

Probabilistic ‘Generalization’ of Functions and Dimension-based Uniform Convergence Results

Martin Anthony
Department of Mathematics
London School of Economics and Political Science
Houghton Street
London WC2A 2AE
`m.anthony@lse.ac.uk`

Abstract

In this largely expository article, we highlight the significance of various types of ‘dimension’ for obtaining uniform convergence results in probability theory and we demonstrate how these results lead to certain notions of generalization for classes of binary-valued and real-valued functions. We also present new results on the generalization ability of certain types of artificial neural networks with real output.

1 Introduction

There are many approaches to the notion of ‘generalization’ in theories of machine learning and psychology. In this article, we review several mathematical approaches and we present some new results on the generalisation ability of

simple artificial neural networks. We start by describing one popular mathematical approach to generalization in the context of concept learning. We emphasise how this model of generalization relates to results in probability theory. We then examine how one might extend this notion of generalization to a more general framework, that of generalizing real functions. Again, we link these models with results in probability theory. We demonstrate concrete applications of one of the models by presenting some new results on the generalization ability of certain types of neural networks.

2 PAC-generalization

The models of generalization we discuss in this article are based on what has become known as the ‘probably approximately correct’, or PAC, model of computational learning theory (or statistical learning theory). This model was introduced by Valiant [20]. In Valiant’s formulation, much stress was placed on the computational complexity of learning algorithms, which is not something we shall address here. The main probabilistic tools which have become useful for the analysis of this model and its variants have their roots in the work of Vapnik and others (see [21, 23, 22]). The books [4, 14, 16] contain general discussions of PAC learning.

In its simplest form, the PAC model of learning may be described as follows. There is a set of *examples* X , and a *target function* $t : X \rightarrow \{0, 1\}$. It is known that t belongs to some set \mathcal{C} of functions, but that is all that is known about it. There is assumed to be some fixed (but unknown) probability measure μ on the set X of examples. (More precisely, we mean a probability measure defined on a fixed σ -algebra Σ of subsets of X . It is usually assumed that X is a complete separable metric space, in which case we take Σ to be the Borel algebra on X .) In this framework, a ‘learner’ receives a *training sample*

$$\mathbf{s} = \mathbf{x}(t) = ((x_1, t(x_1)), (x_2, t(x_2)), \dots, (x_m, t(x_m))) \in (X \times \{0, 1\})^m,$$

a sequence of labelled examples. For the PAC model, it is assumed that the examples x_1, x_2, \dots are drawn independently at random from X , according to μ . The aim is to find a good approximation to t from a set \mathcal{H} of functions (possibly different from \mathcal{C}). The class \mathcal{H} must satisfy some measurability conditions in order for the following definitions and results to be valid; these conditions are quite natural and a discussion may be found in [8]. A *learning algorithm* is a mapping L from samples of the form $\mathbf{x}(t)$, where $\mathbf{x} \in \bigcup_{m=1}^{\infty} X^m$ and $t \in \mathcal{C}$. Generally, the *error* of $h \in \mathcal{H}$ with respect to t and μ is defined to be $\text{er}_{\mu}(h) = \mu(\{x \in X : h(x) \neq t(x)\})$. With these notations, the definition of PAC learning is as follows.

Definition 2.1 Suppose that L is a learning algorithm as described above. We say that L is probably approximately correct (or PAC) if, given $\epsilon, \delta \in (0, 1)$, there is $m_L(\epsilon, \delta)$ such that for any probability measure μ on X and any target function $t \in \mathcal{C}$,

$$\mu^m(\{\mathbf{x} \in X^m : \text{er}_\mu(L(\mathbf{x}(t))) > \epsilon\}) < \delta,$$

for $m \geq m_L(\epsilon, \delta)$. (In other words, with probability at least $1 - \delta$, for a large enough sample, $L(\mathbf{x}(t))$ has error less than ϵ .)

Note that a PAC learning algorithm must work for every probability measure μ and every possible target function t , and that $m_L(\epsilon, \delta)$ depends on neither the probability measure nor the target. PAC learning is therefore ‘distribution-independent’ learning.

If $\mathcal{C} \subseteq \mathcal{H}$ then it is possible (and perhaps sensible) to use a learning algorithm L with the property that the function $h = L(\mathbf{x}(t))$ satisfies $h(x_i) = t(x_i)$ for $1 \leq i \leq m$; in other words, the output function of the learning algorithm agrees with the target function on the examples it saw during training. We say that such an L is *consistent* [8]. We make the following definition.

