Efficient Agnostic Learning of Neural Networks with Bounded Fan-in*

Wee Sun Lee[†], Student Member, IEEE

Peter L. Bartlett[‡], Member, IEEE

Robert C. Williamson[§], Member, IEEE

Abstract

We show that the class of two layer neural networks with bounded fan-in is efficiently learnable in a realistic extension to the Probably Approximately Correct (PAC) learning model. In this model, a joint probability distribution is assumed to exist on the observations and the learner is required to approximate the neural network which minimizes the expected quadratic error. As special cases, the model allows learning real-valued functions with bounded noise, learning probabilistic concepts and learning the best approximation to a target function that cannot be well approximated by the neural network. The networks we consider have real-valued inputs and outputs, an unlimited number of threshold hidden units with bounded fan-in, and a bound on the sum of the absolute values of the output weights. The number of computation

*This work was supported by the Australian Research Council and the Australian Telecommunications and Electronics Research Board. The material in this paper was presented in part at the 6th Australian Conference on Neural Networks, Sydney, 6th-8th February, 1995.

†Electrical Engineering Department, University College UNSW, Australian Defence Force Academy, Canberra, ACT 2600, Australia. This work was completed while this author was at the Department of Systems Engineering, Australian National University.

‡Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, ACT 0200, Australia.

§Department of Engineering, Australian National University, Canberra, ACT 0200, Australia.

steps of the learning algorithm is bounded by a polynomial in $1/\epsilon$, $1/\delta$, n and B where ϵ is the desired accuracy, δ is the probability that the algorithm fails, n is the input dimension and B is the bound on both the absolute value of the target (which may be a random variable) and the sum of the absolute values of the output weights. In obtaining the result, we also extended some results on iterative approximation of functions in the closure of the convex hull of a function class and on the sample complexity of agnostic learning with the quadratic loss function.

Index Terms - Artificial neural networks, agnostic learning, computational learning theory, iterative approximation, polynomial-time learning algorithm, rate of convergence, bounded fan-in neural networks.

1 Introduction

Theoretical studies of feedforward neural networks have provided both positive and negative results on what is learnable by these networks. On the positive side, there are good approximation and estimation results. Two layer feedforward neural networks have been shown to be universal approximators [12], capable of approximating virtually any function of interest provided sufficiently many hidden units are available. Furthermore, for a large class of functions, such as functions with a bound on the first moment of the magnitude distribution of the Fourier transform, integrated squared error of order $O(1/k)$, where k is the number of hidden units, can be achieved [5]. The sample complexity required for learning such networks has also been shown to grow at most polynomially in the size of the networks [11, 6]. These results imply that for such functions, the sample complexity provides no obstacle to efficient (polynomial time) learning with neural networks.

On the other hand, negative results due to Judd [15], and Blum and Rivest [7] indicate that learning may become difficult as the size of the problem increases because of the computational complexity of finding a network that will perform well. These results indicate some of the limits to the complexity of problems that can be solved with such networks. Obviously, to make the problem tractable, more assumptions and restrictions on the networks are required.

By considering fixed sized networks, Maass [21] has shown that such networks can be learned in time bounded by a polynomial in the bit-length of the inputs, the bit-length of the weights, and $1/\epsilon$, where ϵ is the bound on the error of the network. Maass's results were obtained in the agnostic learning model described by Haussler [11] and further developed by Kearns, Schapire and Sellie [17]. This model is a realistic extension of the Probably Approximately Correct (PAC) learning model in computational learning theory. In the agnostic model, no a priori assumptions about the nature of the target concept or target function are made. Instead a joint probability distribution is assumed to exist on the observations and the learner is required to approximate the function in the class which minimizes the expected value of a loss function. The generality of the model means that the results would also hold for learning real-valued functions with noise, learning probabilistic concepts and learning the best approximation to a target function that cannot be well approximated by the neural network.

In this paper, we show that provided the fan-in of the hidden units and the sum of the absolute values of the output weights are bounded, two layer feedforward neural networks (with an unlimited number of threshold hidden units) are learnable in the agnostic model with the quadratic loss function. Furthermore the time complexity grows only polynomially with the input dimension and the sum of absolute values of the output weights (although it is exponential in the fan-in). This implies that networks with a polynomial number of hidden units with bounded fan-in (such that the magnitude of each output weight is bounded by a fixed constant) are efficiently learnable in the agnostic model. Recently, Koiran [18] has independently shown that for a fixed input dimension, functions which can be well approximated by two layer neural networks can be efficiently learned with small bounded noise. However, he did not consider an agnostic setting which allows more general learning problems, such as learning probabilistic concepts, regression and learning the best approximation, to be solved.

In obtaining the result, we extend existing results on approximation and estimation of functions.

For the approximation of functions, Jones [14] and Barron [5] have shown that if the target function is in the closure of the convex hull of a set of function, then by iteratively adding an element of the set of functions so as to minimize the distance to the target function, the squared distance to the target will decrease at a rate of $O(1/k)$, where k is the number of elements added. We extend this result to show that the squared distance will approach that of the best approximation in the convex hull of the set of functions even when the target function is not in the closure of the convex hull, at the same rate of $O(1/k)$. The fact that the target does not have to be in the convex hull allows learning in the agnostic model. The iterative approximation result also allows efficient learning of a network with a small number of hidden units by iteratively adding hidden units until a larger network with the required error is obtained. Using a larger hypothesis class to learn a smaller function class efficiently is a technique commonly used in computational learning theory (see e.g. [21]). It also gives some theoretical support to the common practice of using more hidden units to make learning easier even when the problem can be solved using fewer hidden units (see e.g. [26]).

For the estimation of functions, we extend the results of Barron [6, 3], McCaffrey and Gallant [22], Haussler [11] and Pollard [25]. Barron [6, 3] and McCaffrey and Gallant [22] have shown that for two layer neural networks satisfying a Lipschitz condition, the rate of convergence of $\| \hat{f} - f^* \|^2$ for a particular estimator $\hat{f}$, is $O(\| f_k - f^* \|^2 + kd \log(nm)/m)$ when $\| f_k - f^* \|^2$ converges to zero as $k \to \infty$. Here $f^*(x) = \mathbf{E}(Y|X = x)$, k is the number of hidden units, n is the input dimension, d is the fan-in, m is the sample size, f_k is the network of k hidden units which gives the best approximation to f^* and $\|\hat{f} - f^*\|^2 = \int(\hat{f}(x) - f^*(x))^2 dP(x)$ for an arbitrary probability distribution P. However, in this paper, we use networks with linear threshold functions as hidden units; this class of neural networks does not satisfy a Lipschitz condition. This is also the case for networks with any sigmoidal hidden units if the input values or weights are not restricted to a bounded range. Furthermore, in some cases our function class may not be rich enough to approximate the conditional expectation or target function well. This will often be true if we bound the fan-in of our hidden units. A realistic goal is to require our estimator to converge to the best approximation in the class. It is possible to use the results of Haussler [11] and Pollard [25] to obtain a rate of convergence for $\| \hat{f} - f^* \|^2 - \| f_a - f^* \|^2$ of order $O(\| f_k - f^* \|^2 - \| f_a - f^* \|^2 + \sqrt{kd \log(nm)}/\sqrt{m})$ when f_a is the neural network which gives the best approximation to $f^*(x) = \mathbf{E}[Y|X = x]$. We show that when the neural networks used for estimation form a convex class of functions, the rate of convergence is $O(\| f_k - f^* \|^2 - \| f_a - f^* \|^2 + kd \log(nm)/m)$.

By putting the approximation and estimation bounds together a bound of order $O(\sqrt{d \ln(mn)}/\sqrt{m})$ on the difference between the expected loss of the estimated network and the best network is obtained. Unlike the computational complexity, the sample complexity does not grow exponentially with the fan-in. Barron [4] has shown that for the function class considered here (when the fan-in is not bounded), the rate of convergence of $\|\hat{f} - f^*\|^2$ for any estimator $\hat{f}$ cannot be better than $\Omega(1/m^{(n+2)/(2n+2)})$, where n is the input dimension. Since networks with bounded fan-in include networks with unbounded fan-in with input dimensions smaller than the fan-in, this shows that the rate is close to the best possible when the fan-in of the network is large.

