The Importance of Convexity in Learning with Squared Loss

Wee Sun Lee* Peter L. Bartlett† Robert C. Williamson‡

August 26, 1997

Abstract

We show that if the closure of a function class F under the metric induced by some probability distribution is not convex, then the sample complexity for agnostically learning F with squared loss (using only hypotheses in F) is $\Omega(\ln(1/\delta)/\epsilon^2)$ where $1-\delta$ is the probability of success and ϵ is the required accuracy. In comparison, if the class F is convex and has finite pseudo-dimension, then the sample complexity is $O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$. If a non-convex class F has finite pseudo-dimension, then the sample complexity for agnostically learning the closure of the convex hull of F, is $O\left(\frac{1}{\epsilon}\left(\frac{1}{\epsilon}\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$. Hence, for agnostic learning, learning the convex hull provides better approximation capabilities with little sample complexity penalty.

Index Terms - Sample complexity, agnostic learning, convex hull, artificial neural networks, computational learning theory.

*School of Electrical Engineering, University College UNSW, Australian Defence Force Academy, Canberra, ACT 2600.

†Department of Systems Engineering, RSISE, Australian National University, Canberra, ACT 0200, Australia.

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

1 Introduction

We study the number of examples necessary for learning using the squared loss function when the function class used for learning does not necessarily match the phenomenon being learned. The learning model we use, commonly called agnostic learning [9, 12, 16] is an extension of the popular Probably Approximately Correct (PAC) learning model in computational learning theory [1]. Unlike the PAC model, in agnostic learning we do not assume the existence of a target function which belongs to a known class of functions. Instead, we assume that the phenomenon is described by a joint probability distribution on $\mathcal{X} \times \mathcal{Y}$, where $\mathcal{X}$ is the domain and the range $\mathcal{Y}$ is a bounded interval in $\mathbb{R}$. This is often more realistic in practice where measurements are often noisy and very little is known about the target function.

Under these assumptions, there may not be a function in the class we are using that has small error. Instead, we aim to produce a hypothesis with performance close to that of the best function in the class. For a learning algorithm to agnostically learn a function class F, we require that for any probability distribution P on $\mathcal{X} \times \mathcal{Y}$, the algorithm draws a sequence of i.i.d. examples from P and produces a hypothesis f from F such that with probability at least $1 - \delta$, the expected loss of f is no more than ϵ away from the best expected loss of functions in F. The sample complexity is the smallest number of examples required for any learning algorithm to agnostically learn F.

Many learning problems are special cases of agnostic learning. In learning real-valued functions, the inputs are random but the targets are produced by a function in the class. In regression problems, both the input and outputs are random but the conditional expectation is known to be in the class. An agnostic learning algorithm can also be used to learn the best approximation to the target function when the target function is not in the class.

Table 1 shows some of the known results for learning with squared loss. (The technical conditions such as pseudo-dimension and covering number are described in Section 2.)

Learning Problem	*Sample Complexity*
Learning real-valued functions (Function classes with finite pseudo-dimension)	$O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$ [20, 9]
Regression (Function classes with finite L_∞-covering numbers)	$O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$ [2, 18]
Agnostic learning (Convex function classes with finite pseudo-dimension)	$O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$ [this paper]
Agnostic learning (When the closure of the function class is not convex)	$\Omega\left(\frac{\ln(1/\delta)}{\epsilon^2}\right)$ [this paper]
Agnostic learning (Convex hulls of function classes with finite pseudo-dimension)	$O\left(\frac{1}{\epsilon}\left(\frac{1}{\epsilon}\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$ [this paper]

Table 1: Sample complexity for learning with squared loss.

For learning real-valued function classes with finite pseudo-dimension, results by Pollard [20] and Haussler [9] can be used to bound the sample complexity by $O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$. For classes with finite L_∞-covering numbers, with zero mean noise and a target function selected from F (regression), Barron [2] and McCaffrey and Gallant [18] have shown that the sample complexity can be bounded by $O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$. For the agnostic case, we show that for *convex* function classes with finite pseudo-dimension, the sample complexity is bounded by $O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$. We also show that if the closure of the function class is *not convex* then a sample size of $\Omega\left(\frac{\ln(1/\delta)}{\epsilon^2}\right)$ is necessary for agnostic learning.

Given that the sample complexity for learning is larger when the closure of the function class is not convex, it is natural to try to learn the convex hull of the function class. We show that for non-convex function classes with finite pseudo-dimension, the sample complexity of learning the closure of the convex hull of the function class is bounded by $O\left(\frac{1}{\epsilon}\left(\frac{1}{\epsilon}\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$ examples. This means that for agnostic learning, using the convex hull of the function class instead of the function class itself gives better approximation with little penalty on the sample complexity. This result is of practical interest since for many commonly used function classes (such as single hidden layer neural networks with a fixed number of hidden units), the closure of the function class (under the metric induced by some probability distribution) is not convex. Also of interest is the fact that learning the convex hull is computationally not significantly more difficult than learning the function class itself [14] because it is possible to iteratively learn each component of the convex combination.

In Section 2, we give formal definitions for the learning model and concepts used in this paper. The lower bound for the sample complexity for agnostically learning non-convex function classes is given in Section 3. We give the result on learning convex function classes and the convex hull of non-convex function classes in Section 4 and discuss the results in Section 5. The results in this paper are given in terms of sample complexity for learning to a certain accuracy (as is commonly done in computational learning theory). Analogous results can be obtained in terms of the accuracy of estimators with respect to the sample size (as is commonly done in statistics and information theory).

