2), \mathbf{E}(f^2)) \leq \alpha \text{ and } d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2), \mathbf{E}(f^2)) \leq \alpha\right\}
\end{aligned}$$

$$\geq \frac{1}{2} P^m \{ \boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, d^1_{\nu,\nu_c}(\mathbf{E}(f), \hat{\mathbf{E}}_{\boldsymbol{z}}(f)) \geq \alpha \text{ and } d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2), \mathbf{E}(f^2)) \leq \alpha \}.$$

□

The following two Lemmas will be useful for proving Lemma A.9.

Lemma A.7 *Suppose $m \geq 1$ is arbitrary and let U be the uniform distribution over $\{-1, 1\}$. For a fixed $\boldsymbol{z}\boldsymbol{z}' \in \mathcal{Z}^{2m}$, any function f, and random variables U_i, $i = 1, \ldots, m$ drawn independently from U,*

$$m\mathbf{Var}\left(\frac{1}{m} \sum_{i=1}^{m} U_i(f(z_i) - f(z_i')) \right) \leq 3\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2) + 3\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2).$$

Proof.

$$\begin{aligned}
m\mathbf{Var}\left(\frac{1}{m} \sum_{i=1}^{m} U_i(f(z_i) - f(z_i')) \right) &= \frac{1}{m} \sum_{i=1}^{m} f(z_i)^2 + \frac{1}{m} \sum_{i=1}^{m} f(z_i')^2 - \frac{2}{m} \sum_{i=1}^{m} f(z_i) f(z_i') \\
&\leq \frac{1}{m} \sum_{i=1}^{m} f(z_i)^2 + \frac{1}{m} \sum_{i=1}^{m} f(z_i')^2 + 2\sqrt{\frac{1}{m} \sum_{i=1}^{m} f(z_i)^2} \sqrt{\frac{1}{m} \sum_{i=1}^{m} f(z_i')^2} \\
&\leq \frac{1}{m} \sum_{i=1}^{m} f(z_i)^2 + \frac{1}{m} \sum_{i=1}^{m} f(z_i')^2 + 2 \max \left\{ \frac{1}{m} \sum_{i=1}^{m} f(z_i)^2, \frac{1}{m} \sum_{i=1}^{m} f(z_i')^2 \right\} \\
&\leq 3\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2) + 3\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2).
\end{aligned}$$

□

Lemma A.8 *Suppose $m \geq 1$ is arbitrary and let U be the uniform distribution over $\{-1, 1\}$. Then, for a fixed $\boldsymbol{z}\boldsymbol{z}' \in \mathcal{Z}^{2m}$ and any function f,*

$$\begin{aligned}
U^m &\left\{ \boldsymbol{u} \in \{-1, 1\}^m : \exists f \in \mathcal{F}, \frac{1}{m} \sum_{i=1}^{m} u_i(f(z_i) - f(z_i')) \geq \alpha + \epsilon \right\} \\
&\qquad \leq U^m \left\{ \boldsymbol{u} \in \{-1, 1\}^m : \exists f \in \mathcal{G}_\epsilon, \frac{1}{m} \sum_{i=1}^{m} u_i(f(z_i) - f(z_i')) \geq \alpha \right\}
\end{aligned}$$

where $\mathcal{G}_\epsilon$ is an l_1 ϵ-cover of $\mathcal{F}_{|\boldsymbol{z}\boldsymbol{z}'}$.

Proof. Suppose $\frac{1}{m} \sum_{i=1}^{m} u_i(f(z_i) - f(z_i')) \geq \alpha + \epsilon$. There exists a $g \in \mathcal{G}_\epsilon$ such that

$$\frac{1}{m} \sum_{i=1}^{m} |g(z_i) - f(z_i)| + \frac{1}{m} \sum_{i=1}^{m} |g(z_i') - f(z_i')| < \epsilon.$$

Hence

$$\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m} u_i(g(z_i)-g(z_i'))\\
=\ & \frac{1}{m}\sum_{i=1}^{m} u_i(g(z_i)-f(z_i)+f(z_i)-f(z_i')+f(z_i')-g(z_i'))\\
=\ & \frac{1}{m}\left[\sum_{i=1}^{m} u_i(f(z_i)-f(z_i'))+\sum_{i=1}^{m} u_i(g(z_i)-f(z_i)-g(z_i')+f(z_i'))\right]\\
\geq\ & \frac{1}{m}\left[\sum_{i=1}^{m} u_i(f(z_i)-f(z_i'))-\sum_{i=1}^{m} |g(z_i)-f(z_i)|+|g(z_i')-f(z_i')|\right]\\
\geq\ & \alpha+\epsilon-\epsilon=\alpha.
\end{aligned}$$

□

In the following lemma, we bound the probability involving the difference between the two empirical averages that arose in Lemma A.6, by a probability involving the difference between the empirical averages of two fixed sequences, when each component of the first sequence is randomly interchanged with the corresponding component of the other sequence. This probability depends only on the sequences involved, and thus is bounded by a function of the l_1 covering number.

Lemma A.9 *Let $\mathcal{F}$ be a permissible class of functions with $|f(z)| \leq K_1$ for all $f \in \mathcal{F}$ and $z \in \mathcal{Z}$. Suppose $K_2 \geq 1$ and the distribution P of z is such that $\mathbf{E}f(z) \geq 0$ and $\mathbf{E}(f^2) \leq K_2\mathbf{E}(f)$ for all $f \in \mathcal{F}$. Assume $\nu, \nu_c > 0, 0 < \alpha \leq 1/2$. Then for $m \geq \max\left\{\frac{4(K_1+K_2)}{\alpha^2(\nu+\nu_c)}, \frac{K_1^2}{\alpha^2(\nu+\nu_c)}\right\}$,*

$$P^{2m}\left\{\boldsymbol{z}\boldsymbol{z}' \in \mathcal{Z}^{2m} : \exists f \in \mathcal{F}, \frac{\hat{\mathbf{E}}_{\boldsymbol{z}'}(f)-\hat{\mathbf{E}}_{\boldsymbol{z}}(f)}{(\nu+\nu_c)+\mathbf{E}(f)} \geq \frac{\alpha}{2} \text{ and } d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2),\mathbf{E}(f^2)) \leq \alpha \text{ and } d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2),\mathbf{E}(f^2)) \leq \alpha\right\} \leq \sup_{\boldsymbol{z}\in\mathcal{Z}^{2m}} N\left(\frac{\alpha\nu_c}{4}, \mathcal{F}_{|\boldsymbol{z}}, l_1\right)\exp\left(-\frac{3\alpha^2\nu m}{4K_1+162K_2}\right). \quad \text{(A.5)}$$

Proof. We are interested in $\boldsymbol{z}$ and $\boldsymbol{z}'$ such that there exists an f with

$$\hat{\mathbf{E}}_{\boldsymbol{z}'}(f)-\hat{\mathbf{E}}_{\boldsymbol{z}}(f) \geq \frac{\alpha(\nu+\nu_c)}{2}+\frac{\alpha\mathbf{E}(f)}{2} \quad \text{(A.6)}$$

and

$$|\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2)-\mathbf{E}(f^2)| \leq \alpha(\nu+\nu_c)+\alpha\mathbf{E}(f^2)+\alpha\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2) \quad \text{(A.7)}$$

and

$$|\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2)-\mathbf{E}(f^2)| \leq \alpha(\nu+\nu_c)+\alpha\mathbf{E}(f^2)+\alpha\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2). \quad \text{(A.8)}$$

When that happens, $(1+\alpha)\mathbf{E}(f^2) \geq (1-\alpha)\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2) - \alpha(\nu+\nu_c)$ and similarly for $\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2)$, so we have

$$\begin{aligned}
\hat{\mathbf{E}}_{\boldsymbol{z}'}(f) \quad - \quad & \hat{\mathbf{E}}_{\boldsymbol{z}}(f) \geq \frac{\alpha(\nu+\nu_c)}{2} + \frac{\alpha\mathbf{E}(f)}{2} \\
\geq \quad & \frac{\alpha(\nu+\nu_c)}{2} + \frac{\alpha\mathbf{E}(f^2)}{2K_2} \\
\geq \quad & \frac{\alpha(\nu+\nu_c)}{2} + \frac{\alpha}{2}\left[\frac{(1-\alpha)\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2) - \alpha(\nu+\nu_c)}{2K_2(1+\alpha)} + \frac{(1-\alpha)\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2) - \alpha(\nu+\nu_c)}{2K_2(1+\alpha)}\right] \\
\geq \quad & \frac{\alpha(\nu+\nu_c)}{2} - \frac{\alpha^2(\nu+\nu_c)}{2K_2(1+\alpha)} + \frac{\alpha(1-\alpha)(3\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2) + 3\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2))}{12(1+\alpha)K_2}. \qquad \text{(A.9)}
\end{aligned}$$

The fact that the random variables are independent means that the probability in (A.5) remains unchanged when each component of $\boldsymbol{z}$ is randomly interchanged with the corresponding component of $\boldsymbol{z}'$. Let U be the uniform distribution over $\{-1, 1\}$. We have

$$\begin{aligned}
& P^{2m}\Bigg\{\boldsymbol{z}\boldsymbol{z}' \in \mathcal{Z}^{2m} : \exists f \in \mathcal{F}, \frac{\hat{\mathbf{E}}_{\boldsymbol{z}'}(f) - \hat{\mathbf{E}}_{\boldsymbol{z}}(f)}{(\nu+\nu_c) + \mathbf{E}(f)} \geq \frac{\alpha}{2} \text{ and } d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2), \mathbf{E}(f^2)) \leq \alpha \text{ and} \\
& \qquad d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2), \mathbf{E}(f^2)) \leq \alpha\Bigg\} \\
\leq \quad & P^{2m}\Bigg\{\boldsymbol{z}\boldsymbol{z}' \in \mathcal{Z}^{2m} : \exists f \in \mathcal{F}, \hat{\mathbf{E}}_{\boldsymbol{z}'}(f) - \hat{\mathbf{E}}_{\boldsymbol{z}}(f) \geq \frac{\alpha(\nu+\nu_c)}{2} - \frac{\alpha^2(\nu+\nu_c)}{2K_2(1+\alpha)} + \\
& \qquad \frac{\alpha(1-\alpha)(3\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2) + 3\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2))}{12(1+\alpha)K_2}\Bigg\} \quad \text{(using (A.9))} \\
= \quad & P^{2m} \times U^m \Bigg\{\boldsymbol{z}\boldsymbol{z}' \in \mathcal{Z}^{2m}, \boldsymbol{u} \in \{-1,1\}^m : \exists f \in \mathcal{F}, \frac{1}{m}\sum_{i=1}^m u_i(f(z_i) - f(z_i')) \geq \\
& \qquad \frac{\alpha(\nu+\nu_c)}{2} - \frac{\alpha^2(\nu+\nu_c)}{2K_2(1+\alpha)} + \frac{\alpha(1-\alpha)(3\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2) + 3\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2))}{12(1+\alpha)K_2}\Bigg\} \\
\leq \quad & \sup_{\boldsymbol{z}\boldsymbol{z}' \in \mathcal{Z}^{2m}} U^m \Bigg\{\boldsymbol{u} \in \{-1,1\}^m : \exists f \in \mathcal{F}, \frac{1}{m}\sum_{i=1}^m u_i(f(z_i) - f(z_i')) \geq \\
& \qquad \frac{\alpha(\nu+\nu_c)}{2} - \frac{\alpha^2(\nu+\nu_c)}{2K_2(1+\alpha)} + \frac{\alpha(1-\alpha)(3\hat{\mathbf{E}}_{\boldsymbol{z}}(f^2) + 3\hat{\mathbf{E}}_{\boldsymbol{z}'}(f^2))}{12(1+\alpha)K_2}\Bigg\} \\
\leq \quad & \sup_{\boldsymbol{z}\boldsymbol{z}' \in \mathcal{Z}^{2m}} U^m \Bigg\{\boldsymbol{u} \in \{-1,1\}^m : \exists f \in \mathcal{G}_c, \frac{1}{m}\sum_{i=1}^m u_i(f(z_i) - f(z_i')) \geq \\
& \qquad \frac{\alpha\nu}{2} - \frac{\alpha^2\nu}{2K_2(1+\alpha)} + \frac{\alpha(1-\alpha)(m\mathbf{Var}(\frac{1}{m}\sum_{i=1}^m u_i(f(z_i) - f(z_i'))))}{12(1+\alpha)K_2}\Bigg\} \qquad \text{(A.10)}
\end{aligned}$$

using Lemmas A.8 and A.7, where $\mathcal{G}_c$ is an $\alpha\nu_c/4$-cover of $\mathcal{F}_{|\boldsymbol{z}}$. (Note that $\frac{\alpha\nu_c}{4} \leq (\frac{\alpha\nu_c}{2} - \frac{\alpha^2\nu_c}{2K_2(1+\alpha)})$ for $\alpha \leq 1$).

Now $|u_i(f(z_i) - f(z_i'))| \leq 2K_1$. Set $h = \frac{2K_1}{3}$ to satisfy the condition in Lemma A.4. Let $c = \xi h = \frac{2K_1\xi}{3}$ in Lemma A.4. We want $\frac{\alpha(1-\alpha)}{12(1+\alpha)K_2} = \frac{\xi}{2(1-c)} = \frac{\xi}{2(1-2K_1\xi/3)}$. So $\xi = \frac{3\alpha(1-\alpha)}{18(1+\alpha)K_2+2K_1\alpha(1-\alpha)}$. Now set $\frac{\alpha\nu}{2} - \frac{\alpha^2\nu}{2K_2(1+\alpha)} = \frac{\tau}{\xi m}$. This gives $\tau/m = \frac{3K_2(1-\alpha^2)\alpha^2\nu - 3\alpha^3(1-\alpha)\nu}{36K_2^2(1+\alpha)^2+4\alpha(1-\alpha^2)K_1K_2} > \frac{3\alpha^2\nu}{162K_2+4K_1}$ for $0 < \alpha \leq 1/2$.

With these settings, the expression in (A.10) is less than

$$\sup_{\boldsymbol{z}\in\mathcal{Z}^{2m}} N\left(\frac{\alpha\nu_c}{4}, \mathcal{F}_{|\boldsymbol{z}}, l_1\right) \exp\left(-\frac{3\alpha^2\nu m}{4K_1 + 162K_2}\right).$$

□

The following lemma is useful for bounding the second term on the right hand side of Equation (A.1), using Theorem A.2.

