On Efficient Agnostic Learning of Linear Combinations of Basis Functions

Wee Sun Lee
Dept. of Systems Engineering,
RSISE, Aust. National University,
Canberra, ACT 0200, Australia.
WeeSun.Lee@anu.edu.au

Peter L. Bartlett
Dept. of Systems Engineering,
RSISE, Aust. National University,
Canberra, ACT 0200, Australia.
Peter.Bartlett@anu.edu.au

Robert C. Williamson
Department of Engineering,
Australian National University,
Canberra, ACT 0200, Australia.
Bob.Williamson@anu.edu.au

Abstract

We consider efficient agnostic learning of linear combinations of basis functions when the sum of absolute values of the weights of the linear combinations is bounded. With the quadratic loss function, we show that the class of linear combinations of a set of basis functions is efficiently agnostically learnable if and only if the class of basis functions is efficiently agnostically learnable. We also show that the sample complexity for learning the linear combinations grows polynomially if and only if a combinatorial property of the class of basis functions, called the fat-shattering function, grows at most polynomially. We also relate the problem to agnostic learning of $\{0,1\}$-valued function classes by showing that if a class of $\{0,1\}$-valued functions is efficiently agnostically learnable (using the same function class) with the discrete loss function, then the class of linear combinations of functions from the class is efficiently agnostically learnable with the quadratic loss function.

1 Introduction

In this paper, we study efficient (polynomial-time) learning of linear combinations of basis functions in a robust extension of the popular Probably Approximately Correct (PAC) learning model in computational learning theory [4]. The learning model we use, commonly called agnostic learning [9, 15, 18] is robust with respect to noise and mismatches between the model and the phenomenon being modelled. The only assumption made about the phenomenon is that it can be represented by a joint probability distribution on $X \times Y$ where X is the domain and Y is a bounded subset of $\mathbb{R}$. This model more adequately captures many of the features of practical learning problems where measurements are often noisy and very little is known about the target functions. Under these assumptions, we cannot expect our learning algorithm to always produce a hypothesis with small error. Instead, we demand that with high probability, the hypothesis produced is close to optimal in the class of functions we are using. The error of a hypothesis on an observation is measured using a loss function and we demand that the expected loss of the hypothesis is close to the smallest expected loss of functions in the class. The generality of the agnostic learning model allows, as special cases, learning real-valued functions, learning the conditional expectation when the observations are noisy, learning the best approximation when the target function is not in the class and also learning probabilistic concepts [14].

The function classes we study include some which are widely used in practice. Linear combinations of basis functions can be considered as a generalization of two layer neural networks. Instead of the usual sigmoidal hidden units, arbitrary bounded function classes satisfying mild measurability constraints are allowed to be basis function classes. This includes radial basis functions and polynomial basis functions (with some restrictions on the inputs). We do not bound the number of basis functions in a linear combination but instead insist that the sum of absolute values of the weights of the linear combination be bounded. This work is an extension of results in [17] where it was shown that the class of linear combinations of linear threshold functions with bounded fan-in is efficiently agnostically learnable. Related works include that of Koiran [16] which considered learning two layer neural networks with piecewise linear activation functions (but not in the agnostic model) and that of Maass [18] on agnostic learning of fixed sized multilayer neural networks with piecewise polynomial activation functions.

We say that a function class $F \subset [-B,B]^X$ is efficiently agnostically learnable if there exists an hypothesis class $H \subset [-B,B]^X$, and a learning algorithm which produces an hypothesis h from H such that with probability at least $1-\delta$, the expected loss of h is no more than ϵ away from the expected loss of the best function in F and the algorithm runs in time polynomial in $1/\epsilon$, $1/\delta$, B, T (a bound on the observation range) and the appropriate complexity parameters. Commonly used complexity parameters include the dimension of input space and the number of parameters parametrizing the function class F. The hypothesis class H does not necessarily have to be the same as the function class F. This allows agnostic learning of a function class using a larger hypothesis class, a situation which has been shown to have computational advantages in certain cases for learning in the PAC framework.

In Section 3, we show that with the quadratic loss function, linear combinations of basis functions with bounded sum of absolute values of weights are efficiently agnostically learnable if and only if the basis function class is efficiently agnostically learnable. This means that to show that the class of linear combinations is efficiently agnostically learnable, we need only to show that the basis function class is efficiently agnostically learnable. Efficiently agnostically learnable classes of basis functions include function classes which can be enumerated in time polynomial in the sample size and complexity parameters, and fixed sized neural networks with piecewise polynomial activation functions [18]. Similarly, a (representation independent) hardness result for learning the basis function class would imply a hardness result for the class of linear combinations of functions from the basis function class.

In Section 4, we use a combinatorial property of the basis function class called the fat-shattering function [14, 2, 6] to bound the covering number of the linear combinations, and provide sample complexity bounds (which are better than those obtained using the results in Section 3). Gurvits and Koiran [8] have also provided sample complexity bounds for the case when the basis functions are $\{0,1\}$-valued functions by bounding the fat-shattering function of the convex closure of the basis function class. We also use lower bound results from [6] to show that the class of linear combinations of basis functions has sample complexity which grows polynomially with $1/\epsilon$ and the complexity parameters if and only if the fat-shattering function of the class of basis functions grows at most polynomially with $1/\epsilon$ and the complexity parameters.

In Section 5, we give some examples of function classes which are efficiently agnostically learnable. Section 6 shows that if a class F of $\{0,1\}$-valued basis functions can be learned efficiently and agnostically with $\{0,1\}$-valued targets, the discrete loss function and F as the hypothesis class, then the class of linear combinations of the basis functions can be efficiently agnostically learned with the quadratic loss function. This relates efficient agnostic learning of linear combinations of $\{0,1\}$-valued basis functions to the work on agnostic learning of $\{0,1\}$-valued functions. Unfortunately, most of the results on agnostic learning of $\{0,1\}$-valued functions are negative [15, 11]. In Section 7, we discuss the hardness results as well some open problems.