Definition 2.2 Let \mathcal{H} be a set of functions from some set X to $\{0, 1\}$. Then \mathcal{H} PAC-generalizes if for all $\epsilon, \delta \in (0, 1)$ there is $m_0(\epsilon, \delta)$ such that, for any $t : X \rightarrow \{0, 1\}$ and any probability measure μ on X , if $m \geq m_0(\epsilon, \delta)$ then, with μ^m -probability at least $1 - \delta$, a sample $\mathbf{x} \in X^m$ is such that the following holds:

$$\begin{aligned} & h \in \mathcal{H} \text{ and } h(x_i) = t(x_i) \text{ for } i = 1, 2, \dots, m \\ \implies & \text{er}_\mu(h) = \mu(\{x \in X : h(x) \neq t(x)\}) < \epsilon. \end{aligned}$$

It is clear that if \mathcal{H} PAC-generalizes, if $\mathcal{C} \subseteq \mathcal{H}$, and if L is a consistent learning algorithm, then L is a PAC learning algorithm; see [8, 4].

More generally, when it is not the case that $\mathcal{C} \subseteq \mathcal{H}$, or when \mathcal{C} is unknown, it will be impossible to find a consistent learning algorithm. To deal with this (and with other considerations, such as there being no well-defined target function), the model can be extended by considering probability measures on $X \times \{0, 1\}$, rather than probability measures on X coupled with functions from X to $\{0, 1\}$. (Any probability measure μ on X together with a function $t : X \rightarrow \{0, 1\}$ can be represented in the obvious way by a probability measure P on $X \times \{0, 1\}$; see [8, 5].) In this more general model (which is discussed in [8, 10], for example), the error of $h \in \mathcal{H}$ with respect to a probability measure P on $X \times \{0, 1\}$ is taken to be

$$\text{er}_P(h) = P(\{(x, y) \in X \times \{0, 1\} : h(x) \neq y\}).$$

In this context, a learning algorithm takes as input a P^m -random sample

$$\mathbf{s} = ((x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)).$$

In this more general setting there may be no function in \mathcal{H} with zero error, so we modify slightly the aim of learning: we now hope to produce a function $h \in \mathcal{H}$ with near-minimal error with respect to the measure P .

Definition 2.3 *A learning algorithm L is said to be probably approximately optimal if for any $\epsilon, \delta \in (0, 1)$ there is $m_L(\epsilon, \delta)$ such that, given any probability measure P on $X \times \{0, 1\}$, if $m \geq m_L(\epsilon, \delta)$ then with P^m -probability at least $1 - \delta$, $\mathbf{s} \in (X \times \{0, 1\})^m$ is such that*

$$\text{er}_P(L(\mathbf{s})) < \inf_{h \in \mathcal{H}} \text{er}_P(h) + \epsilon.$$

(A more general definition than this can be given, where the aim is to produce a function whose error is almost as small as the smallest one could hope to find in a class \mathcal{F} , which might be different from \mathcal{H} . This is known as *agnostic learning* and is not something we shall discuss further here. See [13, 15].)

By analogy with a consistent algorithm for the simple PAC framework discussed earlier, it might be thought that a good approach to the present problem is to use a learning algorithm L which chooses $h \in \mathcal{H}$ which appears, on the basis of the sample \mathbf{s} , to have small error. Formally, we define the *observed error* of $h \in \mathcal{H}$ on a sample \mathbf{s} to be

$$\text{er}_{\mathbf{s}}(h) = \frac{1}{m} |\{i : h(x_i) \neq y_i\}|,$$

and we should hope that if \mathbf{s} is large enough then, with high probability, the observed error is close to the (actual) error of h . We make the following more general definition of PAC-generalization.

Definition 2.4 *Let \mathcal{H} be a set of functions from X to $\{0, 1\}$. Then \mathcal{H} PAC-generalizes if for any $\epsilon, \delta \in \{0, 1\}$, there is $m_0(\epsilon, \delta)$ such that for any probability measure P on $X \times \{0, 1\}$,*

$$P^m \left(\left\{ \mathbf{s} \in (X \times \{0, 1\})^m : \sup_{h \in \mathcal{H}} |\text{er}_{\mathbf{s}}(h) - \text{er}_P(h)| \geq \epsilon \right\} \right) < \delta$$

for $m \geq m_0(\epsilon, \delta)$. (In other words, for $m \geq m_0(\epsilon, \delta)$, with probability at least $1 - \delta$, if we take a random sample drawn according to P then for every $h \in \mathcal{H}$, the observed error and the (actual) error differ by less than ϵ .)