Bounding the fan-in of the hidden units allows us to find an algorithm to agnostically learn the networks in polynomial time. However, by bounding the fan-in, we also lose the universal approximation capabilities of the networks. For boolean functions, Minsky and Papert [23] 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, for learning to recognize handwritten digits, good results can be achieved using low degree polynomials (which can be approximated by networks with small fan-in) as shown by Guyon, Boser and Vapnik [8].

In Section 2 we describe the agnostic learning model and the neural networks we are using and give a statement of the main result. We prove the iterative approximation result in Section 3 and give bounds on the sample complexity in Section 4. In Section 5 we give an algorithm for loading the network and finally we prove the main result in Section 6.

2 Learning Model and Statement of Result

We follow closely the agnostic learning model of Kearns, Schapire and Sellie [17]. Let $\mathcal{X}$ be a set called the *domain* and let the points in $\mathcal{X}$ be called *instances*. In our case the set $\mathcal{X}$ will be a subset of $\mathbb{R}^n$. Let $\mathcal{Y}$ be a set called the *observed range*. In our case, $\mathcal{Y}$ is a subset of $\mathbb{R}$ with magnitude bounded by B. We call the pair (x, y) an observation. The *assumption class* $\mathcal{A}$ is a class of probability distributions on $\mathcal{X} \times \mathcal{Y}$. $\mathcal{A}$ is used to represent the assumptions on the phenomenon that is being learned. In this paper, we let $\mathcal{A}$ be the class of all probability distributions on $\mathcal{X} \times \mathcal{Y}$, of which the following are special cases.

In *regression*, $y \in \mathcal{Y}$ 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 $\mathcal{X}$ to $\mathcal{Y}$ and $A_{D,f} \in \mathcal{A}$ is formed from a distribution D over $\mathcal{X}$ and a function $f \in F$. Observations drawn from $A_{D,f}$ have the form $(x, f(x))$ where x is drawn randomly from D. In learning *probabilistic concepts*, we have $\mathcal{Y} = \{0, 1\}$. Again, there is a class F of functions, mapping $\mathcal{X}$ to $[0, 1]$ and $A_{D,f} \in \mathcal{A}$ is formed from a distribution D over $\mathcal{X}$ and a function $f \in F$. Observations drawn from $A_{D,f}$ are of the form (x, b), where x is drawn randomly from D and $b = 1$ with probability $f(x)$ and $b = 0$ with probability $1 - f(x)$.

Let $\mathcal{Y}'$ be a subset of $\mathbb{R}$ with magnitude bounded by B. The *hypothesis class* $\mathcal{H}$ is a class of functions from $\mathcal{X}$ to $\mathcal{Y}'$. The learning algorithm will attempt to model the behaviour from $\mathcal{A}$ with functions from $\mathcal{H}$. Since the behaviour from $\mathcal{A}$ cannot necessarily 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 the learning algorithm produces a hypothesis $h \in \mathcal{H}$ with performance at least close to the "best" function in $\mathcal{H}$.

The range of the hypothesis class is not necessarily the same as that of the observations (although we use the same bound for both ranges in this paper). If the range of the observations is known (as assumed), clamping the output of the hypothesis which is outside the range to the closest point in the range after learning can only improve the performance of the hypothesis. This is particularly useful in learning probabilistic concepts, where it is desirable to have the output in the range $[0,1]$.

The "best" function in $\mathcal{H}$ is the function which minimizes the expected value of a *loss function* L which maps from $\mathcal{Y}' \times \mathcal{Y}$ to a bounded subset of $\mathbb{R}$. Given a function $h: \mathcal{X} \to \mathcal{Y}'$, the *loss* of h is denoted $L_h(x,y) := L(h(x),y)$. In this paper, we use the *quadratic loss* function $L_h(x,y) := (h(x)-y)^2$. The quadratic loss can be bounded by a constant M because both the observed range and the range of the hypothesis are bounded. Given observations drawn from $A \in \mathcal{A}$, 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 $\mathcal{H}$, we define $opt(\mathcal{H}) := \inf_{h\in\mathcal{H}} \mathbf{E}[L_h]$. 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 [16] and [17].

We now define the class of neural networks considered in this paper.

Definition 1 *The class of bounded fan-in neural networks with k hidden units and sum of absolute values of the output weights bounded by B is defined as*

$$\mathcal{N}_{B,k} := \left\{ x \mapsto \sum_{i=1}^{k} w_i \sigma(v_i \cdot x + v_{i0}) : \sum_{i=1}^{k} |w_i| \le B \text{ and at most } d \text{ of the coordinates } v_{ij} \text{ of } v_i \text{ are nonzero} \right\}$$

where σ is the step function, $\sigma(z) = 1$ for $z > 0$ and $\sigma(z) = 0$ otherwise, $x \in \mathbb{R}^n$, and $v_i \cdot x = \sum_{j=1}^{n} v_{ij} x_j$. Each $\sigma(v_i \cdot x + v_{i0})$, $1 \le i \le k$ is called a hidden unit.

Define the class of bounded fan-in neural networks with sum of absolute values of output weights bounded by B as $\mathcal{N}_B := \bigcup_{k=1}^{\infty} \mathcal{N}_{B,k}$.

Theorem 1 *Let the domain $\mathcal{X}$ be a subset of $\mathbb{R}^n$, the observed range $\mathcal{Y}$ be a subset of $\mathbb{R}$ with* *magnitude bounded by B, the assumption class $\mathcal{A}$ be the class of all probability distributions on $\mathcal{X} \times \mathcal{Y}$ and B be the bound on the sum of absolute values of output weights of $\mathcal{N}_B$. Then there exists a function $m(\epsilon, \delta, n, B)$ bounded by a fixed polynomial in $1/\epsilon$, $1/\delta$, n and B and an algorithm such that, for any $A \in \mathcal{A}$, given any $\epsilon > 0$, $0 < \delta < 1$, $n \geq 1$ and $B > 0$, the algorithm draws $m(\epsilon, \delta, n, B)$ observations, halts in time bounded by another fixed polynomial in $1/\epsilon$, $1/\delta$, n and B, and outputs a hypothesis $h \in \mathcal{N}_B$ such that with probability at least $1 - \delta$, $\mathbf{E}[L_h] \leq opt(\mathcal{N}_B) + \epsilon$.*

For simplicity we use the uniform cost model of computation where real numbers occupy one unit of space and standard arithmetic operations (addition, multiplication etc.) take one unit of time (see [1]). With some modification, a similar result can also be shown to hold in the logarithmic cost model [1] where real numbers are represented in finite precision and operations on them are charged time proportional to the number of bits of precision.

The notation we use closely follows the usual notation used in Computational Learning Theory [2]. We would like to point out the differences between our notation and the notation conventionally used by information theorists and statisticians. We have used m in place of N for the sample size, n in place of d for the dimension of the input space (d is used for the fan-in of the hidden units) and k for the number of hidden units. To satisfy both sets of readers, we give both the sample complexity and risk bounds (expected squared error) in our results.

3 Iterative Approximation

The iterative approximation result in this section is an extension of the results of Jones [14] and Barron [5]. They showed that if a function is in the closure of the convex hull of a 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 the 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 $O(1/k)$.