Lemma A.10 *Let $\mathcal{F}$ be a class of functions with $|f(z)| \leq K_1$ for all $f \in \mathcal{F}$ and $z \in \mathcal{Z}$. Let $\mathcal{F}^2 = \{f^2 : f \in \mathcal{F}\}$ and $\boldsymbol{z} \in \mathcal{Z}^m$. Then for all $\epsilon > 0$,*

$$N(\epsilon, \mathcal{F}^2_{|\boldsymbol{z}}, l_1) \leq N\left(\frac{\epsilon}{2K_1}, \mathcal{F}_{|\boldsymbol{z}}, l_1\right).$$

Proof. For any $f, g \in \mathcal{F}$ we have

$$\hat{\mathbf{E}}_{\boldsymbol{z}}|f^2 - g^2| \leq \hat{\mathbf{E}}_{\boldsymbol{z}}|f + g||f - g| \leq 2K_1\hat{\mathbf{E}}_{\boldsymbol{z}}|f - g|.$$

Hence if $T = \{f_1, \ldots, f_N\}$ is an $\epsilon/2K_1$-cover for $\mathcal{F}_{|\boldsymbol{z}}$, $T^2 = \{f_1^2, \ldots, f_N^2\}$ is an ϵ-cover for $\mathcal{F}^2_{|\boldsymbol{z}}$.
□

We are now ready to state a uniform convergence result with the condition that the second moment of the random variable can be bounded by a linear function of the expectation.

Theorem A.11 *Let $\mathcal{F}$ be a permissible class of functions with $|f(z)| \leq K_1$ for all $f \in \mathcal{F}$ and $z \in \mathcal{Z}$. Let $K_2 \geq 1$ and P be a probability distribution on $\mathcal{Z}$ such that $\mathbf{E}f(z) \geq 0$ and $\mathbf{E}(f^2) \leq K_2\mathbf{E}(f)$ for all $f \in \mathcal{F}$. Assume $\nu, \nu_c > 0$ and $0 < \alpha \leq 1/2$. Then for $m \geq \max\left\{\frac{4(K_1+K_2)}{\alpha^2(\nu+\nu_c)}, \frac{K_1^2}{\alpha^2(\nu+\nu_c)}\right\}$,*

$$P^m\{\boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, d^1_{\nu,\nu_c}(\mathbf{E}(f), \hat{\mathbf{E}}_{\boldsymbol{z}}(f)) \geq \alpha\} \leq \sup_{\boldsymbol{z}\in\mathcal{Z}^{2m}} 2N\left(\frac{\alpha\nu_c}{4}, \mathcal{F}_{|\boldsymbol{z}}, l_1\right) \times$$
$$\exp\left(-\frac{3\alpha^2\nu m}{4K_1 + 162K_2}\right) + \sup_{\boldsymbol{z}\in\mathcal{Z}^{2m}} 4N\left(\frac{\alpha\nu_c}{4K_1}, \mathcal{F}_{|\boldsymbol{z}}, l_1\right) \exp(-\alpha^2\nu m/2K_1^2).$$

Proof. From equation (A.1),

$$
\begin{aligned}
&P^m\{\boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, d^1_{\nu,\nu_c}(\mathbf{E}(g_f), \hat{\mathbf{E}}_{\boldsymbol{z}}(g_f)) \geq \alpha\} \\
&\leq P^m\{\boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, d^1_{\nu,\nu_c}(\mathbf{E}(g_f), \hat{\mathbf{E}}_{\boldsymbol{z}}(g_f)) \geq \alpha \text{ and } d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(g_f^2), \mathbf{E}(g_f^2)) \leq \alpha\} + \\
&\qquad P^m\{\boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(g_f^2)), \mathbf{E}(g_f^2)) > \alpha\}.
\end{aligned}
$$

From Lemma A.6 and Lemma A.9, the first term on the right hand side is bounded by $\sup_{\boldsymbol{z} \in \mathcal{Z}^{2m}} 2N\left(\frac{\alpha\nu_c}{4}, \mathcal{F}_{|\boldsymbol{z}}, l_1\right) \exp\left(-\frac{3\alpha^2\nu m}{4K_1+162K_2}\right)$. From Theorem A.2 and Lemma A.10, the second term on the right hand side is bounded by $\sup_{\boldsymbol{z} \in \mathcal{Z}^{2m}} 4N\left(\frac{\alpha\nu_c}{4K_1}, \mathcal{F}_{|\boldsymbol{z}}, l_1\right) \exp(-\alpha^2\nu m/2K_1^2)$.
□

We now show that if either $f^* = f_a \in \mathcal{F}$ or $\mathcal{F}$ is closure-convex, then $\mathbf{E}(g_f^2) \leq K_2\mathbf{E}(g_f)$ for some constant K_2.

Lemma A.12 *Let $\mathcal{F}$ be a class of functions with $|f(x)| \leq T$ for every $f \in \mathcal{F}$ and $x \in \mathcal{X}$. Let $|y| \leq T$ for every $y \in \mathcal{Y}$. Let X and Y be randomly generated according to some joint probability distribution P and suppose f_a in the closure of $\mathcal{F}$ is such that $\int(f_a(x) - f^*(x))^2 dP_{\mathcal{X}}(x) = \inf_{f\in\mathcal{F}} \int(f(x) - f^*(x))^2 dP_{\mathcal{X}}(x)$ where $f^*(x) = \mathbf{E}[Y|X = x]$. Assume $\mathcal{F}$ is closure-convex or $f^* = f_a \in \mathcal{F}$. Then for every $f \in \mathcal{F}$*

$$
\begin{aligned}
\mathbf{E}[((y - f(x))^2 - (y - f_a(x))^2)^2] &\leq 16T^2\mathbf{E}(f(x) - f_a(x))^2 \\
&\leq 16T^2\mathbf{E}[(y - f(x))^2 - (y - f_a(x))^2]. \quad \text{(A.11)}
\end{aligned}
$$

Proof. For the first part of inequality (A.11),

$$
\begin{aligned}
\mathbf{E}[((y - f(x))^2 - (y - f_a(x))^2)^2] &= \mathbf{E}[((2y_i - f(x_i) - f_a(x_i))(f_a(x_i) - f(x_i)))^2] \\
&\leq 16T^2\mathbf{E}[(f(x) - f_a(x))^2].
\end{aligned}
$$

For the second part of inequality (A.11), we have

$$
\begin{aligned}
&\mathbf{E}[(y - f(x))^2 - (y - f_a(x))^2] \\
&= \mathbf{E}[(y - f_a(x))^2 + (f_a(x) - f(x))^2 + 2(y - f_a(x))(f_a(x) - f(x)) - (y - f_a(x))^2] \\
&= \mathbf{E}[f_a(x) - f(x))^2 + 2(y - f^*(x) + f^*(x) - f_a(x))(f_a(x) - f(x))] \\
&= \mathbf{E}[f_a(x) - f(x))^2] + 2\mathbf{E}[(f^*(x) - f_a(x))(f_a(x) - f(x))].
\end{aligned}
$$

We need only to show that $\mathbf{E}[(f^*(x) - f_a(x))(f_a(x) - f(x))] \geq 0$. This is automatically true if $f^* = f_a$. For the case where $\mathcal{F}$ is closure-convex, let $\bar{\mathcal{F}}$ be the closure of $\mathcal{F}$. Then $\bar{\mathcal{F}}$ is convex and $f_a \in \bar{\mathcal{F}}$. From convexity, $f \in \bar{\mathcal{F}}$ implies $\alpha f + (1-\alpha) f_a \in \bar{\mathcal{F}}$ for $\alpha \in [0, 1]$. Since f_a is the best approximation in $\bar{\mathcal{F}}$,

$$
\begin{aligned}
&\mathbf{E}[(f^* - f_a(x))^2] \\
&\quad \leq \mathbf{E}[(f^*(x) - \alpha f(x) - (1-\alpha) f_a(x))^2] \\
&\quad = \mathbf{E}[(f^*(x) - f_a(x) + \alpha(f_a(x) - f(x)))^2] \\
&\quad = \mathbf{E}[(f^*(x) - f_a(x))^2 + \alpha^2 (f_a(x) - f(x))^2 + 2\alpha(f^*(x) - f_a(x))(f_a(x) - f(x)).]
\end{aligned}
$$

This gives $\mathbf{E}(f^*(x) - f_a(x))(f_a(x) - f(x)) \geq -\alpha \mathbf{E}(f(x) - f_a(x))^2/2$ for all $\alpha \in [0, 1]$, which implies $\mathbf{E}(f^*(x) - f_a(x))(f_a(x) - f(x)) \geq 0$. □

Lemma A.13 *Let $\mathcal{F}$ be a class of functions with $|f(x)| \leq T$ for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$. Suppose $|y| \leq T$ for all $y \in \mathcal{Y}$. Let $\mathcal{G} = \{g_f : g_f(x, y) = (y - f(x))^2 - (y - f_a(x))^2, f \in \mathcal{F}\}$, where f_a is an arbitrary function. Let $\boldsymbol{z} \in \mathcal{Z}^m = (\mathcal{X} \times \mathcal{Y})^m$. Then*

$$N(\epsilon, \mathcal{G}_{|\boldsymbol{z}}, l_1) \leq N(\epsilon/4T, \mathcal{F}_{|\boldsymbol{z}}, l_1).$$

Proof. For any $f, g \in \mathcal{F}$ we have

$$
\begin{aligned}
&\frac{1}{m} \sum_{i=1}^{m} [(y_i - f(x_i))^2 - (y_i - f_a(x_i))^2 - (y_i - g(x_i))^2 + (y_i - f_a(x_i))^2] \\
&= \frac{1}{m} \sum_{i=1}^{m} [(y_i - f(x_i))^2 - (y_i - g(x_i))^2] \\
&= \frac{1}{m} \sum_{i=1}^{m} [(2y_i - f(x_i) - g(x_i))(g(x_i) - f(x_i))] \\
&\leq \frac{4B}{m} \sum_{i=1}^{m} |g(x_i) - f(x_i)|.
\end{aligned}
$$

Hence if $\mathcal{T} = \{f_1, \ldots, f_N\}$ is an $\epsilon/4T$-cover for $\mathcal{F}_{|\boldsymbol{z}}$, $\mathcal{T}_{\mathcal{G}} = \{g_{f_1}, \ldots, g_{f_N}\}$ is an ϵ-cover for $\mathcal{G}_{|\boldsymbol{z}}$.
□

We now have everything we need to prove Theorem A.1.

Proof.(Theorem A.1) In Theorem A.11, K_1 can be set to $8C^2$. Using the convexity of $\mathcal{F} = \bigcup_{k=1}^{\infty} \mathcal{F}_k$ and Lemma A.12, K_2 can be set to to $16C^2$. Using Lemma A.13, the right hand side of

Theorem A.11 can be bounded by

$$\begin{aligned}
&\sup_{\boldsymbol{z}\in\mathcal{Z}^{2m}} 2N\left(\frac{\alpha\nu_c}{16C}, \mathcal{F}_{k|\boldsymbol{z}}, l_1\right)\exp\left(-\frac{3\alpha^2\nu m}{2624C^2}\right)+\\
&\qquad \sup_{\boldsymbol{z}\in\mathcal{Z}^{2m}} 4N\left(\frac{\alpha\nu_c}{128C^3}, \mathcal{F}_{k|\boldsymbol{z}}, l_1\right)\exp(-\alpha^2\nu m/128C^4)\\
&\qquad \le \sup_{\boldsymbol{z}\in\mathcal{Z}^{2m}} 6N\left(\frac{\alpha\nu_c}{128C^3}, \mathcal{F}_{k|\boldsymbol{z}}, l_1\right)\exp(-\alpha^2\nu m/875C^4).
\end{aligned}$$

□

A.2 Proof of Lemma 3.9

To prove Lemma 3.9, we will bound the packing number of the function class. (See Chapter 2 for the definition of packing number.) For a set T and (pseudo) metric ρ, it is easily seen that $N(\epsilon, S, \rho) \le M(\epsilon, S, \rho)$.

We will bound $M(\epsilon, \mathcal{F}, L_1(P))$ for all P in terms of the fat-shattering function of $\mathcal{F}$. This provides a bound on $N(\epsilon, \mathcal{F}, L_1(P))$ for all P which bounds $N(\epsilon, \mathcal{F}_{|\boldsymbol{x}}, l_1)$ for any finite sequence of points $\boldsymbol{x}$ (via the isometry between the two metric spaces $(\mathcal{F}_{|\boldsymbol{x}}, d_{l_1})$ and $(\mathcal{F}, d_{L^1(P_{|\boldsymbol{x}})})$, where $P_{|\boldsymbol{x}}$ is the empirical distribution on $\boldsymbol{x}$). We use techniques due to Haussler (1992) which go back to Pollard (1984) and Dudley (1978). The following result follows trivially from a generalization of Sauer's lemma by Alon et al. (1993).

Theorem A.14 ((Alon et al. 1993)) *Let $\mathcal{F}$ be a class of $[0,1]$-valued functions defined on $\mathcal{X}$, $0 < \epsilon < 1$, and $m \ge \log y + 1$ (with $\log$ denoting base 2 logarithm), where*

$$y = \sum_{i=1}^{\mathrm{fat}_{\mathcal{F}}(\epsilon/4)} \binom{m}{i} b^i$$

and $b = \lceil 2/\epsilon \rceil + 1$. Then for all $\boldsymbol{x} \in \mathcal{X}^m$,

$$M(\epsilon, \mathcal{F}_{|\boldsymbol{x}}, l_\infty) \le 2(mb^2)^{\log y}.$$

Corollary A.15 *Let $\mathcal{F}$ be defined as in Theorem A.14 and $0 < \epsilon \le 1$, $d = \mathrm{fat}_{\mathcal{F}}(\epsilon/4)$ and $m \ge 4d \log \frac{8d}{\epsilon}$. Then for all $\boldsymbol{x} \in \mathcal{X}^m$,*

$$M(\epsilon, \mathcal{F}_{|\boldsymbol{x}}, l_\infty) \le \exp\left(\frac{2}{\ln 2} d \ln^2 \frac{16m}{\epsilon^2}\right).$$

Proof. If $d = 0$ then $M(\epsilon, \mathcal{F}_{|\boldsymbol{x}}, l_\infty) = 1$. Assume $d \geq 1$ and let y be defined as in Theorem A.14. First we want to show that if $m \geq 4d \log \frac{8d}{\epsilon}$, then $m \geq \log y + 1$. We have $b < 4/\epsilon$ and

$$\begin{aligned} \log y + 1 &< \log \sum_{i=1}^{d} \binom{m}{i} \left(\frac{4}{\epsilon}\right)^i + 1 \\ &< \log \left(d \left(\frac{4m}{\epsilon}\right)^d \right) + 1 \\ &= d \log \left(\frac{4m}{\epsilon}\right) + \log d + 1 \\ &\leq 2d \log \left(\frac{4m}{\epsilon}\right). \end{aligned}$$

It is easy to see that $2d \log \left(\frac{4m}{\epsilon}\right)$ grows more slowly than m for $m \geq 4d \log \left(\frac{8d}{\epsilon}\right)$. Furthermore, $4d \log \frac{8d}{\epsilon} = 2d \log \left(\frac{8d}{\epsilon}\right)^2 \geq 2d \log \left(\frac{4m}{\epsilon}\right)$ when $m = 2d \log \left(\frac{8d}{\epsilon}\right)^2$. Hence, if $m \geq 4d \log \frac{8d}{\epsilon}$ then $m \geq \log y + 1$.