2 Definitions and Learning Model

2.1 Agnostic Learning model

Our agnostic learning model is based on the agnostic learning model described by Kearns, Schapire and Sellie [15]. Let X be a set called the *domain* and the points in X be called *instances*. Let Y be the *observed range*. We call the pair $(x,y) \in X \times Y$ an observation. The *assumption class* $\mathcal{A}$ is a class of probability distributions on $X \times Y$ and is used to represent the assumptions about the phenomenon that is being learned. The results in this paper hold when $\mathcal{A}$ is the class of all probability distributions on $X \times Y$. The following are special cases, for which our results automatically hold.

In *regression*, $y \in Y$ represents a noisy measurement of some real valued quantity, and the desired quantity to be learned is the conditional expectation of y given x. In the case where there is no noise, we have *function learning* where there is a class F of functions mapping X to Y and $A_{D,f} \in \mathcal{A}$ is formed from a distribution D over X and a function $f \in F$. Observations drawn from $A_{D,f}$ have the form $(x, f(x))$ where x is drawn randomly according to D. In learning *probabilistic concepts* [14], we have $Y = \{0,1\}$. Again, there is a class F of functions, mapping X to $[0,1]$ and $A_{D,f} \in \mathcal{A}$ is formed from a distribution D over X and a function $f \in F$. Observations drawn from $A_{D,f}$ are of the form (x,b), where x is drawn randomly according to D and $b = 1$ with probability $f(x)$ and $b = 0$ with probability $1 - f(x)$.

Let Y' be a bounded subset of $\mathbb{R}$. The *touchstone class* $\mathcal{T}$ and the *hypothesis class* $\mathcal{H}$ are two classes of functions from X to Y'. The learning algorithm will attempt to model the behaviour from $\mathcal{A}$ with functions from $\mathcal{H}$. Since the behaviour from $\mathcal{A}$ cannot neccessarily be well approximated by functions from $\mathcal{H}$, we require a method of judging whether the hypothesis is acceptable. This is done by requiring that learning algorithm produces a hypothesis $h \in \mathcal{H}$ with performance close to the "best" $t \in \mathcal{T}$. The touchstone class $\mathcal{T}$ is introduced for computational reasons. It is often the case that although we are unable to find an efficient algorithm for learning a touchstone class, we can find an efficient algorithm for learning a larger hypothesis class which contains the touchstone class. The resulting hypothesis from the learning algorithm may not belong to the touchstone class but its performance will at least be close to the performance of the best function from the touchstone class.

For the "best" function $t \in \mathcal{T}$, we use the function which minimizes the expected value of a *loss function* $L : Y' \times Y \to \mathbb{R}^+$. Given a function h, the loss of h on (x,y) is $L_h(x,y) = L(h(x),y)$. The main loss function used in this paper is the *quadratic loss function* $Q(y',y) = (y'-y)^2$ but the *discrete loss function* $Z(y',y) = 0$ if $y' = y$ and $Z(y',y) = 1$ if $y' \neq y$, is also used. Given observations drawn according to $A \in \mathcal{A}$, the *expected loss* is $\mathbf{E}_{(x,y)\in A}[L_h(x,y)]$ which we denote $\mathbf{E}[L_h]$ when A is clear from the context. For a class $\mathcal{T}$, we define *opt*$(\mathcal{T}) = \inf_{h\in\mathcal{H}} \mathbf{E}[L_h]$. The quadratic loss function is a natural choice to use for regression because the function with the minimum quadratic loss is the conditional expectation. For learning probabilistic concepts, the quadratic loss has some desirable properties as shown in [14] and [15].

The results in this paper hold in both the uniform cost and logarithmic cost models of computation (see [1]). In the uniform cost model, real numbers occupy one unit of space and standard arithmetic operations (addition, multiplication etc.) take one unit of time. In the logarithmic cost model, real numbers are represented in finite precision and operations on them are charged time proportional to the number of bits of precision. We assume that the observation range $Y = [-T, T]$ is known and the hypothesis and touchstone classes have outputs in the range $Y' = [-B, B]$. Hence, the learning problem can be indexed by B and T with $\mathcal{T} = \bigcup_{B \in \mathbb{R}^+} \mathcal{T}_B$ and $\mathcal{H} = \bigcup_{B \in \mathbb{R}^+} \mathcal{H}_B$. In the logarithmic cost model of computation, we also index the problem by r and s where $\mathcal{A} = \bigcup_{r \in \mathbb{N}} \bigcup_{T \in \mathbb{R}} \mathcal{A}_{rT}$ with the bit length of any observation drawn from $A \in \mathcal{A}_{rT}$ bounded by a polynomial in r and $\mathcal{T} = \bigcup_{s \in \mathbb{N}} \bigcup_{B \in \mathbb{R}^+} \mathcal{T}_{Bs}$ where any $f \in \mathcal{T}_{Bs}$ has a representation whose bit length is bounded by a polynomial in s. The domain X, distribution class $\mathcal{A}$ and function class $\mathcal{T}$ are also often indexed by some other natural *complexity parameters* such as the dimension of the input space or the number of parameters parametrizing the function class.