It can be seen fairly easily that if \mathcal{H} PAC-generalizes, then a learning algorithm L which chooses h minimising the observed error on the training sample is probably approximately optimal; see [10]. (It is clear that if this definition holds true for \mathcal{H} then so does the earlier definition, Definition 2.2, of PAC-generalization. In this sense, the present definition subsumes the previous one.)

3 Uniform convergence of probabilities

Blumer *et al.* [8] observed that ‘uniform convergence’ results in probability theory, such as those obtained by Vapnik and Chervonenkis [23], are immediately applicable to PAC-generalization.

Let Z be a set and Λ a σ -algebra of subsets of Z . Suppose that $\mathcal{E} \subseteq \Lambda$ is a set of events. For $\mathbf{z} = (z_1, z_2, \dots, z_m) \in Z^m$, denote by $\hat{P}_{\mathbf{z}}(A)$ the relative frequency of event $A \in \mathcal{E}$ on the sample \mathbf{z} ,

$$\hat{P}_{\mathbf{z}}(A) = \frac{1}{m} |\{i : z_i \in A\}|.$$

(Although the relative frequency, as defined, does not depend explicitly on a probability measure P as the notation would suggest, this notation is useful since we shall be interested in the relative frequencies of events on P -random samples.) Standard results in probability theory assure us that, given any $A \in \mathcal{E}$, the relative frequency $\hat{P}_{\mathbf{z}}(A)$ converges in probability to the probability $P(A)$. However, we shall require something far stronger. One says that the *relative frequencies (of events in \mathcal{E}) converge uniformly to probabilities* if there is a function $f(m, \epsilon)$ with the property that for every $\epsilon \in (0, 1)$, $f(m, \epsilon) \rightarrow 0$ as $m \rightarrow \infty$, and which is such that

$$\forall P, P^m \left(\left\{ \mathbf{z} \in Z^m : \sup_{A \in \mathcal{E}} |\hat{P}_{\mathbf{z}}(A) - P(A)| \geq \epsilon \right\} \right) < f(m, \epsilon),$$

where ‘ $\forall P$ ’ means for all probability measures P on (Z, Λ) . Note that there are two senses in which the convergence is uniform: the rate of convergence of $\hat{P}_{\mathbf{z}}(A)$ to $P(A)$ can be bounded by a quantity $f(m, \epsilon)$ which is independent of both the probability measure P and the event $A \in \mathcal{E}$.

It is fairly straightforward to see how such a uniform convergence result implies PAC-generalization (an observation made, for example, in [8, 10]). For (using the notation of the previous section), take $Z = X \times \{0, 1\}$ and take Λ to be the product σ -algebra $\Sigma \times 2^{\{0,1\}}$. For each $h \in \mathcal{H}$, let

$$E_h = \{(x, y) \in X \times \{0, 1\} : h(x) \neq y\}$$

and let $\mathcal{E} = \{E_h : h \in \mathcal{H}\}$. With P as in the discussion of PAC-generalization, $P(E_h) = \text{er}_P(h)$ and, for $\mathbf{s} \in Z^m$, $\hat{P}_{\mathbf{s}}(E_h) = \text{er}_{\mathbf{s}}(h)$. Now, if the relative frequencies of the events in \mathcal{E} converge uniformly to their probabilities, then for all $\epsilon \in (0, 1)$, given any $\delta \in (0, 1)$, there is $m_0(\epsilon, \delta)$ such that for $m \geq m_0(\epsilon, \delta)$,

$$\forall P, P^m \left(\left\{ \mathbf{z} \in Z^m : \sup_{A \in \mathcal{E}} |\hat{P}_{\mathbf{z}}(A) - P(A)| \geq \epsilon \right\} \right) < \delta.$$

But this means precisely that, for $m \geq m_0(\epsilon, \delta)$, and for any probability measure P on $X \times \{0, 1\}$,

$$P^m \left(\left\{ \mathbf{s} \in (X \times \{0, 1\})^m : \sup_{h \in \mathcal{H}} |\text{er}_{\mathbf{s}}(h) - \text{er}_P(h)| \geq \epsilon \right\} \right) < \delta,$$

which is exactly what is required (see Definition 2.4) for \mathcal{H} to PAC-generalize.