Theorem 2 *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_1 is chosen to satisfy*

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

and iteratively, f_k is chosen to satisfy

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

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 $k \geq 1$,

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

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

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

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

$$\| \alpha f_{k-1} + \overline{\alpha} g - f \|^2 - \| f^* - f \|^2$$

$$=\| \alpha f_{k-1} + \overline{\alpha} g - f^* \|^2 + 2(\alpha f_{k-1} + \overline{\alpha} g - f^*, f^* - f)$$

$$=\| \alpha (f_{k-1} - f^*) + \overline{\alpha}(g - f^*) \|^2 + 2(\alpha f_{k-1} + \overline{\alpha} g - f^*, f^* - f)$$

$$= \alpha^2 \| f_{k-1} - f^* \|^2 + \overline{\alpha}^2 \| g - f^* \|^2$$

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

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

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

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

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

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

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

we get

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

Substituting into (2) and letting δ go to 0, we get

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

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

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

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

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

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

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

Then (1) follows directly. $\square$

Recently, Koiran [18] 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 [18], he obtained bounds of the form $\| f - f_k \|^2 - d_f^2 \leq 2Cd_f/\sqrt{k} + C^2/k$ where $C > \sqrt{b^2 + d_f^2}$. In comparison, our bound of $O(1/k)$ is asymptotically better than Koiran's bound of $O(1/\sqrt{k})$. The constant in our bound is also independent of the target function, unlike the constant in Koiran's bound.

The class $\mathcal{N}_B$ of bounded fan-in neural networks with the sum of absolute values of weights bounded by B is the convex hull of $G = \{x \mapsto w\sigma(v_i \cdot x + v_{i0}) \colon |w| = B\} \cup \{x \mapsto w \colon |w| = B\}$, where only d coordinates v_{ij} of v_i is nonzero. Since $\| g \| \leq B$ for every $g \in G$, Theorem 2 can be applied to G to bound the rate of convergence of the iterative approximation to the best approximation in $\mathcal{N}_B$ in the Hilbert space with inner product $(f, g) = \int_X f(x)g(x)dP(x)$ where P is a probability measure.

4 Sample Complexity

In the previous section, we have shown that it is possible to approximate functions in $\mathcal{N}_B$ accurately with relatively few hidden units by iteratively adding hidden units to greedily reduce the approximation error. However for learning purposes, we also need to estimate the values of the weights used in the approximation from information contained in a finite set of examples sampled from an unknown distribution. For efficient learning, we also require that the number of examples needed grows only polynomially with the relevant parameters.

When the hidden units satisfy a Lipschitz condition, Barron [6] and McCaffrey and Gallant [22] have shown that the rate of convergence for $\| \hat{f} - f^* \|^2$ is $O(\| f_k - f^* \|^2 + kd\log(nm)/m)$. The Lipschitz condition is used to get an L_∞ cover for the function class. However sigmoidal hidden units with no bound on the weights and linear threshold hidden units do not have finite L_∞ covers. Instead of L_∞ covering numbers, we use l_1 covering numbers which can be bounded when certain combinatorial properties of the function class are known. We show that the rate of convergence is $O(\| f_k - f^* \|^2 - \| f_a - f^* \|^2 + kd\log(nm)/m)$ when the neural networks we are using form a convex class of functions.

For $m \in \mathbb{N}$ and v, $w \in \mathbb{R}^m$, let

$$d_{l_1}(v, w) := \frac{1}{m}\sum_{i=1}^{m} |v_i - w_i|.$$

For $U \subseteq \mathbb{R}^m$, $\epsilon > 0$, we say $C \subseteq \mathbb{R}^m$ is an l_1 *ϵ-cover* of U if for all $v \in U$, there exists $w \in C$ such that $d_{l_1}(v, w) \leq \epsilon$. The l_1 covering number $N(\epsilon, U, l_1)$ is the size of the smallest l_1 ϵ-cover of U.

If $\mathcal{Z}$ is a set, $f: \mathcal{Z} \to \mathbb{R}$ and $\bar{z} \in \mathcal{Z}^m$, define $f_{|\bar{z}} = (f(z_1), \ldots, f(z_m)) \in \mathbb{R}^m$. If F is a set of functions from $\mathcal{Z}$ to $\mathbb{R}$, define $F_{|\bar{z}} := \{f_{|\bar{z}}: f \in F\}$. Let $\hat{\mathbf{E}}_{\bar{z}}(f) = \frac{1}{m}\sum_{i=1}^{m} f(z_i)$.

Theorem 3 *Let $F = \bigcup_{k=1}^{\infty} F_k$ be a convex class of functions mapping from $\mathcal{X}$ to $\mathcal{Y}'$ such that each F_k is permissible*[1]*, and $|f(x)| \leq B$ for all $f \in F$ and $x \in \mathcal{X}$. Let $|y| \leq B$ for all $y \in \mathcal{Y}'$ and P be an arbitrary probability distribution on $\mathcal{Z} = \mathcal{X} \times \mathcal{Y}'$. Let $\bar{F}$ be the closure of f in the space with inner product $\langle f, g \rangle = \int f(x)g(x)dP_X(x)$. Let $C = \max\{B, 1\}$. 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 $f_a \in \bar{F}$ is such that $\int (f_a(x) - f^*(x))^2 dP_X(x) = \inf_{f \in F} \int (f(x) - f^*(x))^2 dP_X(x)$. Then for $m \geq 1$ and each k,*

$$\begin{aligned} P^m &\left\{ \bar{z} \in \mathcal{Z}^m : \exists f \in F_k, \frac{\mathbf{E}(g_f) - \hat{\mathbf{E}}_{\bar{z}}(g_f)}{\nu + \nu_c + \mathbf{E}(g_f)} \geq \alpha \right\} \\ &\qquad \leq \sup_{\bar{z} \in \mathcal{Z}^{2m}} 6N\left(\frac{\alpha \nu_c}{128 C^3}, F_{k|\bar{z}}, l_1\right) \exp(-3\alpha^2 \nu m/(2624 C^4)) \end{aligned}$$

where P^m denotes the product measure.

The constants in the theorem have not been optimized. The proof is long and can be found in the Appendix. The theorem is also true when F is not convex but $f_a = f^*$.

[1]This is a mild measurability condition satisfied by most function classes used for learning, including the one used in this paper. See Haussler [11] for details.

In order to use this result, we need to bound the l_1 covering number of the networks we use for approximating the functions. We first bound the covering number of one hidden unit using a combinatorial parameter called the pseudo-dimension.

Let F be a class of functions mapping from $\mathcal{X}$ to $\mathbb{R}$ and let $x_1, \ldots, x_m \in \mathcal{X}$. We say $x_1, \ldots, x_m$ are *shattered* by 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 $f \in F$ such that for each i,

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

The *pseudo-dimension* is defined as

$$\dim_P(F) = \max\{m \in \mathbb{N} \colon \exists x_1, \ldots, x_m,\ F \text{ shatters } x_1, \ldots, x_m\}$$

if such a maximum exists, and ∞ otherwise.

The pseudo-dimension of a linear threshold unit with n inputs is known to be $n+1$ [11]. For linear functions the pseudo-dimension is k when k is the number of inputs. To bound the covering number with the pseudo-dimension, we use the following lemma from Haussler [11].

Lemma 1 (Pollard [24], Haussler [11]) *Let F be a class of functions from a set $\mathcal{Z}$ into $[0, M]$ where $M > 0$ and suppose $\dim_P(F) = d_p$ for some $1 \leq d_p < \infty$. Then for all $0 < \epsilon \leq M$ and any finite sequence $\bar{z}$ of points in $\mathcal{Z}$,*

$$N(\epsilon, F_{|\bar{z}}, l_1) < 2\left(\frac{2eM}{\epsilon} \ln \frac{2eM}{\epsilon}\right)^{d_p}.$$

The following lemma will be useful in proving the result.

Lemma 2 (Vapnik and Chervonenkis [30], Sauer [27]) *Let F be a class of functions from $S = \{1, 2, \ldots, m\}$ into $\{0,1\}$ with $|F| > 1$ and let d_p be the length of the longest sequence of points z from S such that $F_{|z} = \{0,1\}^{d_p}$. Then, for $m \geq d_p$,*

$$|F| \leq \sum_{i=0}^{d_p} \binom{m}{i} \leq (em/d_p)^{d_p}$$

where e is the base of the natural logarithm.