Finally,

$$\begin{aligned} \ln M(\epsilon, \mathcal{F}_{|\boldsymbol{x}}, l_\infty) &\leq \ln 2 + \log y \ln \frac{16m}{\epsilon^2} \\ &< \ln 2 + \left(d \log \frac{4m}{\epsilon} + \log d \right) \ln \frac{16m}{\epsilon^2} \\ &< \frac{2d}{\ln 2} \ln \frac{4m}{\epsilon} \ln \frac{16m}{\epsilon^2} \\ &< \frac{2d}{\ln 2} \ln^2 \left(\frac{16m}{\epsilon^2} \right). \end{aligned}$$

□

Lemma A.16 *Let $\mathcal{F}$ be a family of functions from a set $\mathcal{X}$ into $[0,1]$ and let P be a probability distribution on $\mathcal{X}$. Let $\boldsymbol{X} = (\boldsymbol{X}_1, \ldots, \boldsymbol{X}_m)$ be a random vector in $\mathcal{X}^m$ drawn at random according to P^m. Then for all* $0 < \epsilon \leq 1$,

$$\mathbf{E}(M(\epsilon/2, \mathcal{F}_{|\boldsymbol{X}}, l_\infty)) \geq M(\epsilon, \mathcal{F}, L_1(P))(1 - M(\epsilon, \mathcal{F}, L_1(P))e^{-\epsilon^2 m/2}).$$

Proof. Choose $\epsilon > 0$. Let $\mathcal{G}$ be an ϵ-separated subset of $\mathcal{F}$ (with respect to $d_{L_1(P)}$), with $|\mathcal{G}| = M(\epsilon, \mathcal{F}, L_1(P))$. Then

$$\begin{aligned} \mathbf{E}(M(\epsilon/2, \mathcal{F}_{|\boldsymbol{X}}, l_\infty)) \geq\; & \mathbf{E}(|\{f \in \mathcal{G} \colon \forall g \in \mathcal{G}, g \neq f, \exists i \in \{1, \ldots, m\} \\ & |f(\boldsymbol{X}_i) - g(\boldsymbol{X}_i)| > \epsilon/2\}|) \end{aligned}$$

$$
\begin{aligned}
&= \sum_{f \in \mathcal{G}} \Pr(\forall g \in \mathcal{G}, g \neq f, \exists i \in \{1, \ldots, m\} |f(\boldsymbol{X}_i) - g(\boldsymbol{X}_i)| > \epsilon/2) \\
&= \sum_{f \in \mathcal{G}} (1 - \Pr(\exists g \in \mathcal{G}, g \neq f : \forall i \in \{1, \ldots, m\} \\
&\qquad |f(\boldsymbol{X}_i) - g(\boldsymbol{X}_i)| \leq \epsilon/2)) \\
&\geq \sum_{f \in \mathcal{G}} (1 - |\mathcal{G}| \max_{g \in \mathcal{G}, g \neq f} \Pr(\forall i \in \{1, \ldots, m\} |f(\boldsymbol{X}_i) - g(\boldsymbol{X}_i)| \leq \epsilon/2)).
\end{aligned}
$$

First, note that if $|f(\boldsymbol{X}_i) - g(\boldsymbol{X}_i)| \leq \epsilon/2$ for every $i \in \{1, \ldots, m\}$, $\frac{1}{m}\sum_{i=1}^{m} |f(\boldsymbol{X}_i) - g(\boldsymbol{X}_i)|$ must be less than or equal to $\epsilon/2$.

We have from the definition of $\mathcal{G}$,

$$\forall f, g \in \mathcal{G}, f \neq g, \int_{\mathcal{X}} |f(\zeta) - g(\zeta)| P(\zeta) d\zeta > \epsilon.$$

So we have $\int_{\mathcal{X}} |f(\zeta) - g(\zeta)| P(\zeta) d\zeta - \frac{1}{m}\sum_{i=1}^{m} |f(\boldsymbol{X}_i) - g(\boldsymbol{X}_i)| > \epsilon/2$ when $|f(\boldsymbol{X}_i) - g(\boldsymbol{X}_i)| \leq \epsilon/2$ for every $i \in \{1, \ldots, m\}$. Hoeffding's inequality (Hoeffding 1963) implies

$$\Pr\left(\int_{\mathcal{X}} |f(\zeta) - g(\zeta)| P(\zeta) d\zeta - \frac{1}{m}\sum_{i=1}^{m} |f(\boldsymbol{X}_i) - g(\boldsymbol{X}_i)| > \epsilon/2\right) \leq e^{-\epsilon^2 m/2}$$

Thus

$$
\begin{aligned}
\mathbf{E}(M(\epsilon/2, \mathcal{F}_{|\boldsymbol{X}}, l_\infty)) &\geq \sum_{f \in \mathcal{G}} (1 - |\mathcal{G}| e^{-\epsilon^2 m/2}) \\
&= |\mathcal{G}|(1 - |\mathcal{G}| e^{-\epsilon^2 m/2}).
\end{aligned}
$$

□

Lemma A.17 *Let $\mathcal{F}$ be a family of functions from a set $\mathcal{X}$ into $[0, 1]$. Let $0 < \epsilon \leq 1$ and suppose $\mathcal{F}$ is such that $0 < d = \mathrm{fat}_{\mathcal{F}}(\epsilon/8) < \infty$. Let P be a probability distribution on $\mathcal{X}$. Then*

$$M(\epsilon, \mathcal{F}, L_1(P)) < \exp\left(\frac{8d}{\ln 2} \ln^2\left(\frac{512d}{\epsilon^4 \ln 2}\right)\right). \tag{A.12}$$

Proof. From Corollary A.15 and Lemma A.16 we have for $m \geq 4d \log \frac{16d}{\epsilon}$,

$$\exp\left(\frac{2}{\ln 2} d \ln^2 \frac{64m}{\epsilon^2}\right) \geq M(\epsilon, \mathcal{F}, L_1(P))(1 - M(\epsilon, \mathcal{F}, L_1(P)) e^{-\epsilon^2 m/2})$$

for all probability distributions P on Z. If $\frac{2}{\epsilon^2} \ln(2M(\epsilon, \mathcal{F}, L_1(P)) < 4d \log \frac{16d}{\epsilon}$, then the bound

(A.12) follows trivially. Hence we assume $\frac{2}{\epsilon^2}\ln(2M(\epsilon,\mathcal{F},L_1(P)) \geq 4d\log\frac{16d}{\epsilon}$. Also we assume $m \geq \frac{2}{\epsilon^2}\ln(2M(\epsilon,\mathcal{F},L_1(P))$, so $m \geq 4d\log\frac{16d}{\epsilon}$. With such an m we also have

$$(1 - M(\epsilon,\mathcal{F},L_1(P))e^{-\epsilon^2 m/2}) \geq 1/2.$$

Thus we obtain

$$\exp\left(\frac{2}{\ln 2} d\ln^2\frac{128\ln 2M(\epsilon,\mathcal{F},L_1(P))}{\epsilon^4}\right) \geq \frac{1}{2}M(\epsilon,\mathcal{F},L_1(P)).$$

Hence we have

$$\frac{2}{\ln 2} d\ln^2\left(\frac{128\ln M(\epsilon,\mathcal{F},L_1(P))}{\epsilon^4} + \frac{128\ln 2}{\epsilon^4}\right) + \ln 2 \geq \ln M(\epsilon,\mathcal{F},L_1(P)).$$

This cannot be true for

$$\ln M(\epsilon,\mathcal{F},L_1(P)) \geq \frac{8d}{\ln 2}\ln^2\left(\frac{512d}{\epsilon^4\ln 2}\right).$$

□

Lemma 3.9 follows from Lemma A.17 and the fact that $\mathcal{F}$ can be transformed into a class of $[0,1]$-valued functions by adding T to the functions and then dividing the result by $2T$.

Appendix B

Proofs of Results from Chapter 7

In Section B.1, we give the proof of Theorems 7.1 and 7.2. In Section B.2, we give the proof of Theorem 7.3.

B.1 Proof of Theorems 7.1 and 7.2

Let the target conditional expectation be f^*, that is $f^*(x) = \mathbf{E}[Y|x]$. Let f_a be the best approximation in Γ_q to f^*, that is $\mathbf{E}(Y - f_a(X))^2 = \inf_{f \in \mathcal{F}} \mathbf{E}(Y - f(X))^2$. (For convenience, we will assume $f_a \in \Gamma_q$, otherwise we can always find an $f \in \Gamma_q$ such that $\mathbf{E}(Y - f(X))^2$ is arbitrarily close to $\mathbf{E}(Y - f_a(X))^2$.) The range of the functions in Γ_q is bounded, that is $|f(x)| \leq T$ for every $x \in [-\pi, \pi]^n$ and every $f \in \Gamma_q$. Suppose the absolute value of the target observations is bounded also by T, that is $|Y| < T$. For simplicity, throughout the proofs we will assume that $T \geq 1$, $C \geq 1$ (the bound on the moments) and $M \geq 1$ (the L_1 norm of the functions).

For agnostic learning, we require that the learner produce an hypothesis $\hat{f}$ such that with probability at least $1 - \delta$,

$$\mathbf{E}(Y - \hat{f}(X))^2 - \mathbf{E}(Y - f_a(X))^2 \leq \epsilon.$$

Note that $f_a(x)$ can be represented as

$$f_a(x) = \mathrm{Re} \int_{\mathbb{R}^n} e^{i2\pi u \cdot x} F_a(u) du,$$

where F_a is the Fourier transform of f_a. To evaluate an approximation to the integral, we multiply the Fourier transform with an appropriate window and use a Monte Carlo sampling procedure over a restricted set of parameters. Let $R(r) = \{u \in \mathbb{R}^n : |u_i| < r, i = 1, \ldots, n\}$ and let the window $W(u) = 1$ if $u \in R(r)$ and $W(u) = 0$ otherwise. The windowed approximation to $f_a(x)$ is

$$f_r(x) = \text{Re} \int_{R(r)} e^{i2\pi u \cdot x} F_a(u) du. \tag{B.1}$$

Let $V(r) = (2r)^n$ be the volume of $R(r)$ and let $f_k(x) = \frac{V(r)}{k} \sum_{j=1}^{k} \text{Re } e^{i2\pi U_i \cdot x} F_a(U_j)$ be the function produced by the Monte Carlo procedure by sampling uniformly from $R(r)$. We will use uniform convergence methods to give approximation bounds which are valid for all $x \in [-\pi, \pi]^n$.

The Monte Carlo procedure will give the required approximation if we know the values of $F_a(U_j)$ at the sampled points U_j, $j = 1, \ldots, k$. Unfortunately, we do not know the values of $F_a(U_j)$. However, we can represent Re $e^{i2\pi U_j \cdot x} F_a(u)$ as $a_j \cos(2\pi U_j \cdot x) + b_i \sin(2\pi U_j \cdot x)$ where a_j and b_j are to be learned from the data by minimising the empirical loss. The hypothesis $\hat{f}$ is selected by choosing a linear combination of the sinusoidal basis functions which produces small empirical loss.

It is also possible to use a linear combination of linear threshold functions to approximate the sinusoidal functions and this forms the basis of the proof for efficient learning of the function class Γ_q using linear combinations of linear threshold functions. Similarly, sigmoids can be used to approximate linear threshold functions and hence can be used to learn the function class.

For the proof, it is convenient to rewrite $\mathbf{E}(Y - \hat{f}(X))^2 - \mathbf{E}(Y - f_a(X))^2$ in the following way.

$$\begin{aligned}
&\mathbf{E}(Y - \hat{f}(X))^2 - \mathbf{E}(Y - f_a(X))^2 = \mathbf{E}(f^*(X) - \hat{f}(X))^2 - \mathbf{E}(f^*(X) - f_a(X))^2 \\
= \; &\underbrace{\mathbf{E}(f^*(X) - \hat{f}(X))^2 - \mathbf{E}(f^*(X) - f_k(X))^2}_{\text{Estimation Error}} + \underbrace{\mathbf{E}(f^*(X) - f_k(X))^2 - \mathbf{E}(f^*(X) - f_r(X))^2}_{\text{Monte Carlo Error}} \\
&+ \underbrace{\mathbf{E}(f^*(X) - f_r(X))^2 - \mathbf{E}(f^*(X) - f_a(X))^2}_{\text{Windowing Error}}.
\end{aligned}$$

We will bound the error caused by the windowing procedure $\mathbf{E}(f^*(X) - f_r(X))^2 - \mathbf{E}(f^*(X) - f_a(X))^2$ in Section B.1.1. In Section B.1.2, we will bound the error caused by the Monte Carlo procedure $\mathbf{E}(f^*(X) - f_k(X))^2 - \mathbf{E}(f^*(X) - f_r(X))^2$. In Section B.1.3 we will bound the estimation error from learning the linear combination of the basis functions $\mathbf{E}(f^*(X) - \hat{f}(X))^2 - \mathbf{E}(f^*(X) - f_k(X))^2$. We conclude the proofs by adding these three bounds together to provide a bound for $\mathbf{E}(Y - \hat{f}(X))^2 - \mathbf{E}(Y - f_a(X))^2$ in Section B.1.4.

B.1.1 Windowing Error

In this section, we will bound $\mathbf{E}(f^*(X) - f_r(X))^2 - \mathbf{E}(f^*(X) - f_a(X))^2$, the error caused by windowing the Fourier transform.

We will use the following lemma to bound the mean square error in terms of the Fourier transform of the functions.

Lemma B.1 *Let f be a real-valued function and F be the Fourier transform of f. Let $\int_{\mathbb{R}^n} |F(u)|du < \infty$ and P be a probability distribution. Then $\int_{\mathbb{R}^n} f(x)^2 dP(x) \leq (\int_{\mathbb{R}^n} |F(u)|du)^2$.*

Proof. The result follows from the fact that $|f(x)| \leq \int_{\mathbb{R}^n} |F(u)|du$. □

The following lemma gives the expected squared error between f_a and f_r.

Lemma B.2

$$\mathbf{E}(f_a(X) - f_r(X))^2 \leq \frac{C^2}{(2\pi r)^{2q}}.$$

Proof.

$$\begin{aligned}
\int_{\mathbb{R}^n} |F_a(u) - F_r(u)|du &= \int_{\mathbb{R}^n \setminus R(r)} |F_a(u) - F_r(u)|du \\
&\leq \sum_{i=1}^{n} \int_{|u_i| \geq r} \int_{\mathbb{R}^{n-1}} \frac{|2\pi u_i|^q |F_a(u)|}{|2\pi u_i|^q} du \\
&\leq \sum_{i=1}^{n} \frac{1}{(2\pi r)^q} \int_{\mathbb{R}^n} |2\pi u_i|^q |F_a(u)|du \\
&\leq \frac{C}{(2\pi r)^q}.
\end{aligned}$$

The lemma then follows from Lemma B.1 □

We can now bound $\mathbf{E}(f^*(X) - f_r(X))^2 - \mathbf{E}(f^*(X) - f_a(X))^2$. The following corollary gives the bound we require.