The paper of Vapnik and Chervonenkis [23] gave the first such general uniform convergence result. In order to describe it, we first need the notion of the *growth function* of the set \mathcal{E} of events. For $S \subseteq Z$, let

$$\mathcal{E} \cap S = \{A \cap S : A \in \mathcal{E}\}.$$

Then the growth function $\Pi_{\mathcal{E}} : \mathbb{N} \rightarrow \mathbb{N}$ is given by

$$\Pi_{\mathcal{E}}(m) = \max_{|S|=m} |\mathcal{E} \cap S|.$$

Vapnik and Chervonenkis [23] proved the following result.

Theorem 3.1 (Vapnik and Chervonenkis [23]) *Let \mathcal{E} be a set of events on (Z, Λ) and P any probability measure on (Z, Λ) . Then, for $m \geq 2/\epsilon^2$,*

$$P^m \left(\left\{ \mathbf{z} \in Z^m : \sup_{A \in \mathcal{E}} |\hat{P}_{\mathbf{z}}(A) - P(A)| \geq \epsilon \right\} \right) \leq 4 \Pi_{\mathcal{E}}(2m) e^{-\epsilon^2 m/8}.$$

As it stands, this is not explicitly a uniform convergence result. However, it leads directly to one for classes \mathcal{E} whose *Vapnik-Chervonenkis dimension*—a measure of the ‘richness’ of the class, developed in [23]—is finite. Noting that $\Pi_{\mathcal{E}}(m) \leq 2^m$ for all m , one says that \mathcal{E} has finite Vapnik-Chervonenkis dimension d if $\Pi_{\mathcal{E}}(d+1) < 2^{d+1}$ and $\Pi_{\mathcal{E}}(d) = 2^d$. (Otherwise—that is, if $\Pi_{\mathcal{E}}(m) = 2^m$ for all m —the class has infinite Vapnik-Chervonenkis dimension.) The Vapnik-Chervonenkis dimension of \mathcal{E} , often called the VC-dimension, is denoted $\text{VCdim}(\mathcal{E})$. Vapnik and Chervonenkis observed that if \mathcal{E} has finite VC-dimension d then the growth function $\Pi_{\mathcal{E}}(m)$ is bounded by a polynomial of degree $d+1$. (A more precise bound, known as ‘Sauer’s

Lemma', is given in [19].) Therefore, when \mathcal{E} has finite VC-dimension, the bound given in Theorem 3.1 takes the form of a negative exponential multiplied by a polynomial, and therefore converges to 0 as m tends to infinity. Thus, finite VC-dimension is a sufficient condition for relative frequencies to converge uniformly to probabilities. In fact [23, 8], this is also a necessary condition.

Theorem 3.2 ([23, 8]) *Let \mathcal{E} be a set of subsets of a set Z . Then the relative frequencies of the events in \mathcal{E} converge uniformly to their probabilities if and only if \mathcal{E} has finite VC-dimension.*

We have seen how uniform convergence results apply directly to prove PAC-generalization, by taking the events \mathcal{E} to be the error sets E_h for $h \in \mathcal{H}$. If we identify $\{0, 1\}$ -valued functions with their supports, then we may define the VC-dimension of the set \mathcal{H} of functions. It is very easy to see that the VC-dimensions of $\mathcal{E} = \{E_h : h \in \mathcal{H}\}$ and \mathcal{H} are the same [8]; therefore, a necessary and sufficient condition for \mathcal{H} to PAC-generalize (in the sense of Definition 2.4) is that \mathcal{H} has finite VC-dimension. (On the necessity side, Blumer *et al.* [8] prove a number of stronger assertions, among them that \mathcal{H} PAC-generalizes in the weaker sense of Definition 2.2 only if \mathcal{H} has finite VC-dimension.)

We remark that a number of different uniform convergence results along the lines of Theorem 3.1 have been obtained, some of which, such as those in [21, 10, 5, 8], provide better bounds on the rate of uniform convergence.