The following result gives bound a on the covering number which will be used with the uniform convergence result of Theorem 3.

Lemma 3 *Let $\mathcal{N}_{B,k}$ be the class of bounded fan-in neural networks with k hidden units, fixed output weights and sum of absolute values of output weights bounded by B. Then for $m \geq d$ and any sequence $\bar{x} \in \mathcal{X}^m$,*

$$N(\epsilon, \mathcal{N}_{B,k_{|\bar{x}}}, l_1) \leq 2\left(\frac{emn}{d+1}\right)^{k(d+1)}\left(\frac{2eB}{\epsilon}\ln\frac{2eB}{\epsilon}\right)^k.$$

Proof. Let $G = \{x \mapsto \sigma(v_i \cdot x + v_{i0}) :$ at most d of the coordinates of v_i are nonzero$\}$. So the function class G is the union of $\binom{n}{d} \leq n^d$ function classes ($\binom{n}{d}$ possible choices of non-zero input weights) with pseudo-dimension $d+1$. Fix an arbitrary sequence $\bar{x} \in \mathcal{X}^m$. Using Lemma 2, $|G_{|\bar{x}}| \leq n^d\left(em/(d+1)\right)^{d+1}$.

There are at most $n^{kd}\left(em/(d+1)\right)^{k(d+1)}$ ways of picking $(g_1, \ldots, g_k)$ which will give different functions on $\bar{x}$. Let $f = \sum_{i=1}^k w_1 g_1 + \ldots + w_k g_k$ be an arbitrary function in $\mathcal{N}_{B,k}$. For each possible choice of hidden units $(g_1, \ldots, g_k)$, we want to pick $(w_1', \ldots, w_k')$ such that the function it represents has l_1 distance no more than ϵ from the function represented by $(w_1, \ldots, w_k)$. From Lemma 1, for linear functions with k inputs the size of an ϵ-cover is no more than $2\left((2eB/\epsilon)\ln(2eB/\epsilon)\right)^k$.

Hence for any sequence $\bar{x}$, we can construct an l_1 ϵ-cover with no more than $n^{k(d+1)}(em/(d+1))^{k(d+1)} 2\left((2eB/\epsilon)\ln(2eB/\epsilon)\right)^k$ functions. $\square$

5 Loading Algorithm and Computational Complexity

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 3. Note that in Theorem 2, we have a fixed set of output layer weights for each iteration. The algorithm receives as inputs the number of iterations k, the bound on the sum of magnitudes of output weights B, the fan-in d 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 1). The subroutine and proof of correctness are adapted from Farago and Lugosi [10]. 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 with $1 \leq l \leq m$.

Lemma 4 *The subroutine SPLITTING generates all possible dichotomies of m points using a linear threshold unit with d or fewer nonzero weights in $\mathbb{R}^n$ in $O(2^d d^3 n^{2d} m^{2(d+1)})$ steps.*

Proof. The algorithm goes through $\binom{n}{d} \sum_{i=1}^{d+1} 2^i \binom{m}{i} \leq n^d 2^{d+1} m^{d+1}$ iterations of the innermost loop where it does comparisons and solves linear equations. Each set of linear equations has no more than $d+1$ variables and no more than $d+1$ equations. By Gaussian elimination solving each system takes $O(d^3)$ operations. Each comparison against P takes $O(n^d m^{d+1})$. So the total number of operations is $O(2^d d^3 n^{2d} m^{2(d+1)})$.

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

SPLITTING(A)

$\mathcal{W} := \emptyset$;

$P := \emptyset$;

for all d-tuples $(t_1, \ldots, t_d)$ from $\{1, \ldots, n\}$ with $t_1 < \ldots < t_d$

for all $(d+1)_{\leq}$-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

$(v_0, v_{0_t}) \in \{(v, v_t) \in \mathbb{R}^{n+1} \colon \text{only } v_{t_1}, \ldots, v_{t_d}, v_t \text{ are nonzero}\}$

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 1: Subroutine *SPLITTING*

CONSTRUCT(k,B,d,S)

$f := 0$;

$f' := 0$;

$\mathcal{W} := SPLITTING(S_A)$;

$\mathcal{G} := \{x \mapsto \omega\sigma(v \cdot x - v_0) \colon v \in \mathcal{W}, |\omega| = B\} \cup \{B, -B\}$;

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 2: Algorithm *CONSTRUCT*

First note that any dichotomy in $\mathbb{R}^d$ can be implemented by a hyperplane of the form $b \cdot x + b_t = 0$ with some $(b, b_t) \in \mathbb{R}^{d+1}$. 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[2] 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

$$x_i \cdot v + v_t \geq 1, \qquad x_i \in A \cap H_1$$

$$x_j \cdot v + v_t \leq -1, \qquad x_j \in A \cap H_2$$

This system of inequalities defines a polyhedron in weight space. It is possible to select l inequalities for some $l \leq d+1$ 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 and Lugosi [10]). Since we are considering all possible systems of $d+1$ or fewer equations, all possible dichotomies are generated by the algorithm. $\square$

The pseudo-code for the algorithm *CONSTRUCT* is given in Figure 2. It receives as inputs the number of iterations k, the bound on the sum of magnitudes of output weights B, the fan-in d 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 error on the sequence S). The time complexity of the algorithm is $O(2^d d^3 k n^{2d} m^{2(d+1)})$.

[2] Farago and Lugosi [10] 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 $b \cdot x_3 \geq 3$.

6 Main Results

In this section we prove Theorem 1. First, note that if P is a probability distribution on $\mathcal{X} \times \mathcal{Y}$ and $f^*(x) = \mathbf{E}(Y|X = x)$, then for any $f\colon \mathcal{X} \to \mathbb{R}$,

$$\int (y - f(x))^2 dP(x,y) = \int (y - f^*(x))^2 dP(x,y) + \int (f(x) - f^*(x))^2 dP(x,y).$$

Since $\int (y - f^*(x))^2 dP(x,y)$ is constant, choosing f to minimize $\int (f(x) - f^*(x))^2 dP(x,y)$ is the same as choosing f to minimize $\int (y - f(x))^2 dP(x,y)$.

From a sequence $\bar{z}$ of m points independently sampled from P, the *empirical distribution* is the distribution D such that $\Pr(X = x, Y = y) = \frac{Occur(x,y)}{m}$ where $Occur(x,y)$ is the number of occurences of (x,y) in $\bar{z}$.

The following lemma from [28] is useful in bounding the sample complexity.

Lemma 5 *Let $x, y \in \mathbb{R}^+$. Then*

$$\ln x \leq xy - \ln ey.$$

Theorem 4 *Let the function class $\mathcal{N}_B = \bigcup_{k=1}^{\infty} \mathcal{N}_{B,k}$. Let $m(\epsilon, \delta, n, B)$ be the number of observations used by a learning algorithm such that with probability at least $1 - \delta$, $\mathbf{E}(L_h) \leq opt(\mathcal{N}_B) + \epsilon$. If the iterative approximation scheme of Theorem 2 is used, then*

$$m(\epsilon, \delta, n, B) = O\left(\frac{B^2}{\epsilon}\left(\frac{dB^2}{\epsilon} \ln \frac{Bn}{\epsilon} + \ln \frac{1}{\delta}\right)\right). \tag{3}$$

Let $\hat{f}_k$ be the function produced by the iterative approximation scheme, $f^*(x) = \mathbf{E}[Y|X = x]$ and let f_a be such that $\mathbf{E}[(y - f_a(x))^2] = opt(\mathcal{N}_B)$. Then the expectation over samples of size m of the additional error of the estimator $\hat{f}_k$ is

$$\mathbf{E}[\|\hat{f}_k - f^*\|^2 - \|f_a - f^*\|^2] = O\left(\frac{B^2}{k} + \frac{kdB^2 \log(nm)}{m}\right) \tag{4}$$

and if k is set to $(\frac{m}{d \ln nm})^{\frac{1}{2}}$,

$$\mathbf{E}[\|\hat{f}_k - f^*\|^2 - \|f_a - f^*\|^2] = O\left(B^2 \left(\frac{d \ln(mn)}{m}\right)^{\frac{1}{2}}\right). \tag{5}$$