Corollary B.3 *Assume $r \geq \frac{1}{2\pi}$.*

$$\begin{aligned}
\mathbf{E}(f^*(X) - f_r(X))^2 - \mathbf{E}(f^*(X) - f_a(X))^2 &\leq \frac{C^2}{(2\pi r)^{2q}} + \frac{4TC}{(2\pi r)^q} \\
&< \frac{5TC^2}{(2\pi r)^q}.
\end{aligned}$$

Proof.

$$
\begin{aligned}
&\mathbf{E}(f^*(X) - f_r(X))^2 - \mathbf{E}(f^*(X) - f_a(X))^2 \\
&\qquad = \mathbf{E}(f_a(X) - f_r(X))^2 + 2\mathbf{E}(f^*(X) - f_a(X))(f_a(X) - f_r(X)).
\end{aligned}
$$

We have $|f^*(X) - f_a(X)| \leq 2T$. From Cauchy-Schwarz inequality and Lemma B.2,

$$
\begin{aligned}
\mathbf{E}(f^*(X) - f_a(X))(f_a(X) - f_r(X)) &\leq \sqrt{\mathbf{E}(f^*(X) - f_a(X))^2}\sqrt{\mathbf{E}(f_a(X) - f_r(X))^2} \\
&\leq \frac{2TC}{(2\pi r)^q}.
\end{aligned}
$$

So

$$
\mathbf{E}(f_a(X) - f_r(X))^2 + \mathbf{E}(f^*(X) - f_a(X))(f_a(X) - f_r(X)) \leq \frac{C^2}{(2\pi r)^{2q}} + \frac{4TC}{(2\pi r)^q}.
$$

□

A better bound can be obtained for the case of learning with noise, where $f^* = f_a$.

Corollary B.4 *Assume $f^* = f_a$. Then*

$$
\mathbf{E}(f^*(X) - f_r(X))^2 \leq \frac{C^2}{(2\pi r)^{2q}}.
$$

Proof. The proof follows from the proof of Corollary B.3 since the cross term is zero. □

B.1.2 Monte Carlo Approximation Error

In this section, we bound $\mathbf{E}(f^*(X) - f_k(X))^2 - \mathbf{E}(f^*(X) - f_r(X))^2$, where f_k is the function produced by the Monte Carlo procedure. For the approximation bounds, we use uniform convergence methods to obtain a sup-norm approximation of f_r with f_k.

We use three different types of basis functions and indicate this by setting f_k to f_k^s when using sinusoidal basis functions, to f_k^h when using linear threshold basis functions and to f_k^σ when using sigmoid basis functions.

We will use the following corollary of Corollary 3.3 which permits the random variable to take negative values.

Corollary B.5 *Let $\mathcal{F}$ be a permissible class of functions from $\mathcal{Z}$ to $[-M, M]$. Let P be any probability distribution on $\mathcal{Z}$. For $m \geq 1$, $\nu > 0$ and $0 < \alpha < 1$,*

$$P^m \left\{ \boldsymbol{z} \in \mathcal{Z}^m : \exists f \in \mathcal{F}, \left| \hat{\mathbf{E}}(f) - \mathbf{E}(f) \right| > \epsilon \right\} \\ \leq 8 \max_{\boldsymbol{z}' \in \mathcal{Z}^{2m}} N\left(\frac{\epsilon}{32}, \mathcal{F}_{|\boldsymbol{z}'}, l_1 \right) e^{-\epsilon^2 m / 256 M^2}. \tag{B.2}$$

Proof. Separate the random variables into its positive and negative components and bound them separately to accuracy $\epsilon/2$. The proof follows from the triangle inequality and union bound. □

Sinusoidal Basis Functions

First we consider approximating f_r using a linear combination of sinusoidal basis functions. Considering (B.1), there is a function $\theta: \mathbb{R} \to \mathbb{R}$ such that

$$\begin{aligned} f_r(x) &= \text{Re} \int_{R(r)} e^{i2\pi u \cdot x} |F_a(u)| e^{i\theta(u)} du \\ &= \int_{R(r)} |F_a(u)| \cos(2\pi u \cdot x + \theta(u)) du \\ &= \int_{R(r)} |F_a(u)| (\cos(\theta(u)) \cos(2\pi u \cdot x) - \sin(\theta(u)) \sin(2\pi u \cdot x)) du \\ &= \int_{R(r)} (2r)^n |F_a(u)| (\cos(\theta(u)) \cos(2\pi u \cdot x) - \sin(\theta(u)) \sin(2\pi u \cdot x)) dP_U(u), \end{aligned}$$

where P_U is the uniform distribution over $R(r)$.

We will sample U_i from P_U, and use results on the uniform convergence of empirical averages to the expected values of random variables to show that with high probability, the empirical average $f_k^s(x) = \frac{1}{k}\sum_{i=1}^k (2r)^n |F_a(U_i)| (\cos(\theta(U_i)) \cos(2\pi U_i \cdot x) - \sin(\theta(U_i)) \sin(2\pi U_i \cdot x))$ is a good approximation to $f_r(x)$ for all $x \in [-\pi, \pi]^n$. In order to do that, we first need to bound the covering number of the relevant function class.

Lemma B.6 *Let $\mathcal{G}^s = \{u \mapsto a(u)\cos(2\pi u \cdot x) + b(u)\sin(2\pi u \cdot x) : |x_i| \leq \pi, |u_i| \leq r\}$ where $a(u) = (2r)^n |F_a(u)| \cos(\theta(u))$ and $b(u) = -(2r)^n |F_a(u)| \sin(\theta(u))$. For any $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$, $\boldsymbol{u}_i \in [-r, r]^n$,*

$$N(\epsilon, \mathcal{G}^s_{|\boldsymbol{u}}, l_1) \leq 2 \left(\frac{16\pi^2 e n 2^n r^{n+1} M}{\epsilon} \ln \frac{16\pi^2 e n 2^n r^{n+1} M}{\epsilon} \right)^n.$$

Proof. Note that $|a(u)| \leq (2r)^n M$ and $|b(u)| \leq (2r)^n M$. We also have $|\cos(2\pi\alpha) -$

$\cos(2\pi\beta))| \leq 2\pi|\alpha - \beta|$ and $|\sin(2\pi\alpha) - \sin(2\pi\beta))| \leq 2\pi|\alpha - \beta|$. Let $\boldsymbol{u} \in [-r, r]^m$. Let $\mathcal{L} = \{u \mapsto u \cdot x : |x_i| \leq \pi, |u_i| \leq r\}$ and let $\mathcal{C}$ be an $\epsilon/(4\pi(2r)^n M)$-cover for $\mathcal{L}_{|\boldsymbol{u}}$. Then for any $x \in [-\pi, \pi]^n$, there exists a $c \in \mathcal{C}$ such that

$$
\begin{aligned}
&\frac{1}{m}\sum_{i=1}^{m} |a(u)(\cos(2\pi u_i \cdot x) - \cos(2\pi c_i)) + b(u)(\sin(2\pi u_i \cdot x) - \sin(2\pi c_i))| \\
&\leq \frac{1}{m}\sum_{i=1}^{m} |a(u)||(\cos(2\pi u_i \cdot x) - \cos(2\pi c_i))| + |b(u)||(\sin(2\pi u_i \cdot x) - \sin(2\pi c_i))| \\
&\leq \frac{1}{m}\sum_{i=1}^{m} 2\pi(2r)^n M|u_i \cdot x - c_i| + 2\pi(2r)^n M|u_i \cdot x - c_i| \\
&= 4\pi(2r)^n M \frac{1}{m}\sum_{i=1}^{m} |u_i \cdot x - c_i| \\
&\leq \epsilon.
\end{aligned}
$$

Hence $N(\epsilon, \mathcal{G}^s_{|\boldsymbol{u}}, l_1) \leq N(\epsilon/4\pi(2r)^n M, \mathcal{L}_{|\boldsymbol{u}}, l_1)$. From Lemma 3.8 $N(\epsilon/4\pi(2r)^n M, \mathcal{L}_{|\boldsymbol{u}}, l_1) \leq 2\left(\frac{16\pi^2 en2^n r^{n+1}M}{\epsilon} \ln \frac{16\pi^2 en2^n r^{n+1}M}{\epsilon}\right)^n$. □

Knowing a bound on the covering number, we can now bound the number of basis functions needed for the required approximation.

Lemma B.7 *Let U_i, $i = 1, \ldots, k$ be uniformly sampled from $R(r)$. Then, with probability at least $1 - \delta$, for*

$$k \geq \frac{1024(2r)^{2n}M^2}{\epsilon^2}\left(n \ln\left(\frac{512\pi^2 en2^n r^{n+1}M}{\epsilon} \ln \frac{512\pi^2 en2^n r^{n+1}M}{\epsilon}\right) + \ln\frac{16}{\delta}\right),$$

$|f_r(x) - f^s_k(x)| \leq \epsilon$ *for all* $x \in [-1, 1]^n$.

Proof. Let $\mathcal{G}^s$ be the class $\{u \mapsto g(x, u) : g(x, u) = (2r)^n|F_a(u)|(\cos(\theta(u))\cos(2\pi u \cdot x) - \sin(\theta(u))\sin(2\pi u \cdot x)), |x_i| \leq 1, u \in R(r)\}$. From Corollary B.5 and Lemma B.6, we have

$$
\begin{aligned}
&P^k_U\left\{u^k : \exists x, \left|\frac{1}{k}\sum_{i=1}^{k} g(x, u_i) - \int_{R(r)} g(x, u)dP_U(u)\right| > \epsilon\right\} \\
&\quad \leq 16\left(\frac{512\pi^2 en2^n r^{n+1}M}{\epsilon} \ln \frac{512\pi^2 en2^n r^{n+1}M}{\epsilon}\right)^n \exp(\epsilon^2 k/1024(2r)^{2n}M^2).
\end{aligned}
$$

Setting the right hand side to be less than or equal to δ, we obtain

$$k \geq \frac{1024(2r)^{2n}M^2}{\epsilon^2}\left(n \ln\left(\frac{512\pi^2 en2^n r^{n+1}M}{\epsilon} \ln \frac{512\pi^2 en2^n r^{n+1}M}{\epsilon}\right) + \ln\frac{16}{\delta}\right). \square$$

We can now give a bound on $\mathbf{E}(f^*(X)-f_k^s(X))^2-\mathbf{E}(f^*(X)-f_r(X))^2$. Note that $|f_r(x)| \leq \int_{\mathbb{R}^n}|F_a(u)|du = \int_{[-1/2,1/2]^n}|F_a(u)|du + \int_{\mathbb{R}^n\setminus[-1/2,1/2]^n}|F_a(u)|du$. We have $|F_a(u)| < M$ and $\int_{\mathbb{R}^n\setminus[-1/2,1/2]^n}|F_a(u)|du \leq \int_{\mathbb{R}^n\setminus[-1/2,1/2]^n}\sum_{i=1}^n|2\pi u_i||F_a(u)|du \leq C$. Hence $|f_r(x)| \leq M+C$.

Corollary B.8 *Let $U_i,\ i=1,\dots,k$ be uniformly sampled from $R(r)$. Then, with probability at least $1-\delta$, for*

$$k \geq \frac{1024(2r)^{2n}M^2}{\epsilon^2}\left(n\ln\left(\frac{512\pi^2en2^nr^{n+1}M}{\epsilon}\ln\frac{512\pi^2en2^nr^{n+1}M}{\epsilon}\right)+\ln\frac{16}{\delta}\right),$$

$$\mathbf{E}(f^*(X)-f_k^s(X))^2-\mathbf{E}(f^*(X)-f_r(X))^2 \leq \epsilon^2+2(T+M+C)\epsilon.$$

If $f^*=f_a$ (for learning with noise) and instead

$$k \geq \frac{1024(2r)^{2n}M^2}{\epsilon}\left(n\ln\left(\frac{512\pi^2en2^nr^{n+1}M}{\sqrt{\epsilon}}\ln\frac{512\pi^2en2^nr^{n+1}M}{\sqrt{\epsilon}}\right)+\ln\frac{16}{\delta}\right),$$

then

$$\mathbf{E}(f^*(X)-f_k^s(X))^2-\mathbf{E}(f^*(X)-f_r(X))^2 \leq \epsilon+\frac{2C\sqrt{\epsilon}}{(2\pi r)^q}.$$

Proof.

$$\begin{aligned}&\mathbf{E}(f^*(X)-f_k^s(X))^2-\mathbf{E}(f^*(X)-f_r(X))^2 = \\ &\quad \mathbf{E}(f_r(X)-f_k^s(X))^2+2\mathbf{E}(f^*(X)-f_r(X))(f_r(X)-f_k^s(X)).\end{aligned}$$

We have $|f^*(X)-f_r(X)| \leq T+M+C$. Hence, from Lemma B.7

$$\mathbf{E}(f_r(X)-f_k^s(X))^2+2\mathbf{E}(f^*(X)-f_r(X))(f_r(X)-f_k^s(X)) \leq \epsilon^2+2(T+M+C)\epsilon.$$

For learning with noise, we use the Cauchy-Schwarz inequality to get

$$\begin{aligned}\mathbf{E}(f^*(X)-f_r(X))(f_r(X)-f_k^s(X)) &\leq \sqrt{\mathbf{E}(f^*(X)-f_r(X))^2}\sqrt{\mathbf{E}(f_r(X)-f_k^s(X))^2} \\ &\leq \frac{C\sqrt{\epsilon}}{(2\pi r)^q}\end{aligned}$$

The result follows from Lemma B.2, Lemma B.7 (by replacing ϵ with $\sqrt{\epsilon}$ and the assumption $f^*=f_a$. □

Linear Threshold Basis Functions

Let h be the threshold function, $h(x) = 1$ if $x \geq 0$ and $h(x) = 0$ otherwise. We can write $f_r(x)$ as an integral involving linear threshold functions as follows:

$$\begin{aligned}
f_r(x) &= \int_{R(r)} (2r)^n |F_a(u)| (\cos(\theta(u))\cos(2\pi u\cdot x) - \sin(\theta(u))\sin(2\pi u\cdot x)) dP_U(u) \\
&= \int_{R(r)} (2r)^n |F_a(u)| \left(\cos(\theta(u))\left(\int_{t=-2\pi^2 nr}^{2\pi u\cdot x} -\sin(t)dt + \cos(-2\pi^2 nr)\right)\right. \\
&\qquad \left. - \sin(\theta(u))\left(\int_{t=-2\pi^2 nr}^{2\pi u\cdot x} \cos(t)dt + \sin(-2\pi^2 nr)\right)\right) dP_U(u) \\
&= \int_{R(r)} (2r)^n |F_a(u)| \left(\cos(\theta(u))\left(\int_{t=-2\pi^2 nr}^{2\pi^2 nr} -\sin(t)h(2\pi u\cdot x - t)dt + \cos(-2\pi^2 nr)\right)\right. \\
&\qquad \left. - \sin(\theta(u))\left(\int_{t=-2\pi^2 nr}^{2\pi^2 nr} \cos(t)h(2\pi u\cdot x - t)dt + \sin(-2\pi^2 nr)\right)\right) dP_U(u) \\
&= \int_{R(r)} (2r)^n |F_a(u)| \left(\int_{t=-2\pi^2 nr}^{2\pi^2 nr} -\sin(\theta(u)+t)h(2\pi u\cdot x - t)dt + \right. \\
&\qquad \left. \cos(\theta(u) - 2\pi^2 nr)\right) dP_U(u).
\end{aligned}$$

Let $f_k^h = \frac{1}{k}\sum_{i=1}^k (2r)^n 4\pi nr|F_a(U_i)|(-\sin(\theta(U_i)+\tau_i)h(2\pi U_i\cdot x - \tau_i) + \cos(\theta(U_i) - 2\pi^2 nr))$.