2 Definitions and Learning Model

The agnostic learning model used here is based on the model described by Kearns, Schapire and Sellie [12]. Let $\mathcal{X}$ be a set called the *domain* and let the *range* $\mathcal{Y}$ be a bounded subset of $\mathbb{R}$. We call the pair $(x, y) \in \mathcal{X} \times \mathcal{Y}$ an example. A class F of real-valued functions defined on $\mathcal{X}$ is *agnostically learnable* if there exists a function $m(\epsilon, \delta, \mathcal{Y})$ and an algorithm such that for any bounded range $\mathcal{Y}$ and any probability distribution P on $\mathcal{X} \times \mathcal{Y}$, given $\mathcal{Y}$ and any $0 < \delta \leq 1$ and $\epsilon > 0$, the algorithm draws $m(\epsilon, \delta, \mathcal{Y})$ i.i.d. examples and outputs a hypothesis $h \in F$ such that with probability at least $1 - \delta$, $\mathbf{E}[(Y - h(X))^2] \leq \inf_{f \in F} \mathbf{E}[(Y - f(X))^2] + \epsilon$. Here (X, Y) is a random variable with distribution P. The *sample complexity* is the function $m(\epsilon, \delta, \mathcal{Y})$ with the smallest value for each ϵ, δ and $\mathcal{Y}$ such that an algorithm for agnostically learning F exists. Since a change of the range $\mathcal{Y}$ is equivalent to a rescaling, we ignore the dependence of m on $\mathcal{Y}$ in what follows.

A useful notion for bounding the sample complexity is the pseudo-dimension of a function class. A sequence of points $x_1, \ldots, x_d$ from $\mathcal{X}$ is *shattered* by F if there exists $r \in \mathbb{R}^d$ such that for each $b \in \{0,1\}^d$, there is an $f \in F$ such that for each i, $f(x_i) > r_i$ if $b_i = 1$ and $f(x_i) \leq r_i$ if $b_i = 0$. The *pseudo-dimension* of F, $\dim_P(F) := \max\{d \in \mathbb{N} : \exists x_1, \ldots, x_d,\ F \text{ shatters } x_1, \ldots, x_d\}$ if such a maximum exists and is ∞ otherwise.

We shall consider *uniformly bounded* function classes, by which we mean classes of functions that map to some bounded subset of $\mathbb{R}$. As Lemma 12 below shows, the pseudo-dimension of such a function class can be used to bound the *covering number* of the function class. For a pseudo-metric space (S, ρ), a set $T \subseteq S$ is an ϵ-cover of $R \subseteq S$ if, for all $x \in R$, there is a $y \in T$ with $\rho(x, y) < \epsilon$. The covering number $N(\epsilon, R, \rho)$ denotes the size of the smallest ϵ-cover for R. If $N(\epsilon, R, \rho)$ is finite for all $\epsilon > 0$, we say that R is *totally bounded.*

We now define the various metrics that are found in this paper. For real-valued functions f, g having a bounded range, let $d_{L_\infty}(f, g) := \sup\{|f(x) - g(x)| : x \in \mathcal{X}\}$. For $m \in \mathbb{N}$ and $v, w \in \mathbb{R}^m$, let $d_{l_\infty}(v, w) := \max\{|v_i - w_i| : i = 1, \ldots, m\}$ and $d_{l_1}(v, w) := \frac{1}{m}\sum_{i=1}^m |v_i - w_i|$.

3 Lower Bound for Sample Complexity

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

Definition 1 *Let $P_{\mathcal{X}}$ be a probability distribution on $\mathcal{X}$ and H be the Hilbert space with inner product $\langle f, g \rangle = \int fg \; dP_{\mathcal{X}}$. Let $\|\cdot\|$ denote the induced norm, $\|g\| := \langle g, g \rangle^{1/2}$. We say that F is* closure-convex *if for all $P_{\mathcal{X}}$ on $\mathcal{X}$, the closure of F in the corresponding Hilbert space H is convex.*

Let $\bar{F}$ denote the closure of F.

Note that the closure of a convex function class is convex, hence convex function classes are closure-convex.

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

The idea behind the proof is to show that if the closure of F is not convex, an agnostic algorithm for learning F to accuracy ϵ can be used to estimate the expected value of a Bernoulli random variable to accuracy $k\epsilon$ for some constant k, using the same number of examples. The following lemma (see e.g. [6] for a proof) shows that estimating the expected value of a Bernoulli random variable requires $\Omega(\ln(1/\delta)/\epsilon^2)$ examples, which implies the agnostic learning algorithm also requires $\Omega(\ln(1/\delta)/\epsilon^2)$ examples.

Define a randomized decision rule, $\hat{\alpha}$, as follows. Choose a function $\phi : \bigcup_{m=1}^{\infty}\{0,1\}^m \to [0,1]$, and for $(\xi_1, \ldots, \xi_m) \in \{0,1\}^m$ let $\hat{\alpha}(\xi_1, \ldots, \xi_m)$ be a random variable that takes values in $\{\alpha_1, \alpha_2\}$, with $\mathbf{Pr}(\hat{\alpha}(\xi_1, \ldots, \xi_m) = \alpha_1) = \phi(\xi_1, \ldots, \xi_m)$.

Lemma 3 *Let $\xi_1, \ldots, \xi_m$ be a sequence of i.i.d. $\{0,1\}$-valued random variables satisfying $\mathbf{Pr}(\xi_i = 1) = \alpha$, where $\alpha = \alpha_1 := 1/2 + \gamma/2$ with probability $1/2$ and $\alpha = \alpha_2 := 1/2 - \gamma/2$ with probability $1/2$. If some randomized decision rule $\hat{\alpha}$ satisfies $\mathbf{Pr}\left(\hat{\alpha}(\xi_1, \ldots, \xi_m) \neq \alpha\right) \leq \delta$ (where the probability is over the data sequence, the choice of α, and the randomization of the decision rule $\hat{\alpha}$), then $m = \Omega\left(\frac{\ln 1/\delta}{\gamma^2}\right)$.*

The following lemma shows that we can assume that for any probability distribution $P_{\mathcal{X}}$ and hence Hilbert space H as in Definition 1, the function class F in Theorem 2 is totally bounded.