Proof. The function class $\mathcal{N}_B$ is convex and each $\mathcal{N}_{B,k}$ is permissible. Rescale the function class and the target random variable by dividing by B to give $C = 1$ in Theorem 3. All the following calculations are done with the scaled function class and target random variable. (This rescaling trick allows us to obtain a sample complexity with a better exponent on B.) To get the correct accuracy when the function class is scaled back to the original scale, we need to learn to accuracy ϵ/B^2. Let $\alpha = 1/2$ in Theorem 3 and Lemma 3 to get

$$\begin{aligned}
P^m \Big\{\bar{z} \in \mathcal{Z}^m : \exists f \in \mathcal{N}_{1,k}, \mathbf{E}[(y - f(x))^2 - (y - f_a(x))^2] \\
\geq 2\hat{\mathbf{E}}_{\bar{z}}[(y - f(x))^2 - (y - f_a(x))^2] + \nu + \nu_c\Big\} \\
\leq \sup_{\bar{z}\in\mathcal{Z}^{2m}} 6N\left(\frac{\nu_c}{256}, \mathcal{N}_{1,k|\bar{z}}, l_1\right)\exp(-3\nu m/10496) \\
\leq 2\left(\frac{emn}{d+1}\right)^{k(d+1)}\left(\frac{512e}{\nu_c}\ln\frac{512e}{\nu_c}\right)^k \exp(-3\nu m/10496). \qquad (6)
\end{aligned}$$

Let $f'(x) = \hat{\mathbf{E}}_{\bar{z}}[Y|X = x]$. Then $\hat{\mathbf{E}}_{\bar{z}}[(y - \hat{f}_k(x))^2 - (y - f_a(x))^2] = \hat{\mathbf{E}}_{\bar{z}}[(f'(x) - \hat{f}_k(x))^2 - (f'(x) - f_a(x))^2]$. Note that $\hat{\mathbf{E}}_{\bar{z}}[(f'(x) - \hat{f}_k(x))^2 - (f'(x) - f_a(x))^2] \leq \hat{\mathbf{E}}_{\bar{z}}[(f'(x) - \hat{f}_k(x))^2 - (f'(x) - \hat{f}_a(x))^2]$, where $\hat{f}_k$ is the estimated function and $\hat{f}_a$ is the function in the closure of $\mathcal{N}_B$ which minimizes the empirical error. In Theorem 2, set $c = 2$. To get approximation to within $\epsilon/(4B^2)$, it suffices to have $k = 32B^2/\epsilon$. Setting $\nu = \nu_c = \epsilon/(4B^2)$ and the right hand side of (6) to δ gives us

$$\ln\frac{\delta}{12} = k(d+1)\ln\left(\frac{emn}{d+1}\right) + k\ln\left(\frac{2048eB^2}{\epsilon}\ln\frac{2048eB^2}{\epsilon}\right) - \frac{3\epsilon m}{41984B^2}. \qquad (7)$$

Let $\lambda \in (0, 1)$. Use Lemma 5 to get

$$\ln m \leq \frac{3\lambda\epsilon m}{41984k(d+1)B^2} + \ln\frac{41984k(d+1)B^2}{3e\epsilon\lambda}.$$

Substituting into (7),

$$\frac{3m\epsilon(1-\lambda)}{41984B^2} \leq k(d+1)\ln\frac{en}{d+1} + k(d+1)\ln\frac{41984B^2k(d+1)}{3e\epsilon\lambda} + k\ln\left(\frac{2048eB^2}{\epsilon}\ln\frac{2048eB^2}{\epsilon}\right) + \ln\frac{12}{\delta}.$$

Rearranging and then substituting for k, we see that the value of m such that the right hand side of (6) is equal to δ is

$$\begin{aligned}
m &\leq \frac{41984B^2}{3\epsilon(1-\lambda)}\left(k(d+1)\ln\frac{en}{d+1} + k(d+1)\ln\frac{41984B^2k(d+1)}{3e\epsilon\lambda} + \right.\\
&\qquad \left. k\ln\left(\frac{2048eB^2}{\epsilon}\ln\frac{2048eB^2}{\epsilon}\right) + \ln\frac{12}{\delta}\right)\\
&= O\left(\frac{B^2}{\epsilon}\left(kd\ln\frac{Bkn}{\epsilon} + \ln\frac{1}{\delta}\right)\right)\\
&= O\left(\frac{B^2}{\epsilon}\left(\frac{dB^2}{\epsilon}\ln\frac{Bn}{\epsilon} + \ln\frac{1}{\delta}\right)\right).
\end{aligned}$$

For a sample of size m, we now want to find the expected difference between the expected error of the estimator and the expected error of the best network. Let $t \in \mathbb{R}^+$ and set $\nu = 10496\ln\left(12\left(\frac{emn}{d+1}\right)^{k(d+1)}\left(\frac{512e}{\nu_c}\ln\frac{512e}{\nu_c}\right)^k\right)/(3m) + t$ in (6). By picking $\hat{f}_k$ (which is a function of the sample $\bar{z}$) as described above (and setting $c = 2$ in Theorem 2), we obtain

$$\begin{aligned}
P^m\Big\{\bar{z} \in \mathcal{Z}^m : \mathbf{E}[(y-\hat{f}_k(x))^2 - (y-f_a(x))^2] \geq \frac{8}{k} + \\
10496\ln\left(12\left(\frac{emn}{d+1}\right)^{k(d+1)}\left(\frac{512e}{\nu_c}\ln\frac{512e}{\nu_c}\right)^k\right)/(3m) + t + \nu_c\Big\} \\
\leq \exp(-3tm/10496). \qquad (8)
\end{aligned}$$

Define the random variable

$$\begin{aligned}
r(\bar{z}) = \mathbf{E}[(y-\hat{f}_k(x))^2 - (y-f_a(x))^2] - \frac{8}{k} - \\
10496\ln\left(12\left(\frac{emn}{d+1}\right)^{k(d+1)}\left(\frac{512e}{\nu_c}\ln\frac{512e}{\nu_c}\right)^k\right)/(3m) - \nu_c
\end{aligned}$$

and let $p(r) = \max\{r, 0\}$. Then $\Pr(r \geq t) = \Pr(p(r) \geq t)$ when $t \geq 0$ and $\mathbf{E}r \leq \mathbf{E}p(r)$. We have

$$\begin{aligned}
\mathbf{E}p(r) &= \int_{t=0}^{\infty}\Pr(p(r) \geq t)dt\\
&\leq \int_{t=0}^{\infty}\exp(-tm/10496)dt\\
&= \frac{10496}{3m}.
\end{aligned}$$

Setting $\nu_c = 1/m$ and rescaling (by multiplying the function and target random variable by B) gives us (4). Setting $k = \left(\frac{m}{d\ln(nm)}\right)^{\frac{1}{2}}$ in (4) gives us (5). $\square$

Proof. (Theorem 1)

Select a sample of size $m(\epsilon, \delta, n, B)$ as described in Theorem 4. Run algorithm $CONSTRUCT$ on the sample with $k = (32B^2)/\epsilon$. Since time complexity of the algorithm is $O(2^d d^3 k n^{2d} m^{2(d+1)})$ with d fixed, we have a polynomial time algorithm as desired. □

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 [6]) to get a similar result.

7 Discussion and Open Problems

We have shown that the class of two layer neural networks with bounded fan-in is efficiently learnable in an agnostic learning model. However the number of computation steps required by the algorithm grows exponentially with the fan-in. In [20], we show that efficient agnostic learning of this class of neural network is as hard as learning polynomial-size DNF formulae in the PAC model when the fan-in is unbounded. Whether polynomial-size DNF formulae are learnable is an important open problem in Computational Learning Theory first posed by Valiant [29] in 1984. It is widely believed that polynomial-size DNF formulae are not efficiently learnable in the PAC model [13]. However, even if agnostic learning of two layer neural networks is hard, the question of whether efficient function learning (with no noise or restricted noise models) is possible is still open.