Lemma B.9 *Suppose* (U_i, τ_i), $i = 1, \ldots, k$ *are uniformly sampled from* $R(r)\times[-2\pi^2 nr, 2\pi^2 nr]$. *(Denote the probability distribution by* $P_U \times P_\tau$.*) Then, with probability at least* $1-\delta$, *for*

$$k \geq \frac{16384(2r)^{2n}\pi^4(nrM)^2}{\epsilon^2}\left(n\ln\left(\frac{512e(2r)^n\pi^2 nrM}{\epsilon}\ln\frac{512e(2r)^n\pi^2 nrM}{\epsilon}\right) + \ln\frac{16}{\delta}\right),$$

$|f_r(x) - f_k^h(x)| \leq \epsilon$ *for all* $x \in [-\pi,\pi]^n$.

Proof. Let

$$\begin{aligned}
\mathcal{G}^h &= \{(u,t) \mapsto g(x,u,t) : g(x,u,t) = (2r)^n 4\pi^2 nr|F_a(u)|\left(-\sin(\theta(u)+t)h(2\pi u\cdot x - t) + \right. \\
&\qquad \left. \cos(\theta(u) - 2\pi^2 nr)\right), x \in [-\pi,\pi]^n\}.
\end{aligned}$$

The class $\mathcal{G}^h$ has pseudo-dimension n. From Corollary B.5 and Lemma 3.8, we have for $k \geq 1$,

$$(P_U \times P_\tau)^k\left\{(u,t)^k : \exists x, \left|\frac{1}{k}\sum_{i=1}^k g(x,U_i,\tau_i) - \int_{R(r)} g(x,u,t)dP_U(u)dP_\tau(t)\right| > \epsilon\right\}$$

$$\leq 16\left(\frac{512e(2r)^n\pi^2 nrM}{\epsilon}\ln\frac{512e(2r)^n\pi^2 nrM}{\epsilon}\right)^n \exp(-\epsilon^2 m/(16384(2r)^{2n}\pi^4(nrM)^2)$$

Setting the right hand side to be less than or equal to δ, we see that

$$k \;\geq\; \frac{16384(2r)^{2n}\pi^4(nrM)^2}{\epsilon^2}\left(n\ln\left(\frac{512e(2r)^n\pi^2 nrM}{\epsilon}\ln\frac{512e(2r)^n\pi^2 nrM}{\epsilon}\right)+\ln\frac{16}{\delta}\right)$$

will give the required result. □

The following corollaries give bounds on the error terms we require.

Corollary B.10 *Let* (U_i, τ_i), $i = 1, \ldots, k$ *be uniformly sampled from* $R(r) \times [-2\pi^2 nr, 2\pi^2 nr]$. *Then, with probability at least* $1 - \delta$, *for*

$$k \;\geq\; \frac{16384(2r)^{2n}\pi^4(nrM)^2}{\epsilon^2}\left(n\ln\left(\frac{512e(2r)^n\pi^2 nrM}{\epsilon}\ln\frac{512e(2r)^n\pi^2 nrM}{\epsilon}\right)+\ln\frac{16}{\delta}\right),$$

$$\mathbf{E}(f^*(X) - f_k^h(X))^2 - \mathbf{E}(f^*(X) - f_r(X))^2 \leq \epsilon^2 + 2(T + K + C)\epsilon.$$

If $f^* = f_a$ *(for learning with noise) and instead*

$$k \;\geq\; \frac{16384(2r)^{2n}\pi^4(nrM)^2}{\epsilon}\left(n\ln\left(\frac{512e(2r)^n\pi^2 nrM}{\sqrt{\epsilon}}\ln\frac{512e(2r)^n\pi^2 nrM}{\sqrt{\epsilon}}\right)+\ln\frac{16}{\delta}\right),$$

then

$$\mathbf{E}(f^*(X) - f_k^h(X))^2 - \mathbf{E}(f^*(X) - f_r(X))^2 \leq \epsilon + \frac{2C\sqrt{\epsilon}}{(2\pi r)^q}.$$

The proof is essentially identical to the proof for Corollary B.8.

Sigmoid Basis Functions

The sigmoid function approximates the linear threshold function as the weight size grows. It is possible to bound the error of the approximation as a function of the weight size. Let

$$f^\sigma = \int_{R(r)} (2r)^n |F_a(u)| \left(\int_{t=-2\pi^2 nr}^{2\pi^2 nr} -\sin(\theta(u)+t)\sigma(2\pi^2 u\cdot x - t)dt \right.$$
$$\left. + \cos(\theta(u) - 2\pi^2 nr)\right) dP_U(u)$$

We have, for $\alpha > 0$,

$$\begin{aligned}|f_r(x) - f^\sigma(x)| &\leq \int_{R(r)} (2r)^n |F_a(u)| \left(\int_{t=-2\pi^2 nr}^{2\pi^2 nr} |-\sin(\theta(u)+t)||h(2\pi u \cdot x - t) \right. \\ &\qquad \left. -\sigma(\alpha(2\pi^2 u \cdot x - t))|dt \right) dP_U(u) \\ &\leq \frac{2(M+C)\ln 2}{\alpha}\end{aligned}$$

where h is the threshold function and σ is the standard sigmoid because for any $\alpha > 0$,

$$\int_{-\infty}^{\infty} |h(x) - \sigma(\alpha x)| dx = \frac{2 \ln 2}{\alpha}.$$

Let $f_k^\sigma = \frac{1}{k}\sum_{i=1}^k (2r)^n 4\pi nr |F_a(U_i)|(-\sin(\theta(U_i)+\tau_i)\sigma(2\pi U_i \cdot x - \tau_i) + \cos(\theta(U_i) - 2\pi^2 nr))$. We can bound the error term we require in the following way:

$$\begin{aligned}&\mathbf{E}(f^*(X) - f_k^\sigma(X))^2 - \mathbf{E}(f^*(X) - f_r(X))^2 \\ =\ & \mathbf{E}(f^*(X) - f_k^\sigma(X))^2 - \mathbf{E}(f^*(X) - f^\sigma(X))^2 + \\ &\qquad \mathbf{E}(f^*(X) - f^\sigma(X))^2 - \mathbf{E}(f^*(X) - f_r(X))^2 \\ =\ & \mathbf{E}(f^*(X) - f_k^\sigma(X))^2 - \mathbf{E}(f^*(X) - f^\sigma(X))^2 + \\ &\qquad \mathbf{E}(f_r(X) - f^\sigma(X))^2 + 2\mathbf{E}(f^*(X) - f_r(X))(f_r(X) - f^\sigma(X)) \\ \leq\ & \mathbf{E}(f^*(X) - f_k^\sigma(X))^2 - \mathbf{E}(f^*(X) - f^\sigma(X))^2 + \\ &\qquad \frac{(2(M+C)\ln 2)^2}{\alpha^2} + \frac{4(T+M+C)(M+C)\ln 2}{\alpha}. \qquad \text{(B.3)}\end{aligned}$$

It remains only to bound $\mathbf{E}(f^*(X) - f_k^\sigma(X))^2 - \mathbf{E}(f^*(X) - f^\sigma(X))^2$. We proceed in the same way as in Section B.1.2

Lemma B.11 *Let (U_i, τ_i), $i = 1, \ldots, k$ be uniformly sampled from $R(r) \times [-2\pi^2 nr, 2\pi^2 nr]$. Then, with probability at least $1 - \delta$, for*

$$k \geq \frac{16384(2r)^{2n}\pi^4(nrM)^2}{\epsilon^2}\left(n \ln\left(\frac{512e(2r)^n\pi^2 nrM}{\epsilon} \ln \frac{512e(2r)^n\pi^2 nrM}{\epsilon}\right) + \ln\frac{16}{\delta}\right),$$

$|f^\sigma(x) - f_k^\sigma(x)| \leq \epsilon$ *for all* $x \in [-\pi, \pi]^n$.

Proof. Let

$$\mathcal{G}^\sigma = \{(u,t) \mapsto g(x,u,t) : g(x,u,t) = (2r)^n 2\pi nr |F_a(u)| \left(-\sin(\theta(u)+t)\sigma(2\pi u \cdot x - t) + \right.$$

$\cos(\theta(u) - 2\pi nr))\,, x \in [-\pi, \pi]^n\}.$

The class $\mathcal{G}^\sigma$ has pseudo-dimension n. The rest of the proof is identical to the proof of Lemma B.9. □

Corollary B.12 *Let (U_i, τ_i), $i = 1, \ldots, k$ be uniformly sampled from $R(r) \times [-2\pi^2 nr, 2\pi^2 nr]$. Then, with probability at least $1 - \delta$, for*

$$k \geq \frac{16384(2r)^{2n}\pi^4(nrM)^2}{\epsilon^2}\left(n \ln\left(\frac{512e(2r)^n\pi^2 nrM}{\epsilon}\ln\frac{512e(2r)^n\pi^2 nrM}{\epsilon}\right) + \ln\frac{16}{\delta}\right),$$

$$\mathbf{E}(f^*(X) - f_k^\sigma(X))^2 - \mathbf{E}(f^*(X) - f^\sigma(X))^2 \leq \epsilon^2 + 2(T + M + C + 2(M + C)\ln 2/\alpha)\epsilon.$$

If $f^* = f_a$ (for learning with noise) and instead

$$k \geq \frac{16384(2r)^{2n}\pi^4(nrM)^2}{\epsilon}\left(n \ln\left(\frac{512e(2r)^n\pi^2 nrM}{\sqrt{\epsilon}}\ln\frac{512e(2r)^n\pi^2 nrM}{\sqrt{\epsilon}}\right) + \ln\frac{16}{\delta}\right),$$

then

$$\mathbf{E}(f^*(X) - f_k^\sigma(X))^2 - \mathbf{E}(f^*(X) - f^\sigma(X))^2 \leq \epsilon + \frac{2C\sqrt{\epsilon}}{(2\pi r)^q} + \frac{4\sqrt{\epsilon}(M + C)\ln 2}{\alpha}.$$

Proof. The proof for the first part is similar to the proof for Corollary B.8. For the second part, we have

$$\begin{aligned}
& \mathbf{E}(f^*(X) - f_k^\sigma(X))^2 - \mathbf{E}(f^*(X) - f^\sigma(X))^2 \\
= \; & \mathbf{E}(f^\sigma(X) - f_k^\sigma(X))^2 + 2\mathbf{E}(f^*(X) - f^\sigma(X))(f^\sigma(X) - f_k^\sigma(X)) \\
\leq \; & \mathbf{E}(f^\sigma(X) - f_k^\sigma(X))^2 + 2\mathbf{E}|f^*(X) - f_r(X)||f^\sigma(X) - f_k^\sigma(X)| \\
& \qquad + 2\mathbf{E}|f_r(X) - f^\sigma(X)||f^\sigma(X) - f_k^\sigma(X)|.
\end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned}
\mathbf{E}|f^*(X) - f_r(X)||f^\sigma(X) - f_k^\sigma(X)| &\leq \sqrt{\mathbf{E}(f^*(X) - f_r(X))^2}\sqrt{\mathbf{E}(f^\sigma(X) - f_k^\sigma(X))^2} \\
&\leq \frac{C\sqrt{\epsilon}}{(2\pi r)^q}.
\end{aligned}$$

The result follows from Lemma B.2 and Lemma B.11. □

B.1.3 Estimation Error

With the randomly selected basis functions, all we need to learn is the second layer weights of the linear combinations. With the sinusoidal basis functions, we have $2k$ second layer weights (pseudo-dimension $2k$) while with the linear threshold and sigmoid basis functions we have $k+1$ second layer weights (pseudo-dimension $k+1$).

Since $|f_r(x)| \leq M + C$, if the approximation step in Section B.1.2 is successful $|f_k(x)| \leq M + C + \epsilon$. Let $\epsilon \leq T$. This gives $|f_k(x)| \leq M + C + T$. Least squared optimisation with the constraints $|f_k(x_i)| \leq M + C + T = B$, $i = 1, \ldots, m$ can be done in polynomial time. Note that with these constraints, the function class is still convex.

Rescale the functions and the target random variable by dividing by B. (The rescaling is just to calculate the sample complexity. There is no need for it in the actual algorithm.) In Theorem 3.7, setting $\nu = \nu_c = \epsilon/4B^2$, $\alpha = 1/2$, we get with probability at least $1-\delta$ that $\mathbf{E}(g_f) \leq 2\hat{\mathbf{E}}_{\boldsymbol{z}}(g_f) + \epsilon$ for sample $\boldsymbol{z}$ of size

$$\frac{7000B^2}{\epsilon}\left(d \ln\left(\frac{2048eB^2}{\epsilon} \ln \frac{2048eB^2}{\epsilon}\right) + d\ln 2 + \ln\frac{6}{\delta}\right),$$

where d is the pseudo-dimension of the function class. Recall that $g_f(x,y) = (y - f(x))^2 - (y - f_a(x))^2$ and $\mathbf{E}(g_f) = \mathbf{E}(f^* - f)^2 - \mathbf{E}(f^* - f_a)^2$. If we optimise to within ϵ of the best function, we obtain an expected mean squared error within at most 3ϵ of the best function.

B.1.4 Combining the Error Bounds

We can now combine the bounds from the previous sections to give bounds on the sample complexity and number of basis functions needed for learning the function class. Assuming the sampling and estimation step is successful, the algorithm will be successful. Since the probability of failure at each step is no more than δ, the probability that we will be unsuccessful is no more than 2δ. This can be rescaled to give δ. In each of the following sections, the accuracy will also have to be rescaled to give the desired accuracy ϵ. This does not change the order of the sample complexity and the order of the number of basis functions used.