4 Generalization of real functions

We now discuss how the previous notions of generalization have been extended to classes of real-valued functions. For simplicity, we shall assume, unless indicated otherwise, that our sets \mathcal{H} of functions map from a set X into the real interval $[0, 1]$. (It is easy to modify the theories presented here to deal with function classes mapping into other bounded subsets of the reals, as in [10], for instance.)

Most of what we shall say here can be made significantly more general: indeed, in [10], a theory is developed for function classes mapping into any complete separable metric space.

It is clear that a different approach must be taken for classes of real-valued functions. For example, if one wanted to extend Definition 2.2 to real func-

tions, it would be inappropriate, given a target real function t and a real function h , to define the error of h with respect to t (and a probability measure μ) to be $\mu(\{x \in X : h(x) \neq t(x)\})$. One should not merely be interested in whether $h(x)$ equals $t(x)$, but in *how close* $h(x)$ is to $t(x)$.

In this section we present a number of the models of generalization for real functions which have been studied. Later, we explore the connections between these and uniform convergence results in probability theory, highlighting the importance of certain extensions of the VC-dimension.

4.1 PAC-generalization

Suppose that \mathcal{H} is a set of functions from a set X to $[0, 1]$. In order to measure how close one function is to a target function or, more generally, how well it ‘fits’ a probability measure P on $Z = X \times [0, 1]$, it is useful to use a *loss function* [10]. In this approach, developed extensively by Haussler [10], a loss function is a function $l : [0, 1] \times [0, 1] \rightarrow [0, M]$, for some $M > 0$, and $l(y, y')$ may be thought of as the ‘distance’ between y and y' . Having said this, it should be emphasised that l need not be a metric. For example, a much-used loss function is the *square loss*, given by $l(y, y') = (y - y')^2$, and this is not a metric. Another commonly-used loss function is the *linear loss*, $l(y, y') = |y - y'|$, which is a metric. There are a number of other loss functions which are appropriate for a range of different problems; see the discussion in [10]. From now on, without any great loss of generality, we shall assume that $M = 1$. (Results for general M are given in [10].)

To extend the general definition, Definition 2.4, of PAC-generalization, we assume that we have, as above, a function class \mathcal{H} from X to $[0, 1]$, and a loss function $l : [0, 1] \times [0, 1] \rightarrow [0, 1]$. For the definition to be as general as possible, rather than assume that there is a target function from X to $[0, 1]$ together with a probability measure on X , we assume that we have some unknown probability measure P on $Z = X \times [0, 1]$. Given $h \in \mathcal{H}$, the error of h with respect to P is defined to be the expected value of the quantity $l(h(x), y)$, where $(x, y) \in Z$ is distributed according to P ; that is,

$$\text{er}_P(h) = \mathbb{E}_{(x,y) \sim P}(l(h(x), y)) = \mathbb{E}_P(l(h(x), y)).$$

The corresponding estimate of the error on a sample

$$\mathbf{z} = ((x_1, y_1), (x_2, y_2), \dots, (x_m, y_m)) \in Z^m$$

is

$$\text{er}_{\mathbf{z}}(h) = \frac{1}{m} \sum_{i=1}^m l(h(x_i), y_i).$$

We have the following definition of PAC-generalization in this context. Again, the motivating idea is that we want to be sure that, with high probability, on a large enough sample, the sample-based estimate of the error of any $h \in \mathcal{H}$ is close to the true error of h .

Definition 4.1 *Let \mathcal{H} be a set of functions from a set X into $[0, 1]$. We say that \mathcal{H} PAC-generalizes if for any $\epsilon, \delta \in (0, 1)$, there is $m_0(\epsilon, \delta)$ such that for any probability measure P on $Z = X \times [0, 1]$,*

$$P^m \left(\left\{ \mathbf{z} \in Z^m : \sup_{h \in \mathcal{H}} |\text{er}_{\mathbf{z}}(h) - \text{er}_P(h)| \geq \epsilon \right\} \right) < \delta$$

for $m \geq m_0(\epsilon, \delta)$.

This definition of generalization leads to certain types of ‘learning’ result (and, indeed, one could motivate it by the desire to obtain such results). Specifically, regarding a learning algorithm as a function from $\cup_{m=1}^{\infty} Z^m$ to \mathcal{H} , it is straightforward to show that the following result on (the obvious extension of) probably approximately optimal learning holds.