Our result implies that it is possible to efficiently learn the best function in the class of single hidden layer neural networks with bounded fan-in with a fixed number of hidden units and bounded output weights using the class of single hidden layer neural networks with bounded fan-in and an unbounded number of hidden units. It is interesting to ask whether efficient agnostic learning is possible for networks with a fixed number of hidden units when we are restricted to choosing hypotheses from the same function class. This would determine whether it is computationally advantageous to use a larger network to learn the best function in a smaller network.

If the closure of a function class is not convex (for some input distribution), the sample complexity for agnostic learning cannot be better than $\Omega(\ln(1/\delta)/\epsilon^2)$ [19] if we are restricted to hypotheses from the same function class. It is interesting to note that the closure of neural networks with a fixed number of hidden units is not convex for some input distributions. Since the sample complexity for agnostic learning the convex hull of the network is at most a logarithmic factor larger, it may be advantageous to allow the size of the network to grow as a function of the sample size (to get better approximation) instead of trying to find the best hypothesis within the fixed sized network.

One possible extension to this work is to analyse two layer networks with other basis functions using techniques similar to those used in this paper. It is easy to show that in fixed dimensions and with bounded sum of absolute values of output weights, networks with axis parallel rectangle basis functions are efficiently learnable in our agnostic model. Similarly, two layer networks with a mixture of halfspaces and axis parallel rectangles as basis functions can be shown to be learnable. More generally, let F be a set of basis functions. If there exists an algorithm that can enumerate $F_{|z}$ for an arbitrary sample z, in time polynomial in all the relevant parameters, then two layer networks with basis functions from F are agnostically learnable. Using techniques from Maass [21] to load the networks (as done in Koiran [18]), more function classes can be shown to be agnostically learnable. Some general conditions for agnostically learning such networks are given in [20].

8 Acknowledgements

We would like to thank Andrew Barron for improving the rate in Theorem 2 and suggesting that we aim for the rate in Theorem 3. We would also like to thank Pascal Koiran and David Pollard

for providing us with preprints.

9 Appendix

9.1 Proof of Theorem 7

The proof is similar to that used by Haussler [11] and Pollard [25]. Theorem 3 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 $f \in F$. Haussler's result applies to more general random variables but only when they are nonnegative while Pollard 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 F is used to satisfy this condition. 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

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

The function $d_{\nu,\nu_c}(r,s)$ is a variant of the function d_ν introduced by Haussler [11].

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

$$P^m\{\bar{z} \in \mathcal{Z}^m : \exists f \in F, d^1_{\nu,\nu_c}(\mathbf{E}(g_f), \hat{\mathbf{E}}_{\bar{z}}(g_f)) \geq \alpha\} \tag{9}$$

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

$$P^m\{\bar{z} \in \mathcal{Z}^m : \exists f \in F, d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\bar{z}}(g_f^2)), \mathbf{E}(g_f^2)) > \alpha\}.$$

The two probabilities will be bounded separately. The random variables in the second term on the right hand side of inequality 9 are nonnegative; hence a result similar to Haussler's can be used. With minor modification to the proof, Theorem 3 in [11] becomes

Theorem 5 (Haussler [11]) *Let F be a permissible set of functions on $\mathcal{Z}$ with $0 \le f(z) \le M$ for all $f \in F$ and $z \in \mathcal{Z}$. Assume $\nu, \nu_c > 0$ and $0 < \alpha < 1$. Suppose that $\bar{z}$ is generated by m independent random draws according to any probability measure P on $\mathcal{Z}$. Then*

$$P^m\{\bar{z} \in \mathcal{Z}^m : \exists f \in F, d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\bar{z}}(f), \mathbf{E}(f)) > \alpha\} \le \sup_{\bar{z} \in \mathcal{Z}^{2m}} 4N(\alpha\nu_c/2, F_{|\bar{z}}, l_1)\exp(-\alpha^2\nu m/2M).$$

We will now bound the first term on the right hand side of equation (9). 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 9. 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 12. The following three results will be useful for the proof.

We will use the following result derived by Haussler [11] from Chebyshev's inequality.

Lemma 6 (Haussler [11]) *Let $V_1, \ldots, V_m$ be independent identically distributed random variables with range $0 \le V_i \le M$ and $\mathbf{E}(V_i) = \mu, 1 \le i \le m$. 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 [3], we use the following inequality developed by Craig [9] in his proof of Bernstein's inequality.

Lemma 7 *Craig [9]] Let $V_i, \ldots, V_m$ be independent identically distributed random variables which satisfy $|V_i - \mathbf{E}V_i| \leq 3h$ for $i = 1, \ldots, m$,*

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

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

Lemma 8 *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 7 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

$$\Pr\left(S_V - \hat{S}_V \geq \frac{\tau}{m\xi} + \frac{K_2 \xi S_V}{2(1-c)}\right) \leq \exp(-\tau). \tag{10}$$

Now $|V_i - \mathbf{E}V_i| \leq 2K_1$ so $h = \frac{2K_1}{3}$ satisfies the required condition in Lemma 7. 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 (10) gives the required inequality. $\Box$

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

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

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

$$\mathbf{E}(f) - \hat{\mathbf{E}}_{\bar{z}}(f) \geq \alpha(\nu + \nu_c) + \alpha \mathbf{E}(f) \tag{11}$$

and $d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\bar{z}}(f^2), \mathbf{E}(f^2)) \leq \alpha$. Draw another independent random sample $\bar{z}'$ of length m. From Lemma 8, 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 $\mathbf{E}(f) - \hat{\mathbf{E}}_{\bar{z}'}(f) \geq \alpha(\nu+\nu_c)/2 + \alpha\mathbf{E}(f)/2$ is less than $1/4$. Since $|f(z)^2| \leq K_1^2$, from Lemma 6 we find that for $m \geq \frac{K_1^2}{\alpha^2(\nu+\nu_c)}$, the probability that $d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\bar{z}'}(f^2), \mathbf{E}(f^2)) > \alpha$ 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

$$\mathbf{E}(f) - \hat{\mathbf{E}}_{\bar{z}'}(f)) < \alpha(\nu + \nu_c)/2 + \alpha\mathbf{E}(f)/2 \tag{12}$$

and $d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\bar{z}'}(f^2), \mathbf{E}(f^2)) \leq \alpha$. Subtracting (12) from (11) and using the independence of the samples, we have

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

□

The following two Lemmas will be useful for proving Lemma 12.

Lemma 10 *Let $m \geq 1$ be arbitrary and U be the uniform distribution over $\{-1, 1\}$. For a fixed $\bar{z}\bar{z}' \in \mathcal{Z}^{2m}$, any function f and u_i, $i = 1, \ldots, m$ drawn independently from U,*

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

Proof.

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

$\square$

Lemma 11 *Let $m \geq 1$ be arbitrary and U be the uniform distribution over $\{-1,1\}$. For a fixed $\bar{z}\bar{z}' \in \mathcal{Z}^{2m}$, any function f and u_i, $i=1,\ldots,m$ drawn independently from U,*

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

where G_ϵ is an l_1 ϵ-cover of $F_{|\bar{z}\bar{z}'}$.