Sinusoidal Basis Functions

For agnostic learning, select r in Corollary B.3 so that $\mathbf{E}(f^*(X)-f_r(X))^2-\mathbf{E}(f^*(X)-f_a(X))^2 \leq \frac{5TC^2}{(2\pi r)^q} = \epsilon$. From Corollary B.8, we can get

$$\mathbf{E}(f^*(X) - f_k^s(X))^2 - \mathbf{E}(f^*(X) - f_r(X))^2 \leq \epsilon$$

for (fixed n and q)

$$k = O\left(\frac{(TC^2)^{\frac{2n}{q}} M^2(T + M + C)}{\epsilon^{2+\frac{2n}{q}}} \left(\frac{n^2}{q} \ln\left(\frac{TCM}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right)$$

with n and q fixed.

For learning with noise, select r in Corollary B.4 so that $\mathbf{E}(f^*(X) - f_r(X))^2 \leq \frac{C^2}{(2\pi r)^{2q}} = \epsilon$. From Corollary B.8, we can get

$$\mathbf{E}(f^*(X) - f_k^s(X))^2 - \mathbf{E}(f^*(X) - f_r(X))^2 \leq 3\epsilon$$

for

$$k = O\left(\frac{C^{\frac{2n}{q}} M^2}{\epsilon^{1+\frac{n}{q}}} \left(\frac{n^2}{q} \ln\left(\frac{CM}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right).$$

Finally, the sample complexity for $\mathbf{E}(f^*(X) - \hat{f}(X))^2 - \mathbf{E}(f^*(X) - f_k^s(X))^2 = \epsilon$ is

$$O\left(\frac{B^2}{\epsilon}\left(k \ln\left(\frac{B^2}{\epsilon}\ln\frac{B^2}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right),$$

where $B = M + C + T$.

Linear Threshold Basis Functions

For agnostic learning, select r in Corollary B.3 so that $\mathbf{E}(f^*(X)-f_r(X))^2-\mathbf{E}(f^*(X)-f_a(X))^2 \leq \frac{5TC^2}{(2\pi r)^q} = \epsilon$. From Corollary B.10, we can get

$$\mathbf{E}(f^*(X) - f_k^h(X))^2 - \mathbf{E}(f^*(X) - f_r(X))^2 \leq \epsilon$$

for

$$k = O\left(\frac{(TC^2)^{\frac{2n+2}{q}} n^2 M^2(T + M + C)}{\epsilon^{2+\frac{2n+2}{q}}} \left(\frac{n^2}{q} \ln\left(\frac{TCM}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right).$$

For learning with noise, select r in Corollary B.4 so that $\mathbf{E}(f^*(X) - f_r(X))^2 \leq \frac{C^2}{(2\pi r)^{2q}} = \epsilon$. From Corollary B.10, we can get

$$\mathbf{E}(f^*(X) - f_k^s(X))^2 - \mathbf{E}(f^*(X) - f_r(X))^2 \leq 3\epsilon$$

for

$$k = O\left(\frac{C^{\frac{2n+2}{q}} n^2 M^2}{\epsilon^{1+\frac{n+1}{q}}}\left(\frac{n^2}{q}\ln\left(\frac{CM}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right).$$

Finally, the sample complexity for $\mathbf{E}(f^*(X) - \hat{f}(X))^2 - \mathbf{E}(f^*(X) - f_k^h(X))^2 = \epsilon$ is

$$O\left(\frac{B^2}{\epsilon}\left(k\ln\left(\frac{B^2}{\epsilon}\ln\frac{B^2}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right),$$

where $B = M + C + T$.

Sigmoid Basis Functions

For agnostic learning, select r in Corollary B.3 so that $\mathbf{E}(f^*(X) - f_r(X))^2 - \mathbf{E}(f^*(X) - f_a(X))^2 \leq \frac{5TC^2}{(2\pi r)^q} = \epsilon$. From Corollary B.12, we can get

$$\mathbf{E}(f^*(X) - f_k^\sigma(X))^2 - \mathbf{E}(f^*(X) - f^\sigma(X))^2 \leq \epsilon$$

for

$$k = O\left(\frac{((T+B)C^2)^{\frac{2n+2}{q}} n^2 M^2 (T+M+C)}{\epsilon^{2+\frac{2n+2}{q}}}\left(\frac{n^2}{q}\ln\left(\frac{(T+M+C)(T+B)CM}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right).$$

If $\alpha = \frac{4(T+M+C)(M+C)\ln 2}{\epsilon}$, from (B.3)

$$\begin{aligned}
& \mathbf{E}(f^*(X) - f_k^\sigma(X))^2 - \mathbf{E}(f^*(X) - f_r(X))^2 \\
\leq\ & \mathbf{E}(f^*(X) - f_k^\sigma(X))^2 - \mathbf{E}(f^*(X) - f^\sigma(X))^2 + \\
& \qquad \frac{(2(M+C)\ln 2)^2}{\alpha^2} + \frac{4(T+M+C)(M+C)\ln 2}{\alpha} \\
\leq\ & 3\epsilon
\end{aligned}$$

for small enough ϵ.

For learning with noise, select r in Corollary B.4 so that $\mathbf{E}(f^*(X) - f_r(X))^2 \leq \frac{C^2}{(2\pi r)^{2q}} = \epsilon$.

From Corollary B.12, we can get

$$\mathbf{E}(f^*(X) - f_k^\sigma(X))^2 - \mathbf{E}(f^*(X) - f^\sigma(X))^2 \leq 4\epsilon$$

for

$$k = O\left(\frac{C^{\frac{2n+2}{q}} n^2 M^2}{\epsilon^{1+\frac{n+1}{q}}}\left(\frac{n^2}{q}\ln\left(\frac{CM}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right)$$

if $\alpha \geq \frac{4(M+C)\ln 2}{\sqrt{\epsilon}}$. Set $\alpha = \frac{4(T+M+C)(M+C)\ln 2}{\epsilon}$. From (B.3)

$$\begin{aligned}
& \mathbf{E}(f^*(X) - f_k^\sigma(X))^2 - \mathbf{E}(f^*(X) - f_r(X))^2 \\
\leq\ & \mathbf{E}(f^*(X) - f_k^\sigma(X))^2 - \mathbf{E}(f^*(X) - f^\sigma(X))^2 + \\
& \qquad \frac{(2(M+C)\ln 2)^2}{\alpha^2} + \frac{4(T+M+C)(M+C)\ln 2}{\alpha} \\
\leq\ & 6\epsilon
\end{aligned}$$

for small enough ϵ. Finally, the sample complexity for $\mathbf{E}(f^*(X) - \hat{f}(X))^2 - \mathbf{E}(f^*(X) - f_k^\sigma(X))^2 = \epsilon$ is

$$O\left(\frac{B^2}{\epsilon}\left(k\ln\left(\frac{B^2}{\epsilon}\ln\frac{B^2}{\epsilon}\right) + \ln\frac{1}{\delta}\right)\right),$$

where $B = M + C + T$.

B.2 Proof of Theorem 7.3

The proof of Theorem 7.3 is similar to the proof of the approximation error component of Theorem 7.2. However, instead of just finding a set of basis functions which can be used for approximating a single function, we want to find a set of basis functions which can be used for uniformly approximating all functions in the class.

Let $R(r) = \{u \in \mathbb{Z}^n : u_j \leq r,\ j = 1, \ldots, n\}$. First we bound the error caused by truncating the Fourier series by excluding the terms outside $R(r)$. Let $f^* \in \Gamma_q^s$ be any function in Γ_q^s and let f_r^* be the corresponding function with the truncated Fourier series. Then

$$\begin{aligned}
|f^*(x) - f_r(x)| &= \left|\sum_{u\in\mathbb{Z}^n}(F^*(u) - F_r(u))e^{i2\pi u\cdot x}\right| \\
&\leq \sum_{|u_j|>r, u\in\mathbb{Z}^n} \frac{|2\pi u_j|^q |F^*(u)|}{|2\pi u_j|^q}
\end{aligned}$$

$$\leq \sum_{i=1}^{n} \frac{1}{(2\pi r)^q} \sum_{u \in \mathbb{Z}^n} |2\pi u_j|^q |F^*(u)|$$
$$\leq \frac{C}{(2\pi r)^q}.$$

Setting the truncation error to $\epsilon/2$ we get

$$r = \frac{(2C)^{1/q}}{2\pi\epsilon^{1/q}}. \tag{B.4}$$

Note that there are $(2r+1)^n$ Fourier coefficients in $R(r)$. Let $\tilde{\Gamma}_q$ be the class of functions which are represented by the class of truncated Fourier series of functions in Γ_q. We will now show that if we select a sample $U_1, \ldots, U_k$ (with an appropriate k) from $R(r)$ according to the uniform distribution, the probability that there exists a function f_r^* (with Fourier transform F^*) in $\tilde{\Gamma}_q$ and an $x \in [-\pi, \pi]^n$ such that $|f_r^*(x) - \frac{(2r+1)^n}{k} \sum_{j=1}^k \text{Re } F^*(U_j) e^{i2\pi U_j \cdot x}| > \epsilon/2$ is less than one. The value of k, such that this is true, is $k = O\left(\frac{n^3 M^4 C^{4n/q}}{q^2 \epsilon^{4+\frac{4n}{q}}}\left(\ln^2 \frac{CM}{\epsilon}\right)\right)$. This shows the existence of a set of basis functions of size k which can be used to uniformly approximate all truncated functions in that class to accuracy $\epsilon/2$. The result then follows from the triangle inequality.

To get the uniform convergence result, we require a bound for the covering number of the following function class

$$\mathcal{G} = \{u \mapsto (2r+1)^n |F(u,t)| \cos(\theta(u,t)) \cos(2\pi u \cdot x) - (2r+1)^n |F(u,t)| \times$$
$$\sin(\theta(u,t)) \sin(2\pi u \cdot x) : |x_i| \leq \pi, |u_j| \leq r, u_j \in \mathbb{Z}, t \in \mathcal{T}\}$$

where $\mathcal{T}$ is the set of indices for the Fourier coefficients of functions in Γ_q^s.

We will bound the covering number using the following lemma.

Lemma B.13 *Let* $\mathcal{G} = \{u \mapsto a(u,t)\cos(2\pi u \cdot x) + b(u,t)\sin(2\pi u \cdot x) : |x_i| \leq \pi, |u_j| \leq r, u_j \in \mathbb{Z}, t \in \mathcal{T}\}$ *where* $a(u,t) = (2r+1)^n |F(u,t)| \cos(\theta(u,t))$ *and* $b(u,t) = -(2r+1)^n |F(u,t)| \sin(\theta(u,t))$. *Let* $\mathcal{A} = \{u \mapsto a(u,t) : t \in \mathcal{T}, |u_j| \leq r, u_j \in \mathbb{Z}\}$ *and let* $\mathcal{B} = \{u \mapsto b(u,t) : t \in \mathcal{T}, |u_j| \leq r, u_j \in \mathbb{Z}\}$. *Then for any* $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$,

$$N(\epsilon, \mathcal{G}_{|\boldsymbol{u}}, l_1) \leq 2\left(\frac{32\pi^2 enr(2r+1)^n M}{\epsilon} \ln \frac{32\pi^2 enr(2r+1)^n M}{\epsilon}\right)^n \times$$
$$N(\epsilon/4, \mathcal{A}_{|\boldsymbol{u}}, l_1) N(\epsilon/4, \mathcal{B}_{|\boldsymbol{u}}, l_1).$$

Proof. Note that $|a(u)| \leq (2r+1)^n M$ and $|b(u)| \leq (2r+1)^n M$. We also have $|\cos(2\pi\alpha) - \cos(2\pi\beta))| \leq 2\pi|\alpha - \beta|$ and $|\sin(2\pi\alpha) - \sin(2\pi\beta))| \leq 2\pi|\alpha - \beta|$. Let $\boldsymbol{u} \in [-r, r]^m$. Let $\mathcal{L} = \{u \mapsto u \cdot x : |x_i| \leq \pi, |u_j| \leq r, u_j \in \mathbb{Z}\}$. Let $\mathcal{C}_1$ be an $\epsilon/(8\pi(2r+1)^n M)$-cover for $\mathcal{L}_{|\boldsymbol{u}}$, $\mathcal{C}_2$ be an $\epsilon/4$-cover for $\mathcal{A}_{|\boldsymbol{u}}$ and $\mathcal{C}_3$ be an $\epsilon/4$-cover for $\mathcal{B}_{|\boldsymbol{u}}$. Then for any $x \in [-\pi, \pi]^n$ and any $t \in \mathcal{T}$, there exists $c^1 \in \mathcal{C}_1$, $c^2 \in \mathcal{C}_2$ and $c^3 \in \mathcal{C}_3$ such that

$$
\begin{aligned}
& \frac{1}{m}\sum_{i=1}^{m} |a(\boldsymbol{u}_i, t)\cos(2\pi\boldsymbol{u}_i \cdot x) - c_i^2\cos(2\pi c_i^1) + b(\boldsymbol{u}_i, t)\sin(2\pi\boldsymbol{u}_i \cdot x) - c_i^3\sin(2\pi c_i^1))| \\
= \; & \frac{1}{m}\sum_{i=1}^{m} |a(\boldsymbol{u}_i, t)\cos(2\pi\boldsymbol{u}_i \cdot x) - a(\boldsymbol{u}_i, t)\cos(2\pi c_i^1) + a(\boldsymbol{u}_i, t)\cos(2\pi c_i^1) - c_i^2\cos(2\pi c_i^1) \\
& \quad + b(\boldsymbol{u}_i, t)\sin(2\pi\boldsymbol{u}_i \cdot x) + b(\boldsymbol{u}_i, t)\sin(2\pi c_i^1) + b(\boldsymbol{u}_i, t)\sin(2\pi c_i^1) - c_i^3\sin(2\pi c_i^1)| \\
\leq \; & \frac{1}{m}\sum_{i=1}^{m} |a(\boldsymbol{u}_i, t)(\cos(2\pi\boldsymbol{u}_i \cdot x) - \cos(2\pi c_i^1)) + (a(\boldsymbol{u}_i, t) - c_i^2)\cos(2\pi c_i^1)| \\
& \quad + |b(\boldsymbol{u}_i, t)(\sin(2\pi\boldsymbol{u}_i \cdot x) + \sin(2\pi c_i^1)) + (b(\boldsymbol{u}_i, t) - c_i^3)\sin(2\pi c_i^1)| \\
\leq \; & \frac{1}{m}\sum_{i=1}^{m} |a(\boldsymbol{u}_i, t)||\cos(2\pi\boldsymbol{u}_i \cdot x) - \cos(2\pi c_i^1)| + |a(\boldsymbol{u}_i, t) - c_i^2||\cos(2\pi c_i^1)| \\
& \quad + |b(\boldsymbol{u}_i, t)||\sin(2\pi\boldsymbol{u}_i \cdot x) + \sin(2\pi c_i^1)| + |b(\boldsymbol{u}_i, t) - c_i^3||\sin(2\pi c_i^1)| \\
\leq \; & \frac{1}{m}\sum_{i=1}^{m} 2\pi(2r+1)^n M|\boldsymbol{u}_i \cdot x - c_i^1| + |a(\boldsymbol{u}_i, t) - c_i^2| \\
& \quad + 2\pi(2r+1)^n M|\boldsymbol{u}_i \cdot x - c_i^1| + |b(\boldsymbol{u}_i, t) - c_i^3| \\
\leq \; & \epsilon.
\end{aligned}
$$

Hence $N(\epsilon, \mathcal{G}_{|\boldsymbol{u}}, l_1) \leq N(\epsilon/8\pi(2r+1)^n M, \mathcal{L}_{|\boldsymbol{u}}, l_1)N(\epsilon/4, \mathcal{A}_{|\boldsymbol{u}}, l_1)N(\epsilon/4, \mathcal{B}_{|\boldsymbol{u}}, l_1)$. From Lemma 3.8 $N(\epsilon/8\pi(2r+1)^n M, \mathcal{L}_{|\boldsymbol{u}}, l_1) \leq 2\left(\frac{32\pi^2 enr(2r+1)^n M}{\epsilon}\ln\frac{32\pi^2 enr(2r+1)^n M}{\epsilon}\right)^n$. □

We now bound the covering number for $\mathcal{A}$ and $\mathcal{B}$ (as defined in Lemma B.13). Note that functions from $\mathcal{A}$ are the (scaled) real part of the Fourier series, while functions from $\mathcal{B}$ are the (scaled) imaginary part. For a function $f \in \Gamma_q^s$, let

$$F_A(u, f) = (2r+1)^n \mathrm{Re}\int_{[-\pi,\pi]^n} f(x)e^{-i2\pi u \cdot x}dx = (2r+1)^n\int_{[-\pi,\pi]^n} f(x)\cos(2\pi u \cdot x)dx$$

and

$$F_B(u, f) = (2r+1)^n \mathrm{Im}\int_{[-\pi,\pi]^n} f(x)e^{-i2\pi u \cdot x}dx = (2r+1)^n\int_{[-\pi,\pi]^n} -f(x)\sin(2\pi u \cdot x)dx.$$

We will approximate the integral with a sum of a finite number of terms using the sup-norm approximation method similar to that used in (Barron 1992), and then bound the covering number of the sum.