Definition 1 *A class of functions $\mathcal{T} = \bigcup_{B \in \mathbb{R}^+} \mathcal{T}_B$ is* efficiently agnostically learnable *(with respect to loss function L) if there exists a function class $\mathcal{H} = \bigcup_{B \in \mathbb{R}^+} \mathcal{H}_B$, a function $m(\epsilon, \delta, T, B)$ bounded by a fixed polynomial in $1/\epsilon$, $1/\delta$, B and T and an algorithm such that for any $A \in \mathcal{A}$, given any $0 < \delta \leq 1$, $\epsilon > 0$, $T > 0$ and $B > 0$, the algorithm draws $m(\epsilon, \delta, T, B)$ observations, halts in time bounded by another fixed polynomial in $1/\epsilon$, $1/\delta$, T and B and outputs a hypothesis $h \in \mathcal{H}_B$ such that with probability at least $1 - \delta$, $\mathbf{E}[L_h] \leq opt(\mathcal{T}) + \epsilon$. The hypothesis h must also be evaluable in time polynomial in $1/\epsilon$, $1/\delta$, T and B. If $\mathcal{H} = \mathcal{T}$, then we say $\mathcal{T}$ is* properly efficiently agnostically learnable.

In the logarithmic cost model, we let $\mathcal{T} = \bigcup_{s \in \mathbb{N}} \bigcup_{B \in \mathbb{R}^+} \mathcal{T}_{Bs}$, $\mathcal{A} = \bigcup_{r \in \mathbb{N}} \bigcup_{T \in \mathbb{R}} \mathcal{A}_{rT}$, allow m to depend on ϵ, δ, T, B, s and r, and insist that the algorithm halt and the hypothesis be evaluable in time polynomial in $1/\epsilon$, $1/\delta$, T, B, s and r. For proper efficient learning, when the touchstone class is $\mathcal{T}_{Bs}$, we only insist that the hypothesis be from $\bigcup_{s' \in \mathbb{N}} \mathcal{T}_{Bs'}$ with the number of bits allowed to grow polynomially with $1/\epsilon$.

If the learning problem is indexed by other complexity parameters, we also insist that m, the running time of the algorithm and the evaluation time of the hypothesis be polynomial in those parameters.

For simplicity we work in the uniform cost model.

2.2 Basis Functions and Linear Combinations

Definition 2 *A class of real-valued functions $\mathcal{G}$ is an* admissible *class of basis functions if $\mathcal{G}$ is* permissible[1] *and there exists $b_{\mathcal{G}} > 0$ such that $|g(x)| \leq b_{\mathcal{G}}$ for all $g \in \mathcal{G}$, $x \in X$.*

Definition 3 *Let $\mathcal{G}$ be an admissible class of basis functions. Then for every $K > 0$,
$\mathcal{N}_K^{\mathcal{G}} = \bigcup_{k=1}^{\infty} \left\{ \sum_{i=1}^{k} w_i g_i : g_i \in \mathcal{G}, \sum_{i=1}^{k} |w_i| \leq K \right\}$, is the class of* linear combinations *of functions from $\mathcal{G}$ with the sum of magnitudes of weights bounded by K.*

3 Equivalence in Efficient Learning

In this section we show that the class of bounded linear combinations of functions from an admissible class of basis functions is efficiently agnostically learnable if and only if the class of basis functions is efficiently agnostically learnable. We use the agnostic learning algorithm of the basis function class as a subroutine to learn the linear combinations of basis functions with nothing assumed about the hypothesis class except the bound on the range of its output. We will need an approximation result from [17]. This result is an extension of results by Jones [13] and Barron [5]. A related result is also presented by Koiran [16].

Theorem 4 *Let H be a Hilbert space with norm $\| \cdot \|$. Let G be a subset of H with $\| g \| \leq b$ for each $g \in G$. Let $co(G)$ be the convex hull of G. For any $f \in H$, let $d_f = \inf_{g' \in co(G)} \| g' - f \|$. Suppose that $f_0 = 0$ and iteratively, f_k is chosen to satisfy $\| f_k - f \|^2 \leq \inf_{g \in G} \| \alpha_k f_{k-1} + (1 - \alpha_k) g - f \|^2 + \epsilon_k$, where $\alpha_k = 1 - 1/k$, $c \geq b^2$, and $\epsilon_k \leq \frac{c - b^2}{k^2}$. Then for any $\beta \in (0, 1)$, any $K_\beta \geq \frac{1}{1-\beta}$ and any $k \geq 1$,*

$$\| f - f_k \|^2 - d_f^2 \leq \frac{c K_\beta}{k^\beta}.$$

Let H be the Hilbert space of measurable functions on X with inner product $(f, g) = \int_X f(x) g(x) dP(x)$ where P is some probability distribution on X. Let $\mathcal{G}$ be an admissible class of basis functions and let $\mathcal{N}_K^{\mathcal{G}}$ be the class of linear combinations of functions derived from $\mathcal{G}$; that is $\mathcal{N}_K^{\mathcal{G}}$ is the convex hull of $G = \{wg : |w| = K, g \in \mathcal{G}\}$. For all $g \in G$, $\| g \| < K b_{\mathcal{G}}$, so Theorem 4 can be used to obtain an approximation to the best function in $\mathcal{N}_K^{\mathcal{G}}$.

The above approximation result requires that the target be a function in the same Hilbert space. In agnostic learning, the target can be a random variable. However, with quadratic loss, minimizing the loss with respect to the observations is the same as minimizing the loss with respect to the conditional expectation, which is a function in the Hilbert space when the target is bounded.