The lemma follows from a similar proof to that of the main result in [7]. (Alternatively, it is an immediate corollary of Theorem 11 in [11] and Theorem 2 in [5].)

Lemma 4 *Let F be a class of functions mapping from $\mathcal{X}$ to some bounded subset of $\mathbb{R}$. If F is agnostically learnable, then for every distribution $P_{\mathcal{X}}$ on $\mathcal{X}$, F is totally bounded with respect to the pseudometric $\rho_{P_{\mathcal{X}}}(f_1, f_2) = \sqrt{\int_{\mathcal{X}} (f_1 - f_2)^2 dP_{\mathcal{X}}}$.*

Combining this with the following lemma shows that if F is not closure-convex then either F is not agnostically learnable, or for some distribution there is a uniformly bounded function in the Hilbert space H that has two best approximations in $\bar{F}$. (The existence of a function with two best approximations follows from standard results in approximation theory—see, for example, Corollary 3.13 in [8]—but we give a constructive proof since we also need the function to be uniformly bounded.)

Lemma 5 *Suppose that $P_{\mathcal{X}}$ is a probability distribution on $\mathcal{X}$, H is the corresponding Hilbert space, and $\mathcal{Y}'$ is a bounded interval in $\mathbb{R}$. Let $H_{\mathcal{Y}'}$ denote the set of functions f in H with $f(x) \in \mathcal{Y}'$ for all $x \in \mathcal{X}$. Let F be a totally bounded subset of $H_{\mathcal{Y}'}$. If $\bar{F}$ is not convex, there is a bounded interval $\mathcal{Y}$ in $\mathbb{R}$ and functions $c \in H_{\mathcal{Y}}$, and $f_1, f_2 \in \bar{F}$ satisfying $\|f_1 - f_2\| \neq 0$, $\|c - f_1\| = \|c - f_2\| > 0$, and for all $f \in \bar{F}$, $\|c - f\| \geq \|c - f_1\|$.*

Proof. Since $\bar{F}$ is a totally bounded, closed subset of a Hilbert space, it is compact. (See, for example, Theorem 11.3 in [13].) Since $\bar{F}$ is non-convex, there exists $g, h \in \bar{F}$ and $\alpha \in (0, 1)$ such that $f_c := \alpha g + (1 - \alpha)h$ is not in $\bar{F}$. To show that the desired functions f_1 and f_2 exist, we grow a ball around f_c until it touches a point in $\bar{F}$. The radius of this ball is $\delta := \min\{\|f - f_c\| : f \in \bar{F}\}$. (The minimum exists and is positive because $\bar{F}$ is compact.) If the set $G := \{f \in \bar{F} : \|f - f_c\| = \delta\}$ contains more than one function, we are finished. If G contains only one function f_1, we grow a ball that touches $\bar{F}$ at f_1 and has its centre on the line joining f_1 and f_c, until the ball touches another point in $\bar{F}$. That is, we set $c := tf_c + (1-t)f_1$ with the smallest $t > 1$ such that $\{f \in \bar{F} : \|f - f_1\| > 0, \|f - c\| = \|f_1 - c\|\} \neq \emptyset$. We show that such a t must exist by showing that eventually the ball must include either g or h. Now, since $\|f_1 - c\|^2 = t^2\|f_1 - f_c\|^2$, we have

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

Similarly, $\|g - c\|^2 = \|g - f_1\|^2 + \|f_1 - c\|^2 - 2t\|f_1 - f_c\|^2 + 2t\langle g - f_c, f_1 - f_c\rangle$. Now $\langle h - f_c, f_1 - f_c\rangle$ and $\langle g - f_c, f_1 - f_c\rangle$ must have opposite sign unless they are both zero. In any case, for t large enough, either $\|f_1 - c\| \geq \|h - c\|$ or $\|f_1 - c\| \geq \|g - c\|$ or both. But g and h belong to $\bar{F}$, so a suitable t exists. Since f_1 and f_c are convex combinations of functions in $\bar{F}$, it is clear that c maps to some bounded range $\mathcal{Y}$. $\square$

For a non-convex totally bounded $\bar{F}$, let $f_1, f_2 \in \bar{F}$ and $c \in H$ be as in Lemma 5. We shall use these functions to show that an agnostic learning algorithm for F can be used to estimate the expected value of a Bernoulli random variable. Let Π be the two dimensional plane $\{c + a(f_1 - c) + b(f_2 - c) : a, b \in \mathbb{R}\}$. For $0 < p < 1$ define $f_1^* := pf_1 + (1-p)c$ and $f_2^* := pf_2 + (1-p)c$. Let $f_c := (f_1^* + f_2^*)/2$ and $f_m := c + (\|f_1 - c\|/\|f_c - c\|)(f_c - c))$. Let $f_{d1} := c + (\|f_1 - c\|/\|f_1^* - f_c\|)(f_1^* - f_c)$ and $f_{d2} := c + (\|f_1 - c\|/\|f_2^* - f_c\|)(f_2^* - f_c)$. The constructed functions in Π are illustrated in Figure 1. Note that $f_{d1} - c$ and $f_{d2} - c$ are orthogonal to $f_m - c$.