Lemma B.14 *Let $\mathcal{A}$ be as defined in Lemma B.13. Let $\mathcal{H}_k = \{u \mapsto \frac{b}{k}\sum_{i=1}^{k}(2r+1)^n a_i \cos(2\pi u \cdot \boldsymbol{x}_i) : \boldsymbol{x}_i \in [-\pi,\pi]^n, |u_j| \leq r, u \in \mathbb{Z}^n, a_i \in \{-1,1\}, b \in [-M,M]\}$. Then for any $\boldsymbol{u} = (\boldsymbol{u}_1,\ldots,\boldsymbol{u}_m)$ and*

$$k \geq \frac{1024(2r+1)^{2n}M^2}{\epsilon^2}\left(n\ln(2r+1)+\ln 8\right),$$

$$N(\epsilon, \mathcal{A}_{|\boldsymbol{u}}, l_1) \leq N(\epsilon/2, \mathcal{H}_{k|\boldsymbol{u}}, l_1).$$

Proof. We can represent $F_A(u,f)$ as

$$F_A(u,f) = \int_{[-\pi,\pi]^n}(2r+1)^n M_f \; sign(f(x))\cos(2\pi u\cdot x)P(dx)$$

where $M_f = \int_{[-\pi,\pi]^n}|f(x)|dx \leq M$, $sign(f(x))$ is the sign of $f(x)$ and $P(dx) = |f(x)|dx/M_f$ is a probability distribution.

There are no more than $(2r+1)^n$ values of u that we are interested in. Corollary B.5 shows that

$$P^k\left\{\boldsymbol{x} \in ([-\pi,\pi]^n)^k \colon \exists u, \left|\frac{M_f}{k}\sum_{i=1}^{k}(2r+1)^n sign(f(\boldsymbol{x}_i))\cos(2\pi u\cdot\boldsymbol{x}_i) - F_A(u,f)\right| > \epsilon/2\right\}$$
$$\leq 8(2r+1)^n e^{-\epsilon^2 k/1024(2r+1)^{2n}M_f^2}.$$

Setting the right hand side equal to 1 shows that for

$$k \geq \frac{1024(2r+1)^{2n}M^2}{\epsilon^2}\left(n\ln(2r+1)+\ln 8\right), \tag{B.5}$$

for every $F_A(\cdot,f)$, there is a function of the form $u \mapsto \frac{1}{k}\sum_{i=1}^{k}(2r+1)^n M_f \; sign(f(\boldsymbol{x}_i))\cos(2\pi u \cdot \boldsymbol{x}_i)$ that is within $\epsilon/2$ of it for all the u's we are interested in. The triangle inequality ensures that an $\epsilon/2$ cover for $\mathcal{H}_k$ would also be an ϵ cover for $\mathcal{A}$. □

Lemma B.15 *Let $\mathcal{H}_k = \{u \mapsto \frac{b}{k}\sum_{i=1}^{k}(2r+1)^n a_i\cos(2\pi u\cdot\boldsymbol{x}_i) : \boldsymbol{x}_i \in [-\pi,\pi]^n, |u_j| \leq r, u \in \mathbb{Z}^n, a_i \in \{-1,1\}, b \in [-M,M]\}$. Then for any $\boldsymbol{u} = (\boldsymbol{u}_1,\ldots,\boldsymbol{u}_m)$, ($\boldsymbol{u}_j \in [-r,r]^n$ for $j = 1,\ldots,m$)*

$$N(\epsilon, \mathcal{H}_{k|\boldsymbol{u}}, l_1) \leq \frac{4\pi M(2r+1)^n}{\epsilon}2^{2k}\left(\frac{16\pi^2 e(2r+1)^n nrM}{\epsilon}\ln\frac{16\pi^2 e(2r+1)^n nrM}{\epsilon}\right)^{nk}.$$

Proof. We have $|\cos(2\pi\alpha) - \cos(2\pi\beta))| \leq 2\pi|\alpha - \beta|$. Let $\mathcal{L} = \{u \mapsto u \cdot x : 0 \leq x_i \leq 2\pi, |u_j| \leq r, u_j \in \mathbb{Z}\}$ and let $\mathcal{C}$ be an $\epsilon/(4\pi(2r+1)^n M)$-cover for $\mathcal{L}_{|\boldsymbol{u}}$. Let B be an $\epsilon/(4\pi(2r+1)^n)$-cover for $\{b \colon b \in [-M, M]\}$. Then there exists a $c_{ij} \in \mathcal{C}$ for each $\boldsymbol{u}_j \cdot \boldsymbol{x}_i$ and a $b' \in B$ for each $b \in [-M, M]$ such that

$$
\begin{aligned}
&\frac{1}{m}\sum_{j=1}^{m}\left|\frac{1}{k}\sum_{i=1}^{k}(2r+1)^n a_i(b\cos(2\pi \boldsymbol{u}_j \cdot \boldsymbol{x}_i) - b'\cos(2\pi c_{ij}))\right| \\
&\leq \frac{1}{k}\sum_{i=1}^{k}\frac{1}{m}\sum_{j=1}^{m}(2r+1)^n|(b\cos(2\pi \boldsymbol{u}_j \cdot \boldsymbol{x}_i) - b\cos(2\pi c_{ij}) + b\cos(2\pi c_{ij}) - b'\cos(2\pi c_{ij}))| \\
&\leq \frac{1}{k}\sum_{i=1}^{k}\frac{1}{m}\sum_{j=1}^{m}(2r+1)^n|(b(\cos(2\pi \boldsymbol{u}_j \cdot \boldsymbol{x}_i) - \cos(2\pi c_{ij})) + (b - b')\cos(2\pi c_{ij}))| \\
&\leq \frac{1}{k}\sum_{i=1}^{k}\frac{1}{m}\sum_{i=1}^{m}2\pi(2r+1)^n(M|\boldsymbol{u}_j \cdot \boldsymbol{x}_i - c_{ij}| + \epsilon/(4\pi(2r+1)^n) \\
&= \frac{1}{k}\sum_{i=1}^{k}2\pi(2r+1)^n M\frac{1}{m}\sum_{i=1}^{m}|\boldsymbol{u}_j \cdot \boldsymbol{x}_i - c_{ij}| + \epsilon/2 \\
&\leq \epsilon.
\end{aligned}
$$

The size of B is no more than $4\pi M(2r+1)^n/\epsilon$. Hence $N(\epsilon, \mathcal{G}_{c|\boldsymbol{u}}, l_1) \leq \frac{4\pi M(2r+1)^n}{\epsilon}2^k N(\epsilon/4\pi(2r+1)^n M, \mathcal{L}_{|\boldsymbol{u}}, l_1)^k$. From Lemma 3.8 $N(\epsilon/4\pi(2r+1)^n M, \mathcal{L}_{|\boldsymbol{u}}, l_1) \leq 2\left(\frac{16\pi^2 e(2r+1)^n nrM}{\epsilon}\ln\frac{16\pi^2 e(2r+1)^n nrM}{\epsilon}\right)^n$. $\square$

It is easy to see that the same bound applies to the covering number of $\mathcal{B}$. We can now bound the covering number of $\mathcal{G}$.

Corollary B.16 *Let* $\mathcal{G} = \{u \mapsto a(u,t)\cos(2\pi u \cdot x) + b(u,t)\sin(2\pi u \cdot x) : |x_i| \leq \pi, |u_j| \leq r, u_j \in \mathbb{Z}, t \in \mathcal{T}\}$ *where* $a(u,t) = (2r+1)^n|F(u,t)|\cos(\theta(u,t))$ *and* $b(u,t) = -(2r+1)^n|F(u,t)|\sin(\theta(u,t))$. *Let* $\mathcal{A} = \{u \mapsto a(u,t) : t \in \mathcal{T}, u_j \leq r, u_j \in \mathbb{Z}\}$ *and let* $\mathcal{B} = \{u \mapsto b(u,t) : t \in \mathcal{T}, u_j \leq r, u_j \in \mathbb{Z}\}$. *Then for any* $\boldsymbol{u} = (\boldsymbol{u}_1, \ldots, \boldsymbol{u}_m)$*,*

$$
\begin{aligned}
N(\epsilon, \mathcal{G}_{|\boldsymbol{u}}, l_1) \quad \leq \quad & 2\left(\frac{32\pi^2 enr(2r+1)^n M}{\epsilon}\ln\frac{32\pi^2 enr(2r+1)^n M}{\epsilon}\right)^n \frac{256\pi^2 M^2(2r+1)^{2n}}{\epsilon^2} \\
& 2^{4k}\left(\frac{128\pi^2 e(2r+1)^n nrM}{\epsilon}\ln\frac{128\pi^2 e(2r+1)^n nrM}{\epsilon}\right)^{2nk}
\end{aligned}
$$

where

$$k = \frac{16384(2r+1)^{2n}M^2}{\epsilon^2}\left(n\ln(2r+1) + \ln 8\right).$$

Given the covering number for $\mathcal{G}$, we can now obtain a the uniform convergence result, and hence give the appropriate number of basis functions k.

Let P_U be the uniform distribution on $R(r)$. Let $\tilde{\Gamma}_q^s$ be the class of functions formed from the truncated Fourier series of functions in Γ_q^s. For a function $f \in \Gamma_q^s$, let the corresponding function with truncated Fourier series be $f_r \in \tilde{\Gamma}_q^s$. Using Corollary B.16 and Corollary B.5 we get

$$\begin{aligned}
& P_U^k \Big\{ \boldsymbol{u}^k \colon \exists f \in \Gamma_q^s, x \in [-\pi,\pi]^n, \\
& \quad \left| \frac{1}{k} \sum_{i=1}^{k} (F_A(\boldsymbol{u}_i, f)\cos(2\pi \boldsymbol{u}_i \cdot x) - F_B(\boldsymbol{u}_i, f)\sin(2\pi \boldsymbol{u}_i \cdot x)) - f_r(x) \right| > \epsilon/2 \Big\} \\
& \quad \le 16 \left(\frac{2048\pi^2 enr(2r+1)^n M}{\epsilon} \ln \frac{2048\pi^2 enr(2r+1)^n M}{\epsilon} \right)^n \frac{512^2\pi^2 M^2 (2r+1)^{2n}}{\epsilon^2} \times \\
& \quad 2^{4k'} \left(\frac{8192\pi^2 e(2r+1)^n nrM}{\epsilon} \ln \frac{8192\pi^2 e(2r+1)^n nrM}{\epsilon} \right)^{2nk'} e^{-\epsilon^2 k/1024(2r+1)^{2n}M^2}
\end{aligned}$$

where

$$k' = \frac{67108864(2r+1)^{2n}M^2}{\epsilon^2} \left(n \ln(2r+1) + \ln 8 \right).$$

Setting the right hand side equal to 1 shows that for

$$\begin{aligned}
k \quad > \quad & \frac{1024(2r+1)^{2n}M^2}{\epsilon^2} \left(n \ln \left(\frac{2048\pi^2 enr(2r+1)^n M}{\epsilon} \ln \frac{2048\pi^2 enr(2r+1)^n M}{\epsilon} \right) + \right. \\
& 2 \ln \frac{512\pi M(2r+1)^n}{\epsilon} + 4k' \ln 2 + \\
& \left. 2nk' \ln \left(\frac{8192\pi^2 e(2r+1)^n nrM}{\epsilon} \ln \frac{8192\pi^2 e(2r+1)^n nrM}{\epsilon} \right) + \ln 16 \right),
\end{aligned}$$

there is a function of the form $\frac{1}{k}\sum_{i=1}^{k}(F_A(\boldsymbol{u}_i, f)\cos(2\pi \boldsymbol{u}_i \cdot x) - F_B(\boldsymbol{u}_i, f)\sin(2\pi \boldsymbol{u}_i \cdot x))$ that is within $\epsilon/2$ of $f_r(x)$ for all the $f_r \in \tilde{\Gamma}_q^s$ and $x \in [-\pi,\pi]^n$. Setting r appropriately to give truncation error $\epsilon/2$ (using (B.4)) completes the proof of Theorem 7.3.

References

Aho, A., Hopcroft, J. & Ullman, J. (1974), *The Design and Analysis of Computer Algorithms*, Addison-Wesley, London.

Alon, N., Ben-David, S., Cesa-Bianchi, N. & Haussler, D. (1993), Scale-sensitive dimensions, uniform convergence and learnability, *in* 'Proc. 35th Annu. IEEE Sympos. Found. Comput. Sci.'.

Barron, A. R. (1990), Complexity regularization with applications to artificial neural networks, *in* G. Roussa, ed., 'Nonparametric Functional Estimation', Kluwer Academic, Boston, MA and Dordrecht, the Netherlands, pp. 561–576.

Barron, A. R. (1992), Neural net approximation, *in* 'Proc. 7th Yale Workshop on Adaptive and Learning Systems'.

Barron, A. R. (1993), 'Universal approximation bounds for superposition of a sigmoidal function', *IEEE Trans. on Information Theory* **39**, 930–945.