Theorem 5 *Let $\mathcal{G}$ be an admissible class of basis functions, and let $K > 0$. Then $\mathcal{N}_K^{\mathcal{G}}$ is efficiently agnostically learnable with respect to the quadratic loss function if and only if $\mathcal{G}$ is efficiently agnostically learnable with respect to the quadratic loss function.*

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 $G = \{wg : |w| = K, g \in \mathcal{G}\}$ and for all $g \in G$, $\| g \| < K b_{\mathcal{G}}$. Theorem 4 shows that to get within ϵ of the best expected loss, a number of iterations equal to $k = (c K_\beta / \epsilon)^{1/\beta}$ will do. Set $c = 2K^2 b_{\mathcal{G}}^2$ and $\epsilon_i = K^2 b_{\mathcal{G}}^2 / i^2$ for $1 \leq i \leq k$. Assume that the agnostic algorithm for learning $\mathcal{G}$ produces an hypothesis from $\mathcal{H} \subset [-b_{\mathcal{G}}, b_{\mathcal{G}}]^X$. Let f be the target function (conditional expectation of target y given input x). Minimizing the quadratic loss with respect to the

[1] This is a mild measurability condition satisfied by most function classes used for learning. See Haussler [9] for details.

joint distribution is equivalent to minimizing the quadratic loss with respect to the conditional expectation. To satisfy Theorem 4, at the ith iteration, we must find $h_i \in \mathcal{H}$ such that $\int_{X\times Y}(h_i(x)/i + (1-1/i)f_{i-1}(x) - y)^2 dP(x,y) \leq \inf_{g\in G}\int_{X\times Y}(g(x)/i + (1-1/i)f_{i-1}(x) - y)^2 dP(x,y) + \epsilon_i$, where f_{i-1} is the linear combination which has been found so far. Furthermore

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

We now use the agnostic learning algorithm for $\mathcal{G}$ with respect to the new target random variable which has magnitude bounded by $i(Kb_{\mathcal{G}}+T)/K$ (where T is an upper bound on the magnitude of y). Set confidence to $\delta/2k$ and accuracy to $i^2\epsilon_i/2K^2$. So with probability at least $1-\delta/2k$, the hypothesis h_i produced is such that

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

This has to be done for both $w = K$ and $w = -K$. So at each iteration, we produce two hypotheses from which we have to choose one. If we have no other way of choosing between the two hypotheses, we have to do hypothesis testing. Using Hoeffding's inequality [10], a sample size of $8(Kb_{\mathcal{G}}+T)^4(\ln 2 + \ln\frac{4k}{\delta})/\epsilon_i^2$ is large enough so that the empirical quadratic loss is no more than $\epsilon_i/4$ from the expected quadratic loss for both functions with probability at least $1-\delta/2k$. If we choose the hypothesis which has the smaller empirical loss, the expected loss will be no more that $\epsilon_i/2$ away from the expected loss of the better hypothesis with probability $1-\delta/2k$.

At each iteration, given an efficient agnostic learning algorithm for learning $\mathcal{G}$, we produce an hypothesis which satisfies the requirements of Theorem 4 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 check that if the algorithm for learning $\mathcal{G}$ is polynomial in the relevant parameters, the resulting algorithm for learning $\mathcal{N}_K^{\mathcal{G}}$ will be polynomial in the desired parameters. □

4 Sample Complexity and Proper Efficient Agnostic Learning

We can obtain sample complexity bounds for learning problems if certain combinatorial properties of the class of basis functions such as the pseudo-dimension [9] or the fat-shattering function [14, 6] are known. The pseudo-dimension has been used to obtain sample complexity bounds for learning function classes (and in particular multilayer neural networks) by Haussler [9]. However, finiteness of the pseudo-dimension is not a necessary condition for agnostic learning. For example, the class of non-decreasing functions that map from $[0,1]$ to $[0,1]$ can be shown to be learnable even though it has infinite pseudo-dimension.

A sequence of points $x_1,\ldots,x_d$ from X is γ-shattered by $F \subset [-B,B]^X$ if there exists $r \in [-B,B]^d$ such that for each $b \in \{0,1\}^d$, there is an $f \in F$ such that for each i, $f(x_i) \geq r_i + \gamma$ if $b_i = 1$ and $f(x_i) \leq r_i - \gamma$ if $b_i = 0$. For each γ, let $\text{fat}_F(\gamma) = \max\{d \in \mathbb{N} : \exists x_1,\ldots,x_d,\ F\ \gamma\text{-shatters } x_1,\ldots,x_d\}$ if such a maximum exists and ∞ otherwise.

In this section, we give sample complexity bounds based on the fat-shattering function of the class of basis functions and show that $\mathcal{N}_K^{\mathcal{G}}$ is agnostically learnable with polynomial sample complexity (disregarding computational complexity) if and only if the fat-shattering function of $\mathcal{G}$ grows polynomially with $1/\epsilon$ and the complexity parameters. Using these bounds, we show how an efficient algorithm for proper agnostic learning of linear combinations of basis functions can be constructed from a proper efficient agnostic algorithm for learning the basis functions. The algorithm in this section is significantly different from the algorithm of the previous section in that the sample size is determined and drawn only once instead of being drawn at each addition of a new basis function to the linear combination. The sample complexity obtained is approximately $O(\text{fat}_{\mathcal{G}}(\epsilon)/\epsilon^3)$, ignoring β and log factors. In comparison, if the method in Section 3 is used for proper learning, we obtain a sample size bound of $O(\text{fat}_{\mathcal{G}}(\epsilon)/\epsilon^3 + 1/\epsilon^5)$, ignoring β and log factors.

In [6], 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.

The following is essentially from [6] with minor modifications.

Theorem 6 *Let F be a class of $[0,1]$-valued functions defined on X. Suppose $0 < \gamma < 1$, $0 < \epsilon \leq \gamma/65$, $0 \leq \delta \leq 1/16$ and $d \in \mathbb{N}$. If $\text{fat}_F(\gamma) \geq d > 800$, then no algorithm can agnostically learn F with respect to the quadratic loss function to accuracy $3\epsilon^2$ with probability $1-\delta$ using fewer than $m > \frac{d}{400\log\frac{40}{\gamma}}$ examples.*

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

4.1 Bounding Covering Numbers

We now want to bound the sample complexity for learning $\mathcal{N}_K^{\mathcal{G}}$ in terms of the fat-shattering function for $\mathcal{G}$. Using Theorem 4, we can bound the number of basis functions needed in the linear combination. Hence, we only need to be able to obtain upper bounds on the sample complexity of linear combinations of a fixed but arbitrary number of basis functions.