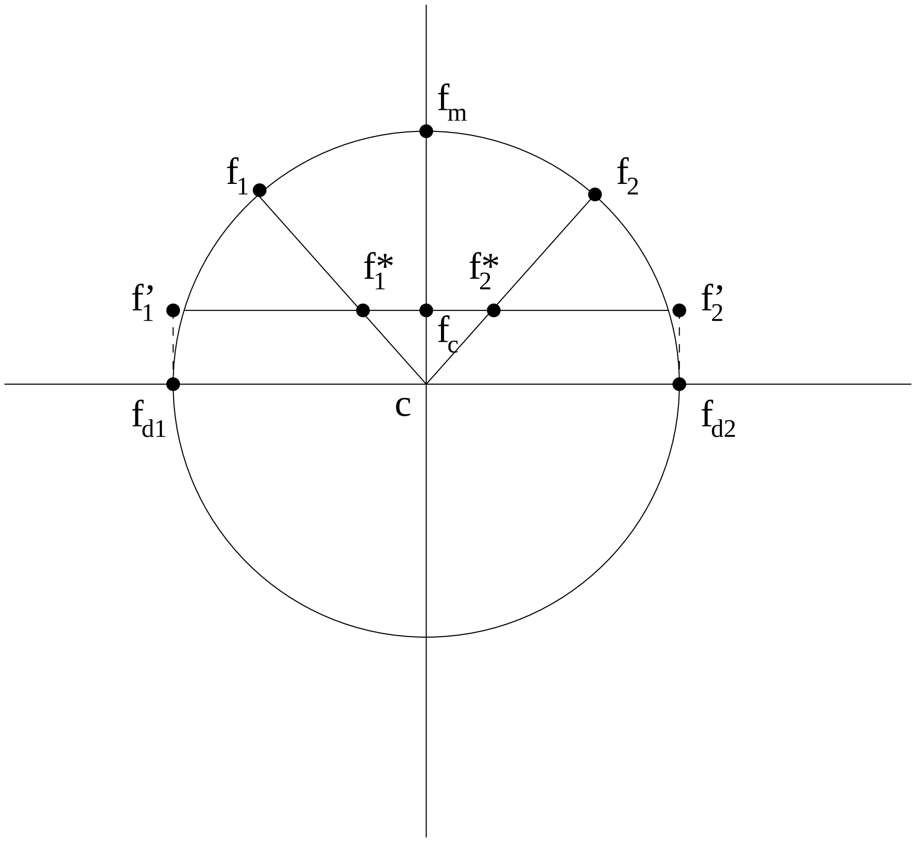

Figure 1: Function class with labelled functions in two dimensions.

Let γ be such that $f_1^* = f_c + \gamma(f_{d_1} - c)$. (Such a γ exists because $f_{d_1} - c = (\|f_1 - c\|/\|f_1^* - f_c\|)(f_1^* - f_c)$ and $f_1^* - f_c$ are collinear.) It is easy to show that $f_2^* = f_c - \gamma(f_{d1} - c)$. The following lemma shows how γ depends on p and on $\epsilon := \|f_m - f_1^*\|^2 - \|f_1 - f_1^*\|^2$, which we shall use in analyzing the agnostic learning algorithm. Notice that ϵ is proportional to γ, and we shall use this value of γ in Lemma 3. The proofs of Lemma 6 and Lemma 7 can be found in the Appendix.

Lemma 6

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

The following lemma will be used to show that if we ensure that either f_1^* or f_2^* is the conditional expectation of y given x, then an agnostic learning algorithm can be used to select between them.

Lemma 7 *Let f_1, f_2, f_1^*, f_2^*, ϵ and the function class F be as defined above. Then for any $\hat{f} \in F$ and $\epsilon > 0$,*

$$\|\hat{f} - f_1^*\|^2 - \|f_1 - f_1^*\|^2 < \epsilon \Rightarrow \|\hat{f} - f_1\| < \|\hat{f} - f_2\| \tag{1}$$

and

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

We can now prove Theorem 2.

Proof (Theorem 2). Let F be a class that is not closure-convex, and suppose that there is an agnostic learning algorithm A for F. Then for any probability distribution on $\mathcal{X} \times \mathcal{Y}$, the algorithm draws m examples and with probability at least $1 - \delta$, it produces $\hat{f} \in F$ such that $\|\hat{f} - f^*\|^2 - \|f_a - f^*\|^2 \leq \epsilon$, where $f^*(x) := \mathbf{E}[Y|X = x]$ and $f_a \in \bar{F}$ is such that $\|f_a - f^*\| = \inf_{f \in F} \|f - f^*\|$. The function f_a is a best approximation in $\bar{F}$ to f^*.

We now argue that if the sample complexity of Algorithm A (for accuracy $\epsilon/2$ and confidence $1 - \delta$) is m, then there is a probability distribution and an algorithm, Algorithm B, (which depends on F and the probability distribution) which, with probability $1-\delta$ over m examples, solves the problem described in Lemma 3, for γ which depends on ϵ according to Lemma 6.

Let $P_{\mathcal{X}}$ be a probability distribution on $\mathcal{X}$ for which $\bar{F}$ is not convex. Then we can assume that F is totally bounded, since if it is not, Lemma 4 implies that the sample complexity is infinite. So we can define c, f_1, f_2, f_1^*, f_2^*, f_{d1}, f_{d2}, ϵ, and γ as above. Let $f_1' := (f_1^* + f_2^*)/2 + (f_{d1} - c)$ and $f_2' := (f_1^* + f_2^*)/2 + (f_{d2} - c)$.