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

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

In the following lemma, we bound the probability involving the difference between the two empirical averages that arose in Lemma 9 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 12 *Let F be a permissible class of functions with $|f(z)| \leq K_1$ for all $f \in F$ and $z \in \mathcal{Z}$. Let $K_2 \geq 1$ and the distribution P of z be such that $\mathbf{E}f(z) \geq 0$ and $\mathbf{E}(f^2) \leq K_2\mathbf{E}(f)$ for all $f \in F$. Assume $\nu, \nu_c > 0, 0 < \alpha \leq 1/2$. Then for $m \geq \max\left\{\frac{4(K_1+K_2)}{\alpha^2(\nu+\nu_c)}, \frac{K_1^2}{\alpha^2(\nu+\nu_c)}\right\}$,*

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

$$\left. d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\bar{z}'}(f^2), \mathbf{E}(f^2)) \leq \alpha \right\} \leq \sup_{\bar{z}\in\mathcal{Z}^{2m}} N\left(\frac{\alpha\nu_c}{4}, F_{|\bar{z}}, l_1\right) \exp\left(-\frac{3\alpha^2\nu m}{4K_1 + 162K_2}\right). \tag{13}$$

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

$$\hat{\mathbf{E}}_{\bar{z}'(f)} - \hat{\mathbf{E}}_{\bar{z}}(f) \geq \frac{\alpha(\nu+\nu_c)}{2} + \frac{\alpha\mathbf{E}(f)}{2} \tag{14}$$

and

$$|\hat{\mathbf{E}}_{\bar{z}}(f^2) - \mathbf{E}(f^2)| \leq \alpha(\nu+\nu_c) + \alpha\mathbf{E}(f^2) + \alpha\hat{\mathbf{E}}_{\bar{z}}(f^2) \tag{15}$$

and

$$|\hat{\mathbf{E}}_{\bar{z}'}(f^2) - \mathbf{E}(f^2)| \leq \alpha(\nu+\nu_c) + \alpha\mathbf{E}(f^2) + \alpha\hat{\mathbf{E}}_{\bar{z}'}(f^2). \tag{16}$$

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

$$\begin{aligned}
\hat{\mathbf{E}}_{\bar{z}'}(f) - \hat{\mathbf{E}}_{\bar{z}}(f) &\geq \frac{\alpha(\nu+\nu_c)}{2} + \frac{\alpha\mathbf{E}(f)}{2} \\
&\geq \frac{\alpha(\nu+\nu_c)}{2} + \frac{\alpha\mathbf{E}(f^2)}{2K_2} \\
&\geq \frac{\alpha(\nu+\nu_c)}{2} + \frac{\alpha}{2}\left[\frac{(1-\alpha)\hat{\mathbf{E}}_{\bar{z}}(f^2) - \alpha(\nu+\nu_c)}{2K_2(1+\alpha)} + \frac{(1-\alpha)\hat{\mathbf{E}}_{\bar{z}'}(f^2) - \alpha(\nu+\nu_c)}{2K_2(1+\alpha)}\right] \\
&\geq \frac{\alpha(\nu+\nu_c)}{2} - \frac{\alpha^2(\nu+\nu_c)}{2K_2(1+\alpha)} + \frac{\alpha(1-\alpha)(3\hat{\mathbf{E}}_{\bar{z}}(f^2) + 3\hat{\mathbf{E}}_{\bar{z}'}(f^2))}{12(1+\alpha)K_2}.
\end{aligned} \tag{17}$$

The fact that the random variables are independent means that the probability in (13) remains unchanged when each component of $\bar{z}$ is randomly interchanged with the corresponding component of $\bar{z}'$. Let U be the uniform distribution over $\{-1,1\}$ with u_i, $i=1,\ldots,m$ drawn independently from U.

$$
\begin{aligned}
& P^{2m}\left\{\bar{z}\bar{z}' \in \mathcal{Z}^{2m} : \exists f \in F, \frac{\hat{\mathbf{E}}_{\bar{z}'}(f) - \hat{\mathbf{E}}_{\bar{z}}(f)}{(\nu+\nu_c)+\mathbf{E}(f)} \geq \frac{\alpha}{2} \text{ and } d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\bar{z}}(f^2), \mathbf{E}(f^2)) \leq \alpha \text{ and} \right. \\
& \qquad \left. d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\bar{z}'}(f^2), \mathbf{E}(f^2)) \leq \alpha \right\} \\
\leq \; & P^{2m}\left\{\bar{z}\bar{z}' \in \mathcal{Z}^{2m} : \exists f \in F, \hat{\mathbf{E}}_{\bar{z}'}(f) - \hat{\mathbf{E}}_{\bar{z}}(f) \geq \frac{\alpha(\nu+\nu_c)}{2} - \frac{\alpha^2(\nu+\nu_c)}{2K_2(1+\alpha)} + \right. \\
& \qquad \left. \frac{\alpha(1-\alpha)(3\hat{\mathbf{E}}_{\bar{z}}(f^2)+3\hat{\mathbf{E}}_{\bar{z}'}(f^2))}{12(1+\alpha)K_2} \right\} \quad \text{(using (17))} \\
= \; & P^{2m} \times U^m \left\{\bar{z}\bar{z}' \in \mathcal{Z}^{2m}, \bar{u} \in U^m : \exists f \in F, \frac{1}{m}\sum_{i=1}^m u_i(f(z_i)-f(z_i')) \geq \right. \\
& \qquad \left. \frac{\alpha(\nu+\nu_c)}{2} - \frac{\alpha^2(\nu+\nu_c)}{2K_2(1+\alpha)} + \frac{\alpha(1-\alpha)(3\hat{\mathbf{E}}_{\bar{z}}(f^2)+3\hat{\mathbf{E}}_{\bar{z}'}(f^2))}{12(1+\alpha)K_2} \right\} \\
\leq \; & \sup_{\bar{z}\bar{z}' \in \mathcal{Z}^{2m}} U^m \left\{\bar{u} \in U^m : \exists f \in F, \frac{1}{m}\sum_{i=1}^m u_i(f(z_i)-f(z_i')) \geq \right. \\
& \qquad \left. \frac{\alpha(\nu+\nu_c)}{2} - \frac{\alpha^2(\nu+\nu_c)}{2K_2(1+\alpha)} + \frac{\alpha(1-\alpha)(3\hat{\mathbf{E}}_{\bar{z}}(f^2)+3\hat{\mathbf{E}}_{\bar{z}'}(f^2))}{12(1+\alpha)K_2} \right\} \\
\leq \; & \sup_{\bar{z}\bar{z}' \in \mathcal{Z}^{2m}} U^m \left\{\bar{u} \in U^m : \exists f \in G_c, \frac{1}{m}\sum_{i=1}^m u_i(f(z_i)-f(z_i')) \geq \right. \\
& \qquad \left. \frac{\alpha\nu}{2} - \frac{\alpha^2\nu}{2K_2(1+\alpha)} + \frac{\alpha(1-\alpha)(m\mathbf{Var}(\frac{1}{m}\sum_{i=1}^m u_i(f(z_i)-f(z_i'))))}{12(1+\alpha)K_2} \right\} \qquad (18)
\end{aligned}
$$

using Lemmas 11 and 10, where G_c is an $\alpha\nu_c/4$-cover of $F_{|\bar{z}}$. (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 7. Let $c = \xi h = \frac{2K_1\xi}{3}$ in Lemma 7. 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 (18) is less than

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

□

The following lemma is useful for bounding the second term on the right hand side of Equation (9), using Theorem 5.