For $n \in \mathbb{N}$, $v, w \in \mathbb{R}^n$, let $d_{l^1}(v,w) = \frac{1}{n}\sum_{i=1}^{n} |v_i - w_i|$.

For a set S with a metric (or pseudo-metric) ρ, an ϵ-cover is a finite set $U \subseteq S$ such that for all $x \in S$, there is a $y \in U$ with $\rho(x,y) \leq \epsilon$. The covering number $N(\epsilon, S, \rho)$ denotes the size of the smallest ϵ-cover for S.

If Z is a set, $h\colon Z \to \mathbb{R}$ and $z \in Z^m$, let $h_{|z} \in \mathbb{R}^m$ denote $(h(z_1), \ldots, h(z_m))$. If H is a set of functions from Z to $\mathbb{R}$, write $H_{|z} = \{h_{|z}\colon h \in H\}$. Given a loss function $L : Y' \times Y \to [0, C]$, define $L_H := \{L_h\colon h \in H\}$. Define $\hat{\mathbf{E}}_z[L_h] = \frac{1}{m}\sum_{i=1}^{m} L_h(z_i)$. We write $\hat{\mathbf{E}}[L_h] = \hat{\mathbf{E}}_z[L_h]$ when the meaning is clear from the context.

Theorem 7 *(Haussler [9])*

Suppose H is a class of functions mapping from X to Y' and L is a loss function $L\colon Y' \times Y \to [0, C]$ such that L_H is permissible. Let P be any probability distribution on $Z = X \times Y$. For $m \geq 1$ and any $0 < \epsilon \leq C$,

$$P^m\left\{z \in Z^m\colon \exists h \in H, \left|\hat{\mathbf{E}}[L_h] - \mathbf{E}[L_h]\right| > \epsilon\right\}$$
$$\leq 4\mathbf{E}(N(\epsilon/16, L_{H_{|z}}, d_{l^1})e^{-\epsilon^2 m/64C^2},$$

where the expectation is over z drawn randomly from Z^{2m} according to P^{2m}.

Let F be a class of functions from X to $[0, C]$ and let P be a probability distribution on X. Let $d_{L^1(P)}$ be the pseudo-metric on F defined by $d_{L^1(P)}(f,g) = \mathbf{E}(|f-g|) = \int_X |f(x) - g(x)| dP(x)$ for all $f, g \in F$.

We bound $N(\epsilon, F, d_{L^1(P)})$ for all P in terms of the fat-shattering function. This provides a bound on $N(\epsilon, F_{|x}, d_{l^1})$ for any finite sequence of points x (via the isometry between $(F_{|x}, d_{l^1})$ and $(F, d_{L^1(P_{|x})})$, where $P_{|x}$ is the empirical distribution on x). We use a generalization of Sauer's lemma by Alon *et al.* [2] and techniques by Haussler [9].

Theorem 8 *Let $G = \{wg\colon |w| = K, g \in \mathcal{G}\}$, let P be a probability distribution on X and let $\mathcal{H}_k = \{\sum_{i=1}^{k} a_i g_i : g_i \in G\}$ where $a_i \geq 0$ are fixed for $1 \leq i \leq k$ and $\sum_{i=1}^{k} a_i =$ 1. Then,*

$$N(\epsilon, \mathcal{H}_k, d_{L^1}(P)) \leq 2^k \exp\left(\frac{8k\mathrm{fat}_{\mathcal{G}}(\epsilon/(8Kb_{\mathcal{G}}))}{\ln 2}\right.$$
$$\left.\ln^2\left(\frac{2048K^4 b_{\mathcal{G}}^4 \mathrm{fat}_{\mathcal{G}}(\epsilon/(8Kb_{\mathcal{G}}))}{\epsilon^4 \ln 2}\right)\right).$$

The proof is omitted from this abstract.

To use Theorem 7, we need to bound the covering number of the L_H (here we use the quadratic loss function $L = Q$). To do that we bound the covering number of the loss function class Q_H in terms of the covering number of H using the following lemma from [6].

Lemma 9 *([6]) Let F be a class of functions from X to $Y' \subseteq [-B, B]$. Let $Y \subseteq [-T, T]$ and so $Q(Y, Y') \subseteq [0, C]$, where $C = (B+T)^2$. Let $x \in X^m$ and $z \in (X \times Y)^m$. Then $N(\epsilon, Q_{F_{|z}}, d_{l^1}) \leq N\left(\frac{\epsilon}{3\sqrt{C}}, F_{|x}, d_{l^1}\right)$ where $z \in (X \times Y)^m$ and $x \in X^m$.*

4.2 The Learning Algorithm

We now give a relationship between the learning problem and an optimization problem on a training sample. The method used is similar to that used in [17] and is based on the technique used by Haussler [9].

For $S = ((x_1, y_1), \ldots, (x_m, y_m))$ and a function class F, let $\widehat{opt}_S(F) = \inf_{f \in F} \frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2$.

Lemma 10 *Let $0 < \epsilon$, $0 < \delta < 1$, let F and H be function classes such that $F \subseteq H$ and let the sample size $m(\epsilon, \delta)$ be such that for any probability distribution P on $X \times Y'$, $P^m\left\{\exists f \in F\colon \left|\hat{\mathbf{E}}[L_f] - \mathbf{E}[L_f]\right| \geq \epsilon/4\right\} \leq \delta/2$. Suppose $|opt(F) - opt(H)| \leq \epsilon/4$ and we have a randomized algorithm which produces $\hat{f} \in F$ for sample S of size m drawn according to P such that $\Pr\left(\left|\hat{\mathbf{E}}[L_{\hat{f}}] - \widehat{opt}_S(F)\right| \geq \epsilon/4\right) \leq \delta/2$. Then $\Pr\left(\left|\mathbf{E}[L_{\hat{f}}] - opt(H)\right| \geq \epsilon\right) \leq \delta$ where all the probabilities are taken over the random samples and the randomization used by the algorithm.*

Proof Sketch. With probability greater than $1 - \delta/2$, we have simultaneously, $\left|\hat{\mathbf{E}}[L_{\hat{f}}] - \mathbf{E}[L_{\hat{f}}]\right| \leq \epsilon/4$ and $\left|\widehat{opt}_S(F) - opt(F)\right| \leq \epsilon/4$. The desired result is obtained using the triangle inequality. □

We now show how an algorithm for properly agnostically learning $\mathcal{G}$ can be used to obtain a randomized algorithm for optimization using $\mathcal{N}_K^{\mathcal{G}}$ as required in Lemma 10.