Algorithm B, in solving the problem described in Lemma 3, receives as input a sequence $\xi_1, \ldots, \xi_m$. It then generates a sequence $x_1, \ldots, x_m \in \mathcal{X}$ independently according to $P_{\mathcal{X}}$. For $i = 1, \ldots, m$, if $\xi_i = 1$, Algorithm B passes $(x_i, f_1'(x_i))$ to Algorithm A; otherwise it passes $(x_i, f_2'(x_i))$. Clearly, the target conditional expectation is f_1^* if $\alpha = \alpha_1$ and f_2^* if $\alpha = \alpha_2$. Suppose that Algorithm A returns $\hat{f}$. If $\|\hat{f} - f_1\| < \|\hat{f} - f_2\|$ then Algorithm B chooses $\alpha = \alpha_1$; otherwise it chooses $\alpha = \alpha_2$.

Now, since A is an agnostic learning algorithm, with probability at least $1-\delta$, if $\alpha = \alpha_1$, $\|\hat{f} - f_1^*\|^2 \leq \|f_1 - f_1^*\|^2 + \epsilon/2$, and if $\alpha = \alpha_2$, $\|\hat{f} - f_2^*\|^2 \leq \|f_2 - f_2^*\|^2 + \epsilon/2$. Lemma 7 ensures that in either case Algorithm B guesses correctly. But Lemma 3 shows that obtaining the correct α with probability $1-\delta$ in this way requires $\Omega(\ln(1/\delta)/\gamma^2)$ examples, which implies that Algorithm A also requires at least $\Omega(\ln(1/\delta)/\gamma^2) = \Omega(\ln(1/\delta)/\epsilon^2)$ examples.

Notice that the definitions of f_1' and f_2' imply that they have a bounded range which depends only on the class F. $\square$

4 Learning Convex Classes

In this section, we look at upper bounds on the sample complexity for learning. We show that if a function class F is uniformly bounded, has finite pseudo-dimension, and is closure-convex then the sample complexity for learning F is $O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon} + \ln\frac{1}{\delta}\right)\right)$. We show that for a uniformly bounded non-convex function class with finite pseudo-dimension, the sample complexity for agnostically learning the convex hull of the class is $O\left(\frac{1}{\epsilon}\left(\frac{1}{\epsilon}\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$, which is not significantly worse (constant and log factors) than agnostically learning the function class itself.

In the following we make use of an assumption called permissibility. This is a measurability condition satisfied by most function classes used for learning. See [19, 21] for details.

Theorem 8 *Suppose F is permissible, has finite pseudo-dimension, and is uniformly bounded. Then*

1. *If F is closure-convex, the sample complexity of agnostically learning F is $O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$.*
2. *The sample complexity of agnostically learning the convex hull of F is*
   $O\left(\frac{1}{\epsilon}\left(\frac{1}{\epsilon}\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$.

The proof of Theorem 8 uses the following result which is taken from [15] with minor modification (and almost identical proof).

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

Theorem 9 *Let $F=\bigcup_{k=1}^\infty F_k$ be a closure-convex class of real-valued functions defined on $\mathcal{X}$ such that each F_k is permissible and $|f(x)|\leq B$ for all $f\in F$ and $x\in\mathcal{X}$. Suppose $\mathcal{Y}\subseteq[-B,B]$, and let 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 $\langle f,g\rangle=\int f(x)g(x)dP_{\mathcal{X}}(x)$. Let $C=\max\{B,1\}$. Assume $\nu,\nu_c>0$ and $0<\alpha\leq 1/2$. Let $f^*(x)=\mathbf{E}[Y|X=x]$ and $g_f(x,y)=(y-f(x))^2-(y-f_a(x))^2$ where $f_a=\operatorname{argmin}_{f\in\bar{F}}\int(f(x)-f^*(x))^2dP_{\mathcal{X}}(x)$. Then for $m\geq 1$ and each k,*

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

where P^m denotes the m-fold product probability measure.

We use the following result (see [15]) to bound the number of terms in the convex combination needed to achieve a desired accuracy. The theorem is an extension of the results of Barron [4] and Jones [10]. We can apply this result to a uniformly bounded function class, considered as a subset of a Hilbert space of the type defined in Definition 1, since any function g with range in $[-B, B]$ satisfies $\|g\| \leq B$.

Theorem 10 *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

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

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

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

To bound the covering number, we need the following result. (See [19, 9].)

Lemma 11 *Let F be a class of functions from a set $\mathcal{Z}$ into $[-M, M]$ where $M > 0$, and suppose $\dim_P(F) = d$ for some $1 \leq d < \infty$. Then for all $0 < \epsilon \leq 2M$ and any finite sequence $\bar{z}$ of points in $\mathcal{Z}$,*

$$N(\epsilon, F_{|\bar{z}}, d_{l_1}) < 2 \left(\frac{4eM}{\epsilon} \ln \frac{4eM}{\epsilon} \right)^d .$$

Let $\mathcal{N}_k^F$ be the class of functions consisting of convex combinations of k functions from F with the convex coefficients given by the iterative procedure suggested by Theorem 10. That is, $\mathcal{N}_1^F = F$, and for $k > 1$,

$$\mathcal{N}_k^F = \left\{ \alpha_k f_{k-1} + (1 - \alpha_k) g : g \in F,\ f_{k-1} \in \mathcal{N}_{k-1}^F \right\},$$

where $\alpha_k = 1 - 2/(k+1)$.