Theorem 4.2 ([10]) *Suppose that \mathcal{H} is a class of functions from X to $[0, 1]$ and that l is a loss function. Suppose also that \mathcal{H} PAC-generalizes (with respect to the loss function l). Then, there is a learning algorithm L such that the following holds: given $\epsilon, \delta \in (0, 1)$, there is $m_L(\epsilon, \delta)$ such that for any probability measure P on $Z = X \times [0, 1]$ and for $m \geq m_L(\epsilon, \delta)$, with probability at least $1 - \delta$, a P^m -random sample $\mathbf{z} \in Z^m$ is such that*

$$\text{er}_P(L(\mathbf{z})) < \inf_{h \in \mathcal{H}} \text{er}_P(h) + \epsilon.$$

4.2 Generalization from interpolation

Another approach to the generalization of real functions is to consider ‘generalization from approximate interpolation’, where we can develop two distinct models [3, 2]. This approach is less general than the loss functions approach, in that it extends Definition 2.2 rather than Definition 2.4. For these models of generalization, we do have a target real function $t : X \rightarrow \mathbb{R}$ together with a probability measure μ on X , and the aim is to find a good approximation to t from \mathcal{H} .

To motivate these definitions of generalization, we recall Definition 2.2, the first, most basic, definition of PAC-generalization for classes of $\{0, 1\}$ -valued

functions. The key part of the definition was the requirement that for $m \geq m_0(\epsilon, \delta)$, with μ^m -probability at least $1 - \delta$, a μ^m -randomly drawn $\mathbf{x} \in X^m$ is such that

$$\begin{aligned} & h \in \mathcal{H} \text{ and } h(x_i) = t(x_i) \text{ for } i = 1, 2, \dots, m \\ \implies & \text{er}_\mu(h) = \mu(\{x \in X : h(x) \neq t(x)\}) < \epsilon. \end{aligned}$$

As we mentioned earlier, it would be too coarse to carry this definition over, as it stands, to real function classes. Suppose that we replace the too-stringent ‘ $h(x) = t(x)$ ’ by the condition ‘ $h(x)$ is within η of $t(x)$ ’, where η is some small, chosen number in $(0, 1)$. That is, we wish to be sure that, with probability at least $1 - \delta$, if h is an η -interpolant of t on the sample, in the sense that $t(x_i) - \eta < h(x_i) < t(x_i) + \eta$ for $i = 1, 2, \dots, m$, then $h(x)$ is within η of $t(x)$ on a set of measure at least $1 - \epsilon$. Then we arrive at the following definition [6, 3].

Definition 4.3 *Let \mathcal{H} be a set of functions from X to $[0, 1]$. We say that \mathcal{H} generalizes from approximate interpolation if for all $\epsilon, \delta, \eta \in (0, 1)$, there is $m_0(\eta, \epsilon, \delta)$ such that, for all probability measures μ on X and all $t : X \rightarrow \mathbb{R}$, if $m \geq m_0(\eta, \epsilon, \delta)$ then with μ^m -probability at least $1 - \delta$, a sample $\mathbf{x} = (x_1, x_2, \dots, x_m) \in X^m$ is such that the following implication holds for every $h \in \mathcal{H}$:*

$$|h(x_i) - t(x_i)| < \eta (1 \leq i \leq m) \implies \mu(\{x : |h(x) - t(x)| \geq \eta\}) < \epsilon.$$

This definition may seem to be rather demanding: one expects, with high probability, to be able to deduce from the fact that h is within η of t on a random sample that it is within η of t almost everywhere else (with respect to μ). We may weaken this requirement by requiring such an h to be close to t almost everywhere, but not necessarily *as close as* η . This results in the following definition [2].

Definition 4.4 *Let \mathcal{H} be a set of functions from X to $[0, 1]$. We say that \mathcal{H} weakly generalizes from approximate interpolation if for all $\epsilon, \delta, \eta, \gamma \in (0, 1)$, there is $m_0(\eta, \gamma, \epsilon, \delta)$ such that, for all probability measures μ on X and all $t : X \rightarrow \mathbb{R}$, if $m \geq m_0(\eta, \gamma, \epsilon, \delta)$ then with μ^m -probability at least $1 - \delta$, a sample $\mathbf{x} = (x_1, x_2, \dots, x_m) \in X^m$ is such that the following implication holds for every $h \in \mathcal{H}$:*

$$|h(x_i) - t(x_i)| < \eta (1 \leq i \leq m) \implies \mu(\{x : |h(x) - t(x)| \geq \eta + \gamma\}) < \epsilon.$$

5 Uniform convergence of empiricals

5.1 Definitions and connections with generalization

We have seen how PAC-generalization for $\{0, 1\}$ -valued function classes relates directly to the uniform convergence of relative frequencies to probabilities. Here, we explain how PAC-generalization for classes of real functions relates to results in probability theory on the uniform convergence of empiricals to expectations.