Theorem 11 *Let $\mathcal{G}$ be an admissible function class. Then $\mathcal{N}_K^{\mathcal{G}}$ is properly efficiently agnostically learnable if $\mathcal{G}$ is properly efficiently agnostically learnable. Furthermore the sample complexity for properly efficiently learning $\mathcal{N}_K^{\mathcal{G}}$ is at most*

$$\frac{1024C^2}{\epsilon^2}\left(\frac{8kd}{\ln 2}\left(30 + \ln\left(\frac{C^2K^4b_{\mathcal{G}}^4 d}{\epsilon^4}\right)\right)^2 + k\ln 2 + \ln\frac{8}{\delta}\right)$$

where $k = (8K_\beta K^2 b_{\mathcal{G}}^2)^{1/\beta}/\epsilon^{1/\beta}$, $C = (Kb_{\mathcal{G}} + T)^2$ and $d = \mathrm{fat}_{\mathcal{G}}(\epsilon/(1536\sqrt{C}Kb_{\mathcal{G}}))$.

Proof. The aim of the proof is to set up the conditions such that Lemma 10 holds and to show that this can be done efficiently. Let $G = \{wg \colon |w| = K, g \in \mathcal{G}\}$. In Theorem 4, note that for each k there is a fixed sequence of weights $(a_1, \ldots, a_k)$ with $a_i \geq 0$ for each $i = 1, \ldots, k$ and $\sum_{i=1}^{k} a_i = 1$ such that an approximation rate of cK_β / k^β is achieved. Let $\mathcal{H}_k = \{\sum_{i=1}^{k} a_i g_i : g_i \in G\}$. The range of $\mathcal{H}_k$ is $Y' \subseteq [-b_{\mathcal{G}}K, b_{\mathcal{G}}K]$. Let the range of the observation be $Y \subseteq [-T, T]$ and let $Q(Y, Y') \subseteq [0, C]$, where $\sqrt{C} = b_{\mathcal{G}}K + T$. From a set S of m independently sampled points in $X \times Y$, form an empirical distribution D by weighting each member of S in proportion to the number of times the point appears in S.

Using Theorem 4 we can find the number k such that $|opt(\mathcal{N}) - opt(\mathcal{H}_k)| \leq \epsilon/4$. Using the empirical distribution, D, for the same k there also exists an $\hat{f} \in \mathcal{H}_k$ such that $\left|\hat{\mathbf{E}}[L_{\hat{f}}] - \widehat{opt}_S(\mathcal{N})\right| \leq \epsilon/4$. Since $\widehat{opt}_S(\mathcal{H}_k) \geq \widehat{opt}_S(\mathcal{N})$, we have $\left|\hat{\mathbf{E}}[L_{\hat{f}}] - \widehat{opt}_S(\mathcal{H}_k)\right| \leq \epsilon/4$. From Theorem 4, to obtain an approximation to accuracy $\epsilon/4$, $k = (4K_\beta c)^{1/\beta}/\epsilon^{1/\beta}$ suffice.

Let $c = 2K^2 b_{\mathcal{G}}^2$. From Theorem 4, Theorem 8 and Lemma 9 with $k = (8K_\beta K^2 b_{\mathcal{G}}^2)^{1/\beta}/\epsilon^{1/\beta}$, for any finite sample size,

$$N(\epsilon/64, Q_{\mathcal{H}_k}, d_{l^1}) \quad \leq \quad 2^k \exp\left(\frac{8kd}{\ln 2}\left(30 + \ln\left(\frac{C^2K^4b_{\mathcal{G}}^4 d}{\epsilon^4}\right)\right)^2\right)$$

where $d = \mathrm{fat}_{\mathcal{G}}(\epsilon/(1536\sqrt{C}Kb_{\mathcal{G}}))$.

Thus using Theorem 7, for samples z drawn according to distribution P^m,

$$P^m\left\{z \in Z^m \colon \exists h \in \mathcal{H}_k, \left|\hat{\mathbf{E}}[Q_h] - \mathbf{E}[Q_h]\right| > \epsilon/4\right\} \leq 2^k 4 \exp\left(\frac{8kd}{\ln 2}\left(30 + \ln\left(\frac{C^2K^4b_{\mathcal{G}}^4 d}{\epsilon^4}\right)\right)^2\right) e^{(-\epsilon^2 m/1024C^2)}.$$

Setting the right hand side to $\delta/2$ to satisfy Lemma 10, we obtain the sample complexity bound. Since $\mathcal{G}$ is properly efficiently agnostically learnable, Theorem 6 implies that the fat-shattering function of $\mathcal{G}$ is bounded by a polynomial in $1/\epsilon$ and the complexity parameters, so the sample size bound for learning $\mathcal{N}_K^{\mathcal{G}}$ is polynomial in $1/\epsilon$, $1/\delta$ and the complexity parameters.

Finally we need to 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 $\mathcal{H}_k$. 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 4) with probability greater than $1 - \delta/2k$. Unlike in the proof of Theorem 5, we can test the hypotheses directly using the same sample. Knowing the fat-shattering function of $\mathcal{G}$ enables us to bound the size of the sample required to be sampled according to the empirical distribution. Note that since we are sampling from the empirical distribution, no new observations need to be drawn from the original distribution. Theorem 4 assures us that if we are successful at each iteration, we will be within $\epsilon/4$ of the minimum error on the sample as required in Lemma 10. □

5 Efficiently Learnable Function Classes

In this section, we give some examples of efficiently agnostically learnable function classes.