Barron, A. R. (1994), 'Approximation and estimation bounds for artificial neural networks', *Machine Learning* **14**, 115–133.

Bartlett, P. L. & Long, P. M. (1995), More theorems about scale-sensitive dimensions and learning, *in* 'Proc. 8th Annu. Workshop on Comput. Learning Theory', ACM Press, New York, NY, pp. 392–401.

Bartlett, P. L., Long, P. M. & Williamson, R. C. (1994), Fat-shattering and the learnability of real-valued functions, *in* 'Proc. 7th Annu. ACM Workshop on Comput. Learning Theory', ACM Press, New York, NY, pp. 299–310.

Blum, A. & Rivest, R. (1992), 'Training a 3-node neural network is NP-complete', *Neural Networks* **5**, 117–127.

Blumer, A., Ehrenfeucht, A., Haussler, D. & Warmuth, M. K. (1989), 'Learnability and the Vapnik-Chervonenkis dimension', *J. ACM* **36**(4), 929–965.

Boser, B. E., Guyon, I. M. & Vapnik, V. N. (1992), A training algorithm for optimal margin classifiers, *in* 'Proc. 5th Annu. Workshop on Comput. Learning Theory', ACM Press, New York, NY, pp. 144–152.

Craig, C. C. (1933), 'On the Tchebychef inequality of Bernstein', *Annals of Mathematical Statistics* **4**, 94–102.

Darken, C., Donahue, M., Gurvits, L. & Sontag, E. (1993), Rate of approximation results motivated by robust neural network learning, *in* 'Proc. 6th Annu. Workshop on Comput. Learning Theory', ACM Press, New York, NY, pp. 303–309.

DasGupta, B., Siegelmann, H. T. & Sontag, E. (1995), 'On the complexity of training neural networks with continuous activation function', *IEEE Trans. on Neural Networks* **6**, 1490–1504.

Delyon, B., Juditsky, A. & Benveniste, A. (1995), 'Accuracy analysis for wavelets approximations', *IEEE Trans. on Neural Networks* **6**, 332–348.

Dudley, R. M. (1978), 'Central limit theorems for empirical measures', *Annals of Probability* **6**(6), 899–929.

Eubank, R. (1988), *Spline Smoothing and Nonparametric Regression*, Marcel Dekker, New York.

Farago, A. & Lugosi, G. (1993), 'Strong universal consistency of neural network classifiers', *IEEE Trans. on Information Theory* **39**, 1146–1151.

Fukunaga, K. (1972), *Introduction to Statistical Pattern Recognition*, Academic Press, New York and London.

Goldberg, P. & Jerrum, M. (1993), Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers, *in* 'Proc. 6th Annu. ACM Workshop on Comput. Learning Theory', ACM Press, New York, NY, pp. 361–369.

Gurvits, L. & Koiran, P. (1995), Approximation and learning of convex superposition, *in* 'Computational Learning Theory: EUROCOLT'95'.

Hardle, W. (1990), *Applied Nonparametric Regression*, Cambridge University Press, Cambridge, U.K. and New York.

Haussler, D. (1992), 'Decision theoretic generalizations of the PAC model for neural net and other learning applications', *Inform. Comput.* **100**(1), 78–150.

Hoeffding, W. (1963), 'Probability inequalities for sums of bounded random variables', *Journal of the American Statistical Association* **58**(301), 13–30.

Höffgen, K. & Simon, H. (1992), Robust trainability of single neurons, *in* 'Proc. 5th Annu. Workshop on Comput. Learning Theory', ACM Press, New York, NY, pp. 428–439.

Ibragimov, I. A. & Hasminskii, R. Z. (1980), 'On nonparametric estimation of regression', *Doklady Acad. Nauk SSSR* **252**, 780–784.

Igelnik, B. & Pao, Y. H. (1995), 'Stochastic choice of basis functions in adaptive function approximation and the functional-link net', *IEEE Trans. on Neural Networks* **6**, 1320–1329.

Jerrum, M. (1994), 'Simple translation-invariant concepts are hard to learn', *Inform. Comput.* **113**(2), 300–311.

Jones, L. K. (1992), 'A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training', *The Annals of Statistics* **20**, 608–613.

Judd, J. S. (1990), *Neural Network Design and the Complexity of Learning*, MIT Press.

Karpinski, M. & Macintyre, A. (1995), VC dimension of sigmoidal and general pfaffian networks, Technical Report TR95-055, Electronic Colloquium on Computational Complexity (ECCC) - http://www.eccc.uni-trier.de/eccc/.

Kearns, M. (1989), The Computational Complexity of Machine Learning, PhD thesis, Harvard University Center for Research in Computing Technology. Technical Report TR-13-89. Also published by MIT Press as an ACM Distinguished Dissertation.

Kearns, M. J. & Schapire, R. E. (1994), 'Efficient distribution-free learning of probabilistic concepts', *J. Comput. Syst. Sci.* **48**(3), 464.

Kearns, M. J., Schapire, R. E. & Sellie, L. M. (1994), 'Toward efficient agnostic learning', *Machine Learning* **17**(2), 115.

Kearns, M., Li, M., Pitt, L. & Valiant, L. (1987), On the learnability of Boolean formulae, *in* 'Proc. 19th Annu. ACM Sympos. Theory Comput.', ACM Press, New York, NY, pp. 285–294.

Koiran, P. (1994), Efficient learning of continuous neural networks, *in* 'Proc. 7th Annu. ACM Workshop on Comput. Learning Theory', ACM Press, New York, NY, pp. 348–355.

Le Cun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W. & Jackel, L. D. (1989), 'Backpropagation applied to handwritten zip code recognition', *Neural Computation* **1**, 541–551.

Maass, W. (1995), 'Agnostic PAC-learning of functions on analog neural networks', *Neural Computation* **7**(5), 1054–1078.

McCaffrey, D. F. & Gallant, A. R. (1994), 'Convergence rates for single hidden layer feedforward networks', *Neural Networks* **7**(1), 147–158.

Minsky, M. & Papert, S. (1969), *Perceptrons*, MIT Press, Cambridge, MA.

Nestorov, Y. & Nemirovskii, A. (1994), *Interior-Point Polynomial Algorithms in Convex Programming*, SIAM, Philadelphia.

Nussbaum, M. (1986), 'On nonparametric estimation of a regression function that is smooth in a domain of $\mathbb{R}^k$', *Theory of Probability and its Applications* **31**, 118–125.

Pinkus, A. (1985), *n-Widths in Approximation Theory*, Springer-Verlag, New York.

Pinsker, M. S. (1980), 'Optimal filtering of square-integrable signals on a background of Gaussian noise', *Problems in Information Transmission*.

Pitt, L. & Valiant, L. (1988), 'Computational limitations on learning from examples', *J. ACM* **35**, 965–984.

Pollard, D. (1984), *Convergence of Stochastic Processes*, Springer-Verlag, Berlin.

Pollard, D. (1990), *Empirical Processes: Theory and Applications*, Vol. 2 of *NSF-CBMS Regional Conference Series in Probability and Statistics*, Institute of Math. Stat. and Am. Stat. Assoc.

Pollard, D. (1995), Uniform ratio limit theorems for empirical processes, Submitted to Scandinavian Journal of Statistics.

Pomerleau, D. A. (1989), ALVINN: An autonomous land vehicle in a neural network, *in* D. S. Touretzky, ed., 'Advances in Neural Information Processing, Vol. 1', Morgan Kaufmann, pp. 305–313.

Rumelhart, D. E., Hinton, G. E. & Williams, R. J. (1986), Learning internal representations by error propagation, *in* 'Parallel Distributed Processing – Explorations in the Microstructure of Cognition', MIT Press, chapter 8, pp. 318–362.

Schapire, R. E. (1990), 'The strength of weak learnability', *Machine Learning* **5**(2), 197–227.

Silverman, B. W. (1986), *Density Estimation for Statistics and Data Analysis*, Chapman and Hall, New York.

Stone, C. J. (1982), 'Optimal global rates of convergence for nonparametric estimators', *Annals of Statistics* **10**, 1040–1053.

Tesauro, G. (1990), 'Neurogammon wins computer Olympiad', *Neural Computation* **1**, 321–323.

Tesauro, G. & Sejnowski, T. J. (1989), 'A 'neural' network that learns to play backgammon', *Artificial Intelligence* **39**(3), 357–390.

Valiant, L. G. (1984), 'A theory of the learnable', *Commun. ACM* **27**(11), 1134–1142.

Vapnik, V. N. (1982), *Estimation of Dependences Based on Empirical Data*, Springer-Verlag, New York.

Vapnik, V. N. & Chervonenkis, A. Y. (1971), 'On the uniform convergence of relative frequencies of events to their probabilities', *Theory of Probabability and its Applications* **16**(2), 264–280.

Yukich, J. E., Stinchcombe, M. B. & White, H. (1995), 'Sup-norm approximation bounds for networks through probabilistic methods', *IEEE Trans. on Information Theory* **41**, 1021–1027.

Errata

Pg 5 l -14 An additional reference (Alon et al. 1993) should be added here.

Pg 14 Sec 2.2 l -1 This line should be changed to: "Sinusoidal basis functions: $\{g(x) = \sin(v \cdot x), g(x) = \cos(v \cdot x) : v \in \mathbb{R}^n\}$."

Pg 15 l 2 It should be mentioned that for bounded measurable functions, the essential supremum is usually used in place of the supremum in the definition of the L_∞ norm.

Pg 20 In Table 3.2, assuming that $\text{fat}_{\mathcal{F}}(\epsilon)$ grows polynomially with $1/\epsilon$, the $\text{fat}_{\mathcal{F}}(\epsilon)$ term can be removed from inside the log.

Pg 23 l 3 This line should be changed from "suffices for agnostically learning $\mathcal{F}$" to "suffices for learning $\mathcal{F}$".

Pg 24 l 8 & Pg 25 l 13 Both these lines should be changed to:

$$m \geq \frac{7000T^2}{\epsilon}\left(\ln\left(\max_{\boldsymbol{x} \in \mathcal{X}^{2m}} N\left(\frac{\epsilon}{512T}, \mathcal{F}_{|\boldsymbol{x}}, l_1\right)\right) + \ln\frac{6}{\delta}\right)$$

Pg 27 l -5 This line should be changed from "networks with linear threshold hidden units and ..." to "networks with linear threshold hidden units (or sigmoidal hidden units) and ...".

Pg 31 It should be mentioned that for sets of at least two $\{0, 1\}$-valued functions, an $\Omega(1/\epsilon^2)$ lower bound on the sample complexity for proper agnostic learning follows from the results of the following two papers:

- L. Devroye and G. Lugosi. Lower bounds in pattern recognition and learning. *Pattern Recognition*, 28(7):1011-1018, 1995.
- H. U. Simon. General lower bounds on the number of examples needed for learning probabilistic concepts. *The 1993 Conference on Computational Learning Theory*,

pages 402-412, 1993.

Pg 31 l -1 The term k_2 shold be in the denominator, not the numerator.

Pg 31 l -10 The term "$\Omega(\ln(\delta)/\epsilon^2)$" should be changed to "$\Omega(\ln(1/\delta)/\epsilon^2)$".

Pg 40 Theorem 5.2 The sample complexity bound should be changed to

$$\frac{14000C^2}{\epsilon}\left(\frac{16C^2}{\epsilon}\left(\ln \max_{\boldsymbol{x}\in\mathcal{X}^{2m}}\left(N\left(\frac{\epsilon}{1024CK}, \mathcal{G}_{|\boldsymbol{x}}, l_1\right)+1\right)+\ln 2\right)+\ln\frac{6}{\delta}\right).$$

The bounds in Corollary 5.3 and 5.4 should also be changed accordingly. In the proof, the last equality on page 44, line 9 and 10 is incorrect. To correct the proof, on page 44 line 11, use Theorem 6.1 in place of Lemma 5.5. This results in the sample complexity bound given above.

Pg 42 l 7 The sentence "Let $f = \frac{1}{n}\sum_{i=1}^{k} f_i$..." should be replaced by "Let $f = \frac{1}{k}\sum_{i=1}^{k} f_i$...".

Pg 42 Eqn 5.1 $N(\epsilon, \mathcal{A}^{\mathcal{G}^1}_{K,k|\boldsymbol{x}}, l_1)$ should be replaced with $N(\epsilon, \mathcal{A}^{\mathcal{G}}_{K,k|\boldsymbol{x}}, l_1)$.

Pg 42 l -9 $f_{|\boldsymbol{x}} \in \mathcal{A}^{\mathcal{G}^1}_{K,k}$ should be replaced with $f_{|\boldsymbol{x}} \in \mathcal{A}^{\mathcal{G}}_{K,k}$.

Pg 43 l -10 The sentence "Corollary 5.3 shows ..." should be changed to "Corollary 5.3 and Theorem 4.4 show ...".

Pg 49 l 10 The term α is missing from the numerator of $\frac{4c}{k-1}$.

Pg 51 l 4,8 Bracket missing after Y (should be two close brackets).

Pg 51 l -7 The sentence begining with "Let f be ..." should be removed.

Pg 52 l 10 This line should be changed to

$$= \inf_{g\in\mathcal{G}}\int_{X\times Y}(2wg(x)/(i+1)+(1-2/(i+1))f_{i-1}(x)-y)^2 dP(x,y)+\epsilon_i/2.$$

Pg 56 l -3 The following sentence should be added: "Let f be the target function and let $d_f = \inf_{g'\in co(\mathcal{G})}||g'-f||$, where $co(\mathcal{G})$ is the convex hull of $\mathcal{G}$.".

Pg 57 l 15 The sentence begining with "We will use ..." should be changed to "We will use the agnostic PAC learning algorithm to learn h under a modified distribution with confidence $1-\delta/2k$ and an accuracy which will be determined below."

Pg 59 l -6 The sentence "Let $\mathcal{G} = \bigcup_{n=1}^{\infty}$ where ..." should be changed to "Let $\mathcal{G} = \bigcup_{n=1}^{\infty} \mathcal{G}_n$ where ...".

Pg 80 l -6 "$\mathcal{F}_k = \mathcal{G}$" should be changed to "$\mathcal{F}_k = \mathcal{F}$".

Pg 81 Eqn A.1 In the third line, "$d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(g_f^2)), E(g_f^2))$" should be changed to "$d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\boldsymbol{z}}(g_f^2), E(g_f^2))$".

Pg 88 l -2,-1 "$\hat{\mathbf{E}}[f_a(x) - f(x))^2$..." should be changed to "$\hat{\mathbf{E}}[(f_a(x) - f(x))^2$..." .

Pg 90 l -10 The sentence begining with "The following result..." should be changed to "Corollary A.15 ...".

Pg 99 l -3 "exp$(\epsilon^2 k/1024(2r)^{2n} M^2)$" should be changed to "exp$(-\epsilon^2 k/1024(2r)^{2n} M^2)$".