Lemma 12 *Suppose the pseudo-dimension of a class F of functions mapping $\mathcal{X}$ to $[-B, B]$ is d. Then for any positive integer m and $\bar{x} \in \mathcal{X}^m$, we have*

$$N(\epsilon, \mathcal{N}^F_{k\ |\bar{x}}, d_{l_1}) \leq 2^k \left(\frac{4eB}{\epsilon} \ln \frac{4eB}{\epsilon} \right)^{kd}.$$

Proof. Let $f = \sum_{i=1}^k \beta_i f_i$ be an arbitrary function in $\mathcal{N}^F_{k\ |\bar{x}}$. Let U be an ϵ-cover for $F_{|\bar{x}}$. For each f_i, pick a $g_i \in U$ such that $d_{l_1}(f_i, g_i) \leq \epsilon$. Let $g = \sum_{i=1}^k \beta_i g_i$. Obviously, $d_{l_1}(f, g) \leq \epsilon$. From Lemma 11, we have $|U| < 2\left(\frac{4eB}{\epsilon} \ln \frac{4eB}{\epsilon}\right)^d$. The definition of $\mathcal{N}^F_k$ implies that the coefficients β_i are fixed, and so there are only $|U|^k$ ways to select the function g, which implies the result. □

We now give the proof of Theorem 8.

Proof (Theorem 8). For the first part of the theorem, F is uniformly bounded and closure-convex, and has finite pseudo-dimension d. We can thus use Theorem 9 and we set $F_k = F$ for all positive integers k. Then we can bound the covering number in (3) using Lemma 11. Rescale the function class and target random variable by dividing by B. (This rescaling trick gives a B^2 term instead of a B^4 term in the sample complexity bound.) The ϵ covering number of the scaled function class is the same as the $B\epsilon$ covering number of the unscaled function class. We will work with the scaled function class and target random variable. To get the correct accuracy when the function is scaled back to the original scale, we need to learn to accuracy ϵ/B^2. Let $\alpha = 1/2$ and $\nu = \nu_c = \epsilon/(2B^2)$. As the estimator, choose a function $\hat{f}$ in $\bar{F}$ that has $\hat{\mathbf{E}}_{\bar{z}}(g_{\hat{f}}) \leq \epsilon/(4B^2)$. (Since $g_f(x, y) = (y - f(x))^2 - (y - f_a(x))^2$ and $f_a \in \bar{F}$, this is always possible.) With this choice of $\hat{f}$, if $\mathbf{E}(g_{\hat{f}}) \geq \epsilon/B^2$ then

$$\frac{\mathbf{E}(g_{\hat{f}}) - \hat{\mathbf{E}}_{\bar{z}}(g_{\hat{f}})}{\nu + \nu_c + \mathbf{E}(g_{\hat{f}})} \geq \alpha.$$

Theorem 9 gives a bound on the probability that this occurs. Setting the right hand side of (3) to

δ, we get

$$12\left(\frac{2048eB^2}{\epsilon}\ln\frac{2048eB^2}{\epsilon}\right)^d \exp(-3\epsilon m/(20992B^2)) = \delta.$$

This means that

$$m \geq \frac{20992B^2}{3\epsilon}\left(d\ln\left(\frac{2048B^2}{\epsilon}\ln\frac{2048B^2}{\epsilon}\right) + \ln\frac{12}{\delta}\right)$$

ensures that the probability that $\mathbf{E}(g_{\hat{f}}) \geq \epsilon/B^2$ is no more than δ, so the sample complexity is $O\left(\frac{1}{\epsilon}\left(d\ln\frac{1}{\epsilon} + \ln\frac{1}{\delta}\right)\right)$.

For the second part of the theorem, we set $F_k = \mathcal{N}_k^F$, and again we rescale the function class and target random variable by dividing by B. Let $\alpha = 1/2$ and use Theorem 9 and Lemma 12 with $\nu = \nu_c = \epsilon/(4B^2)$ to get

$$\begin{aligned}
P^m \Big\{\bar{z} \in \mathcal{Z}^m : \exists f \in \mathcal{N}_k^F, \mathbf{E}[(y - f(x))^2 - (y - f_a(x))^2] \geq \\
2\hat{\mathbf{E}}_{\bar{z}}[(y - f(x))^2 - (y - f_a(x))^2] + \epsilon/(2B^2)\Big\} \\
\leq \sup_{\bar{z} \in \mathcal{Z}^{2m}} 6N\left(\frac{\epsilon}{1024B^2}, \mathcal{N}^F_{k_{|\bar{z}}}, d_{l_1}\right)\exp(-3\epsilon m/(41984B^2)) \\
\leq 6 \times 2^k \left(\frac{4096eB^2}{\epsilon}\ln\frac{4096eB^2}{\epsilon}\right)^{kd} e^{-3\epsilon m/(41984B^2)}.
\end{aligned} \tag{5}$$

The learning algorithm we consider chooses a function from $\mathcal{N}_k^F$ in the iterative way suggested by Theorem 10. For k sufficiently large, the empirical error of this function is close to the minimum empirical error over $\mathrm{co}(\bar{F})$, which implies that $2\hat{\mathbf{E}}_{\bar{z}}[(y - \hat{f}_k(x))^2 - (y - f_a(x))^2]$ is small. In particular, we shall ensure that this quantity is no more than $\epsilon/(2B^2)$. Combining this with the above inequality, shows that, with high probability, $\mathbf{E}g_{\hat{f}_k} < \epsilon/B^2$ as desired.

Now, for $z_i = (x_i, y_i) \in \mathcal{Z}$ and $\bar{z} = (z_1, \ldots, z_m)$, we let $\hat{\mathbf{E}}_{\bar{z}}[\phi(x,y)]$ denote the expectation of ϕ under the empirical measure, as above. Define $f'(x) = \hat{\mathbf{E}}_{\bar{z}}[Y|X = x]$ in the obvious way. Let $\hat{f}_a$ be the function in the convex closure which minimizes the empirical error. Then $\hat{\mathbf{E}}_{\bar{z}}[(y - \hat{f}_k(x))^2 - (y - f_a(x))^2] = \hat{\mathbf{E}}_{\bar{z}}[(f'(x) - \hat{f}_k(x))^2 - (f'(x) - f_a(x))^2]$. The definition of $\hat{f}_a$ implies that