5.1 Efficiently Enumerable Function Classes

In [17], it was shown that a linear combination of linear threshold units with a bounded sum of magnitudes of weights is efficiently agnostically learnable if the fan-ins of the linear threshold units are bounded. The complexity parameter in this case is the dimension of the input space and the basis function class is the class of linear threshold units with bounded fan-in. Similarly, in fixed dimension or with bounded fan-in, a linear combination of axis parallel rectangles with bounded sum of magnitudes of weights is efficiently agnostically learnable. These results are generalized in the following corollary which is particularly useful for $\{0, 1\}$-valued basis function classes.

Corollary 12 *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 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. Since the number of functions in $\mathcal{G}_{|x}$ is polynomial in m and the complexity parameter, the fat-shattering function must be bounded by a logarithmic function of 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 $\mathcal{G}$. □

5.2 Neural Networks with Piecewise Polynomial Activation Functions

Let $x \in \mathbb{R}^n$ where n is fixed. Let $\mathcal{G} = \{x \mapsto \phi(w \cdot x - \theta)\}$ where the magnitudes of the threshold θ and each component of the weight vector w are less than W, $\phi(\nu) = 0$ for $\nu \leq 0$, $\phi(\nu) = \nu$ for $0 \leq \nu \leq 1$ and $\phi(\nu) = 1$ for $\nu > 1$. It is not possible to enumerate $\mathcal{G}_{|x}$ because the number of possible outputs is not finite. However, as shown by Koiran [16], it is possible to enumerate all possible combinations of inputs with the linear pieces of the activation function. With a proper parametrization, the optimization problem can be solved by solving a family of quadratic programming problems. Hence, in fixed dimension, a linear combination of functions from $\mathcal{G}$ with bounded sum of magnitudes of weights is properly efficiently agnostically learnable.

Maass [18] has shown that a fixed architecture neural network with an arbitrary number of hidden layers and piecewise polynomial activation functions is efficiently agnostically learnable (with respect to the absolute loss function $\Lambda(y, y') = |y - y'|$ but the result also holds for the quadratic loss function). Hence, the class of linear combinations of such networks with bounded sum of magnitudes of weights is also efficiently agnostically learnable. Note that Maass used a larger hypothesis class to learn this class. Hence the function class has not been shown to be *properly* efficiently agnostically learnable.

6 Relationship with Agnostic PAC learning

Let $\mathcal{G}$ be a class of $\{0,1\}$-valued functions. Let the target function be chosen from the class of all $\{0,1\}$-valued functions on X. Following Kearns, Schapire and Sellie [15], we call proper efficient agnostic learning with discrete loss under these assumptions *agnostic PAC learning*. In this section, we show that if $\mathcal{G}$ is agnostically PAC learnable, then $\mathcal{N}_K^{\mathcal{G}}$ is properly efficiently agnostically learnable.

As shown by Jones [13], the iterative approximation result holds even if the inner product of the basis function with $f_k - f$ (where f is the target function 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 [16]. We use this property and transform the problem of minimizing the inner product on a finite set of examples into the problem of agnostic PAC learning.

The following theorem follows from the proof of Theorem 1 given in [16] with minor changes.

Theorem 13 *Let G be a subset of a Hilbert space H with $\| g \| \leq b$ for each $g \in G$. Let $co(G)$ be the convex hull of G. For any $f \in H$, let $d_f = \inf_{g' \in co(G)} \| g' - f \|$. 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' is chosen to satisfy $(f_{k-1} - f, g') \leq \inf_{g \in G}(f_{k-1} - f, g) + \epsilon_k$ and $\epsilon_k \leq \frac{c^2-(2b+d_f)^2}{k^2}$. Then $\| f - f_k \|^2 - d_f^2 \leq \frac{2cd_f}{\sqrt{k}} + \frac{c^2}{k}$.*

Theorem 14 *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 agnostically PAC learnable.*

Proof Sketch. Again we will set up conditions necessary for Lemma 10. Since the target range is bounded we can easily find a bound for d_f. Using Theorem 13, pick the number of basis functions k in the linear combinations to obtain approximation $\epsilon/4$. Let $G = \{wg \colon |w| = K, g \in \mathcal{G}\}$. In Theorem 13, note that for each k there is a fixed sequence of weights $(a_1, \ldots, a_k)$ with $a_i \geq 0$ for each $i = 1, \ldots, k$ and $\sum_{i=1}^k a_i = 1$ such that the desired approximation rate is achieved. Let $\mathcal{H}_k = \{\sum_{i=1}^k a_i g_i : g_i \in G\}$. Then using the fat-shattering function bound, pick a sample large enough to get the accuracy and confidence required in Lemma 10. If $\mathcal{G}$ is agnostically PAC learnable, then the fat-shattering function (which is the same as the VC-dimension for $\{0,1\}$-valued functions) is polynomial in the complexity parameters [7]. Theorem 13 shows that approximating to accuracy ϵ_i at each iteration with respect to the empirical distribution will provide a hypothesis with the desired error.