Lemma 13 *Let F be a class of functions with $|f(z)| \le K_1$ for all $f \in F$ and $z \in \mathcal{Z}$. Let $F^2 = \{f^2 : f \in F\}$ and $\bar{z} \in \mathcal{Z}^m$. Then for all $\epsilon > 0$,*

$$N(\epsilon, F^2_{|\bar{z}}, l_1) \le N\left(\frac{\epsilon}{2K_1}, F_{|\bar{z}}, l_1\right).$$

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

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

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

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

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

Proof. From equation (9),

$$P^m\{\bar{z}\in\mathcal{Z}^m : \exists f\in F, d^1_{\nu,\nu_c}(\mathbf{E}(g_f),\hat{\mathbf{E}}_{\bar{z}}(g_f)) \ge \alpha\}$$

$$\le P^m\{\bar z \in \mathcal{Z}^m : \exists f \in F, d^1_{\nu,\nu_c}(\mathbf{E}(g_f), \hat{\mathbf{E}}_{\bar z}(g_f)) \ge \alpha \text{ and } d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\bar z}(g_f^2), \mathbf{E}(g_f^2)) \le \alpha\} +$$

$$P^m\{\bar z \in \mathcal{Z}^m : \exists f \in F, d_{\nu,\nu_c}(\hat{\mathbf{E}}_{\bar z}(g_f^2)), \mathbf{E}(g_f^2)) > \alpha\}.$$

From Lemma 9 and Lemma 12, the first term on the right hand side is bounded by $\sup_{\bar z \in \mathcal{Z}^{2m}} 2N\left(\frac{\alpha\nu_c}{4}, F_{|\bar z}, l_1\right) \exp\left(-\frac{3\alpha^2\nu m}{4K_1+162K_2}\right)$. From Theorem 5 and Lemma 13, the second term on the right hand side is bounded by $\sup_{\bar z \in \mathcal{Z}^{2m}} 4N\left(\frac{\alpha\nu_c}{4K_1}, F_{|\bar z}, l_1\right) \exp(-\alpha^2\nu m/2K_1^2)$. □

We now show that convexity is a sufficient condition for $\mathbf{E}(g_f^2) \le K_2\mathbf{E}(g_f)$ for some constant K_2.

Lemma 14 *Let F be a convex class of functions with $|f(x)| \le B$ for every $f \in F$ and $x \in \mathcal{X}$. Let $|y| \le B$ 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 F is such that $\int (f_a(x) - f^*(x))^2 dP_X(x) = \inf_{f\in F} \int (f(x) - f^*(x))^2 dP_X(x)$ where $f^*(x) = \mathbf{E}[Y|X = x]$. Then for every $f \in F$*

$$\begin{aligned}\mathbf{E}[((y - f(x))^2 - (y - f_a(x))^2)^2] &\le 16B^2\mathbf{E}(f(x) - f_a(x))^2 \\ &\le 16B^2\mathbf{E}[(y - f(x))^2 - (y - f_a(x))^2]. \qquad (19)\end{aligned}$$

Proof. For the first part of inequality (19),

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

For the second part of inequality (19), 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 $\mathbf{E}[(f^*(x) - f_a(x))(f_a(x) - f(x))] \geq 0$. Let $\bar{F}$ be the closure of F. Then $\bar{F}$ is convex and $f_a \in \bar{F}$. From convexity, $f \in \bar{F}$ implies $\alpha f + (1-\alpha) f_a \in \bar{F}$ for $\alpha \in [0,1]$. Since f_a is the best approximation in $\bar{F}$,

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

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 $\mathbf{E}(f^*(x) - f_a(x))(f_a(x) - f(x)) \geq 0$. $\square$

Lemma 15 *Let F be a class of functions with $|f(x)| \leq B$ for all $f \in F$ and $x \in \mathcal{X}$. Let $|y| \leq B$ for all $y \in \mathcal{Y}$ and $G = \{g_f : g_f(x,y) = (y - f(x))^2 - (y - f_a(x))^2, f \in F\}$ where f_a is an arbitrary function. Let $\bar{z} \in \mathcal{Z}^m = (\mathcal{X} \times \mathcal{Y})^m$. Then*

$$
N(\epsilon, G_{|\bar{z}}, l_1) \leq N(\epsilon/4B, F_{|\bar{z}}, l_1).
$$

Proof. For any $f, g \in F$ 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 $T = \{f_1, \ldots, f_N\}$ is an $\epsilon/4B$-cover for $F_{|\bar{z}}$, $T_G = \{g_{f_1}, \ldots, g_{f_N}\}$ is an ϵ-cover for $G_{|\bar{z}}$. $\square$

We now have everything we need to prove Theorem 3.

Proof. (Theorem 3) In Theorem 6, K_1 can be set to $8C^2$. Using the convexity of $F = \bigcup_{k=1}^{\infty} F_k$ and Lemma 14, K_2 can be set to to $16C^2$. Using Lemma 15, the right hand side of Theorem 6 can be bounded by

$$\sup_{\bar{z}\in\mathcal{Z}^{2m}} 2N\left(\frac{\alpha\nu_c}{16C}, F_{k|\bar{z}}, l_1\right)\exp\left(-\frac{3\alpha^2\nu m}{2624C^2}\right) + \sup_{\bar{z}\in\mathcal{Z}^{2m}} 4N\left(\frac{\alpha\nu_c}{128C^3}, F_{k|\bar{z}}, d_{l_1}\right)\exp(-\alpha^2\nu m/128C^4)$$
$$\leq \sup_{\bar{z}\in\mathcal{Z}^{2m}} 6N\left(\frac{\alpha\nu_c}{128C^3}, F_{k|\bar{z}}, d_{L_1}\right)\exp(-3\alpha^2\nu m/2624C^4).$$

□

References

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

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

[3] A. R. Barron. Complexity regularization with applications to artificial neural networks. In G. Roussa, editor, *Nonparametric Functional Estimation*, pages 561–576. Kluwer Academic, Boston, MA and Dordrecht, the Netherlands, 1990.

[4] A. R. Barron. Neural net approximation. In *Proc. 7th Yale Workshop on Adaptive and Learning Systems*, 1992.

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

[6] A. R. Barron. Approximation and estimation bounds for artificial neural networks. *Machine Learning*, 14:115–133, 1994.

[7] A. Blum and R. Rivest. Training a 3-node neural network is NP-complete. *Neural Networks*, 5:117–127, 1992.

[8] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In *Proc. 5th Annu. Workshop on Comput. Learning Theory*, pages 144–152. ACM Press, New York, NY, 1992.

[9] C. C. Craig. On the Tchebychef inequality of Bernstein. *Annals of Mathematical Statistics*, 4:94–102, 1933.

[10] A. Farago and G. Lugosi. Strong universal consistency of neural network classifiers. *IEEE Trans. on Information Theory*, 39:1146–1151, 1993.

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

[12] K. Hornik, M. Stinchcombe, and H. White. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3:551–560, 1990.

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

[14] 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.

[15] J. S. Judd. *Neural Network Design and the Complexity of Learning*. MIT Press, 1990.

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

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

[18] 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.

[19] W. S. Lee, P. L. Bartlett, and R. C. Williamson. The importance of convexity in learning with squared loss. Submitted to IEEE Trans. on Information Theory, 1995.

[20] W. S. Lee, P. L. Bartlett, and R. C. Williamson. On efficient agnostic learning of linear combinations of basis functions. In *Proc. 8th Annu. Workshop on Comput. Learning Theory*, pages 369–376. ACM Press, New York, NY, 1995.

[21] W. Maass. Agnostic PAC-learning of functions on analog neural networks. *Neural Computation*, 7(5):1054–1078, 1995.

[22] D. F. McCaffrey and A. R. Gallant. Convergence rates for single hidden layer feedforward networks. *Neural Networks*, 7(1):147–158, 1994.

[23] M. Minsky and S. Papert. *Perceptrons.* MIT Press, Cambridge, MA, 1969.

[24] D. Pollard. *Convergence of Stochastic Processes.* Springer-Verlag, Berlin, 1984.

[25] D. Pollard. Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics, Theory and Applications*, 22(3):271, 1995.

[26] D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning internal representations by error propagation. In *Parallel Distributed Processing – Explorations in the Microstructure of Cognition*, chapter 8, pages 318–362. MIT Press, 1986.

[27] N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (Series A)*, 13:145–147, 1972.

[28] J. Shawe-Taylor, M. Anthony, and N.L. Biggs. Bounding the sample size with the Vapnik-Chervonenkis dimension. *Discrete Applied Mathematics*, 42:65–73, 1993.

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

[30] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probabability and its Applications*, 16(2):264–280, 1971.

Footnotes

1. This is a mild measurability condition satisfied by most function classes used for learning, including the one used in this paper. See Haussler [11] for details.

2. Farago and Lugosi [10] 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 $b \cdot x_3 \geq 3$.

List of Figures