$\hat{\mathbf{E}}_{\bar{z}}[(f'(x)-\hat{f}_k(x))^2-(f'(x)-f_a(x))^2] \le \hat{\mathbf{E}}_{\bar{z}}[(f'(x)-\hat{f}_k(x))^2-(f'(x)-\hat{f}_a(x))^2]$. That is,

$$\hat{\mathbf{E}}_{\bar{z}}[(y-\hat{f}_k(x))^2-(y-f_a(x))^2] \le \hat{\mathbf{E}}_{\bar{z}}[(f'(x)-\hat{f}_k(x))^2-(f'(x)-\hat{f}_a(x))^2].$$

It follows that if we choose $\hat{f}_k$ such that this latter quantity is no more than $\epsilon/(4B^2)$, we have $\mathbf{E}g_{\hat{f}_k} < \epsilon/B^2$. To do this, we apply Theorem 10 with $b^2=1$, $c=2$, and with inner product weighted by the empirical measure. If we choose $\hat{f}_k$ as described in that theorem, and $k=\lceil 32B^2/\epsilon\rceil$, we get the desired result.

Setting the right hand side of (5) to δ, we get

$$k\ln 2 + kd\ln\left(\frac{4096eB^2}{\epsilon}\ln\frac{4096eB^2}{\epsilon}\right) - \frac{3\epsilon m}{41984B^2} = \ln\frac{\delta}{6}.$$

Rearranging,

$$m \quad = \quad \frac{41984B^2}{3\epsilon}\left(kd\ln\left(\frac{4096eB^2}{\epsilon}\ln\frac{4096eB^2}{\epsilon}\right)+k\ln 2+\ln\frac{6}{\delta}\right)$$

will suffice. Substituting for k, the sample complexity is

$$O\left(\frac{1}{\epsilon}\left(\frac{d}{\epsilon}\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right).$$

□

5 Discussion

We have shown that the sample complexity of agnostic learning classes of uniformly bounded functions is bounded by $O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$ if the function class used for learning is convex and has finite pseudo-dimension, but is at least $\Omega(\ln(1/\delta)/\epsilon^2)$ if the closure of the function class is not convex. Furthermore, for non-convex function classes (with finite pseudo-dimension), the sample complexity of learning the convex hull is $O\left(\frac{1}{\epsilon}\left(\frac{1}{\epsilon}\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$.

For some function classes, the rate of growth of the sample complexity of agnostic learning actually improves when learning the convex hull of the function class instead of the function class itself. For example, for any finite class of functions F, the pseudo-dimension of the convex hull is no more than $|F|$. Since finite classes of size at least two are not convex, the sample complexity for agnostically learning F is $\Omega((\ln 1/\delta)/\epsilon^2)$ as opposed to the sample complexity for agnostically learning the convex hull of F, which is $O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$. Hence, for these function classes, using the convex hull for learning gives better approximation capabilities as well as smaller sample complexity for agnostic learning. However, in general, the pseudo-dimension of the class of convex combinations of a function class can only be bounded in terms of the number of terms in the convex combinations. Barron [3] has shown that for linear threshold functions defined on $\mathbb{R}^n$, the sample complexity of learning the convex hull using any estimator cannot be better than $\Omega(1/\epsilon^{(2d+2)/(d+2)})$. Since the pseudo-dimension of linear threshold functions is $d+1$, our sample complexity result of $O\left(\frac{1}{\epsilon}\left(\ln\frac{1}{\epsilon}+\ln\frac{1}{\delta}\right)\right)$ is close to optimal (with respect to ϵ) for learning the convex hull of linear threshold functions when d is large. Hence, in general, we cannot hope to get much better sample complexity bounds for the convex hull. (In fact, the exponent on ϵ could be improved to match Barron's lower bound if one could generalize the improvement over Theorem 10 obtained by Makovoz [17] to the agnostic case, where the conditional expectation is not in the class F. Ideally a constructive generalization may be found, which would have the same advantage of Theorem 10 which is constructive, and hence can be used to define an agnostic learning algorithm for the convex hull that has computational cost not significantly larger than that of agnostically learning the function class itself [14].)

In summary, for agnostic learning, using the convex hull may sometimes greatly improve the performance of the estimators (because of the better approximation) without much penalty in terms of sample complexity (the sample complexity may even be much improved in some cases).

6 Acknowledgements

We would like to thank Andrew Barron for asking the questions which led to the results in this paper. Thanks also to Tamás Linder for helpful suggestions, and to the reviewers for helpful remarks concerning the presentation and proofs. This research was partially supported by the Australian Research Council.

Appendix

Proof (Lemma 6).

Note that $f_1^* - f_c$ is the projection of $pf_1 + (1-p)c - c$ in the direction of $f_{d1} - c$. Thus

$$\begin{aligned}\gamma(f_{d1} - c) &= f_1^* - f_c \\ &= \frac{\langle pf_1 + (1-p)c - c, f_{d1} - c\rangle}{\|f_{d1} - c\|\|f_{d1} - c\|}(f_{d1} - c) \\ &= \frac{p\langle f_1 - c, f_{d1} - c\rangle}{\|f_{d1} - c\|^2}(f_{d1} - c). \qquad (6)\end{aligned}$$

which gives the first equality. To prove the second, first notice that $f_c - c$ is the projection of $f_1^* - c$ in the direction of $f_m - c$, and so

$$\begin{aligned}f_c - c &= \frac{\langle f_1^* - c, f_m - c\rangle}{\|f_m - c\|^2}(f_m - c) \\ &= \frac{\langle pf_1 + (1-p)c - c, f_m - c\rangle}{\|f_m - c\|^2}(f_m - c) \\ &= \frac{p\langle f_1 - c, f_m - c\rangle}{\|f_m - c\|^2}(f_m - c). \qquad (7)\end{aligned}$$