For each iteration i, $1 \leq i \leq k$, we first find a function $g \in \mathcal{G}$ which nearly minimizes $(f_{k-1} - f, wg) = \frac{w}{m}\sum_{i=1}^m (f_{k-1}(x_i) - f(x_i))g(x_i)$ for $w = K$ where f is the conditional expectation on the sample under the empirical distribution. It is possible to show that this is equivalent to agnostic PAC learning of the function h where $h(x_i) = 0$ if $f_{k-1}(x_i) - f(x_i) > 0$ and $h(x_i) = 1$ if $f_{k-1}(x_i) - f(x_i) \leq 0$ under the distribution $P(x_i) = \frac{|f_{k-1}(x_i) - f(x_i)|}{s}$ where $s = \sum_{i=1}^m |f_{k-1}(x_i) - f(x_i)|$. Next, a similar procedure is carried out for $w = -K$. The two functions produced are then compared and the better one chosen. □

7 Discussion and Open Problems

An interesting open problem is to find the limits of the complexity of basis functions which allow efficient agnostic learning. For agnostic PAC learning of basis functions, available results include hardness of learning monomials and halfspaces under the assumption $RP \neq NP$ [15, 11]. This implies that for networks of functions from these classes, it is unlikely that an efficient algorithm can be obtained from the approach given in Sections 4 and 6. However to rule out efficient learning with other methods or hypothesis classes requires representation independent hardness results. In [15], 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. Whether polynomial-size DNF can be learned efficiently has been an open problem in computational learning theory since it was first posed by Valiant [20] in 1984 (the majority view is that polynomial-sized DNF is not likely to be efficiently learnable [12]). Using techniques similar to that in [15], 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.

Theorem 15 *Let $\mathcal{G}$ be 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 Sketch. We will show that there exists a weak learning algorithm (which produces randomized hypotheses) for $p(n)$-term DNF. The result then follows from Schapire's boosting technique [19] for converting a weak learning algorithm into a strong one.

For any target $p(n)$-term DNF, 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. Get a sufficiently large sample. (Use for example the Chernoff bounds [3] to get a sufficient sample size). If significantly more than half of the examples are labelled as positive, use the all one monomial for classification. If significantly more than half the examples are labelled as negative, use the all zero monomial for classification. Otherwise, the probabilities of negative and positive examples are approximately equal. We then use the agnostic learning algorithm to learn the function using $\mathcal{N}_1^{\mathcal{G}}$ with quadratic loss. From Section 3, $\mathcal{N}_1^{\mathcal{G}}$ is efficiently agnostically learnable if $\mathcal{G}$ is efficiently agnostically learnable. Suppose that the probability of a positive example is between $1/4$ and $3/4$. 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 and assume that the hypothesis h is no more that $1/(32p(n))$ away from the optimum. Then from [15] Theorem 6, we have $\Pr[f(x)\neq \$_h(x)]\leq \mathbf{E}[Q_h]+1/4<1/2-1/(32p(n))$, where $\$_h(x)$ is a boolean random variable that is 1 with probability $h(x)$ and zero with probability $1-h(x)$. Hence the algorithm is a weak learning algorithm (which produces randomized hypotheses) for learning $p(n)$-term DNF. □

It would also be interesting to find function classes which are not properly efficiently agnostically learnable but are efficiently agnostically learnable with other hypothesis classes. We are unaware of any such example for agnostic learning, although in PAC learning, some function classes are learnable with larger hypothesis classes but not with the target function classes (e.g. k-term DNF is not properly PAC learnable but is learnable when k-CNF is used as the hypothesis class; see [7]).

8 Acknowledgements

We would like to thank Pascal Koiran for numerous discussions which resulted in substantial improvements in several of the results in this paper. This work was partially supported by the Australian Research Council.

References

[1] A. Aho, J. Hopcroft, and J. Ullman. *The Design and Analysis of Computer Algorithms*. Addison-Wesley, London, 1974.

[2] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence and learnability. In *Proc. 35th Annu. IEEE Sympos. Found. Comput. Sci.*, 1993.

[3] D. Angluin and L. Valiant. Fast probabilistic algorithms for hamiltonian circuits and matching. *J. Comput. Syst. Sci.*, 18:155–193, 1970.

[4] M. Anthony and N. Biggs. *Computational Learning Theory*. Cambridge Tracts in Theoretical Computer Science (30). Cambridge University Press, 1992.

[5] A. Barron. Universal approximation bounds for superposition of a sigmoidal function. *IEEE Trans. on Information Theory*, 39:930–945, 1993.

[6] P. L. Bartlett, P. M. Long, and R. C. Williamson. Fat-shattering and the learnability of real-valued functions. In *Proc. 7th Annu. ACM Workshop on Comput. Learning Theory*, pages 299–310. ACM Press, New York, NY, 1994.

[7] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *J. ACM*, 36(4):929–965, 1989.

[8] L. Gurvits and P. Koiran. Approximation and learning of real-valued functions. In *Proc. 2nd European Conf. on Comput. Learning Theory*, 1995.

[9] D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Inform. Comput.*, 100(1):78–150, September 1992.

[10] W. Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, March 1963.

[11] K. Höffgen and H. Simon. Robust trainability of single neurons. In *Proc. 5th Annu. Workshop on Comput. Learning Theory*, pages 428–439, New York, NY, 1992. ACM Press.

[12] M. Jerrum. Simple translation-invariant concepts are hard to learn. *Inform. Comput.*, 113(2):300–311, September 1994.

[13] L. K. Jones. 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, 1992.

[14] M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. *J. Comput. Syst. Sci.*, 48(3):464, 1994.

[15] M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2):115, 1994.

[16] P. Koiran. Efficient learning of continuous neural networks. In *Proc. 7th Annu. ACM Workshop on Comput. Learning Theory*, pages 348–355. ACM Press, New York, NY, 1994.

[17] W. S. Lee, P. L. Bartlett, and R. C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. Technical report, Department of Systems Engineering, Australian National University, 1994.

[18] W. Maass. Agnostic PAC-learning of functions on analog neural networks. Technical report, Institute for Theoretical Computer Science, Technische Universitaet Graz, Graz, Austria, 1993.

[19] R. E. Schapire. The strength of weak learnability. *Machine Learning*, 5(2):197–227, 1990.

[20] L. G. Valiant. A theory of the learnable. *Commun. ACM*, 27(11):1134–1142, November 1984.