Note also that

$$\|f_1 - f_1^*\|^2 = \|f_1 - pf_1 - (1-p)c\|^2 = (1-p)^2\|f_1 - c\|^2 = (1 - 2p + p^2)\|f_1 - c\|^2. \qquad (8)$$

With that, by Pythagoras Theorem

$$\|f_m - f_1^*\|^2 = \|f_1^* - f_c\|^2 + \|f_m - f_c\|^2$$

$$
\begin{aligned}
&= \|f_1^* - f_c\|^2 + \|f_m - c + c - f_c\|^2 \\
&= \frac{p^2\langle f_1 - c, f_{d1} - c\rangle^2}{\|f_{d1} - c\|^2} + \|f_m - c\|^2 + \frac{p^2\langle f_1 - c, f_m - c\rangle^2}{\|f_m - c\|^2} + 2\langle f_m - c, c - f_c\rangle \\
&\qquad \text{(by (6) and (7))} \\
&= p^2\|f_1 - c\|^2 + \|f_m - c\|^2 - \frac{2p\langle f_1 - c, f_m - c\rangle}{\|f_m - c\|^2}\|f_m - c\|^2 \qquad (9)
\end{aligned}
$$

and the last equality holds because $\frac{p^2\langle f_1-c,f_{d1}-c\rangle^2}{\|f_{d1}-c\|^2} + \frac{p^2\langle f_1-c,f_m-c\rangle^2}{\|f_m-c\|^2} = p^2\|f_1 - c\|^2$ and $f_c - c = \frac{\langle f_1^*-c,f_m-c\rangle(f_m-c)}{\|f_m-c\|^2} = \frac{p\langle f_1-c,f_m-c\rangle(f_m-c)}{\|f_m-c\|^2}$.

From the construction, $\|f_1 - c\|^2 = \|f_m - c\|^2$. Recalling the definition of $\epsilon = \|f_m - f_1^*\|^2 - \|f_1 - f_1^*\|^2$, subtracting (8) from (9), and rearranging we obtain

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

and hence

$$p = \frac{\epsilon}{2\left(1 - \frac{\langle f_1-c,f_m-c\rangle}{\|f_m-c\|^2}\right)\|f_m - c\|^2}. \qquad (10)$$

From the first equality of the claim, Equation (10), and the fact that $\|f_{d1} - c\|^2 = \|f_m - c\|^2$, we infer

$$
\begin{aligned}
\gamma &= \frac{p\langle f_1 - c, f_{d1} - c\rangle}{\|f_{d1} - c\|^2} \\
&= \frac{\epsilon}{2\left(1 - \frac{\langle f_1-c,f_m-c\rangle}{\|f_m-c\|^2}\right)\|f_m - c\|^2}\frac{\langle f_1 - c, f_{d1} - c\rangle}{\|f_{d1} - c\|^2} \\
&= \frac{\epsilon\langle f_1 - c, f_{d1} - c\rangle}{2\left(1 - \frac{\langle f_1-c,f_m-c\rangle}{\|f_m-c\|^2}\right)\|f_m - c\|^4}. \qquad (11)
\end{aligned}
$$

□

Proof (Lemma 7)

Recall that $\|f_m - f_1^*\|^2 - \|f_1 - f_1^*\|^2 = \epsilon$. We show that if $\|\hat{f} - f_2\| \leq \|\hat{f} - f_1\|$ then $\|\hat{f} - f_1^*\| \geq \|f_m - f_1^*\|$, and this implies $\|\hat{f} - f_1^*\|^2 - \|f_1 - f_1^*\|^2 \geq \epsilon$.

We have

$$\|\,\hat{f} - f_1^*\|^2 = \|\hat{f} - f_c + f_c - f_1^*\|^2$$

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

where the inequality follows from the fact that $\hat{f}$ is in F (since f_m is the closest point in $\bar{F}$ to f_c) and the last equality is Pythagoras' theorem. So we need only show that the second term is greater than or equal to zero when $\|\hat{f} - f_2\| \leq \|\hat{f} - f_1\|$.

We have

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

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

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

References

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

[2] 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.

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

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

[5] Peter L. Bartlett, Sanjeev R. Kulkarni, and S. Eli Posner. Covering numbers for real-valued function classes. *IEEE Transactions on Information Theory*, 1997. (to appear).

[6] S. Ben-David and M. Lindenbaum. Learning distributions by their density levels—a paradigm for learning without a teacher. In *Computational Learning Theory: EUROCOLT'95*, pages 53–68, 1995.

[7] G. Benedek and A. Itai. Learnability with respect to fixed distributions. *Theoret. Comput. Sci.*, 86(2):377–389, 1991.

[8] D. Braess. *Nonlinear Approximation Theory.* Springer-Verlag, 1986.

[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] 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.

[11] M. J. Kearns and R. E. Schapire. Efficient distribution-free learning of probabilistic concepts. In *Proc. of the 31st Symposium on the Foundations of Comp. Sci.*, pages 382–391. IEEE Computer Society Press, Los Alamitos, CA, 1990.

[12] M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2):115–141, 1994.

[13] A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis*. Dover, 1970.

[14] 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.

[15] W. S. Lee, P. L. Bartlett, and R. C. Williamson. Efficient agnostic learning of neural networks with bounded fan-in. *IEEE Transactions on Information Theory*, 42(6):2118–2132, 1996.

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

[17] Y. Makovoz. Random approximants and neural networks. *Journal of Approximation Theory*, 85:98–109, 1996.

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

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

[20] D. Pollard. Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics*, 22(3):271–278, 1995.

[21] A.W. van der Vaart and J.A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

List of Figures