Let Z be a set and Λ a σ -algebra on Z . Suppose that \mathcal{F} is a set of random variables on Z , each with range $[0, 1]$. For $\mathbf{z} \in Z^m$, the *empirical estimate* of the random variable $f \in \mathcal{F}$ on \mathbf{z} is

$$\hat{\mathbb{E}}_{\mathbf{z}}(f) = \frac{1}{m} \sum_{i=1}^m f(z_i).$$

One says that the *empiricals (of \mathcal{F}) converge uniformly to expectations* if there is a function $f(m, \epsilon)$ such that for every $\epsilon \in (0, 1)$, $f(m, \epsilon) \rightarrow 0$ as $m \rightarrow \infty$, and

$$\forall P, P^m \left(\left\{ \mathbf{z} \in Z^m : \sup_{f \in \mathcal{F}} |\hat{\mathbb{E}}_{\mathbf{z}}(f) - \mathbb{E}(f)| \geq \epsilon \right\} \right) < f(m, \epsilon),$$

where ‘ $\forall P$ ’ means for all probability measures defined on (Z, Λ) . Note that this is an extension of the notion of the uniform convergence of relative frequencies to probabilities, since the probability and relative frequency of an event are, respectively, the expectation and the empirical estimate of its indicator (or characteristic) function.

Returning to the loss functions approach to function generalization, given any $h \in \mathcal{H}$, let l_h be the function from Z to $[0, 1]$ defined by $l_h(x, y) = l(h(x), y)$. Then, as is easily verified,

$$\text{er}_P(h) = \mathbb{E}_P(l_h), \quad \text{er}_{\mathbf{z}}(h) = \hat{\mathbb{E}}_{\mathbf{z}}(l_h).$$

From this, it is clear that PAC-generalization as defined in Definition 4.1 is equivalent to the uniform convergence of empiricals to expectations for the class of random variables $\mathcal{F} = \{l_h : h \in \mathcal{H}\}$, usually called the *loss space* and denoted $l_{\mathcal{H}}$ [10].

5.2 Covering numbers

The notion of covering numbers turns out to be central to results on uniform convergence of empiricals to expectations. Given a pseudo-metric space (Y, d)

and a subset S of Y , we say that the set $T \subseteq Y$ is an ϵ -cover for S (where $\epsilon > 0$) if, for every $s \in S$ there is $t \in T$ such that $d(s, t) \leq \epsilon$. For a fixed $\epsilon > 0$ we denote by $\mathcal{N}(\epsilon, S, d)$ the cardinality of the smallest ϵ -cover for S . (We define $\mathcal{N}(\epsilon, S, d)$ to be ∞ if there is no such cover.)

Suppose that \mathcal{F} is a set of $[0, 1]$ -random variables defined on Z . For $\mathbf{z} = (z_1, z_2, \dots, z_m)$, we denote by $\mathcal{F}(\mathbf{z})$ the following subset of \mathbb{R}^m :

$$\mathcal{F}(\mathbf{z}) = \{(f(z_1), f(z_2), \dots, f(z_m)) : f \in \mathcal{F}\}.$$

Pollard [18] (see also [10]) obtained the following result.

Theorem 5.1 *Suppose that \mathcal{F} is a permissible¹ set of $[0, 1]$ -valued random variables on Z and that P is any probability measure on Z . Then, for any positive integer m and any $\epsilon > 0$,*

$$\begin{aligned} & P^m \left(\left\{ \mathbf{z} \in Z^m : \sup_{f \in \mathcal{F}} |\hat{\mathbb{E}}_{\mathbf{z}}(f) - \mathbb{E}_P(f)| \geq \epsilon \right\} \right) \\ & \leq 4\mathbb{E}_{P^{2m}} \left(\mathcal{N} \left(\frac{\epsilon}{16}, \mathcal{F}(\mathbf{z}), d \right) \right) \exp \left(-\frac{\epsilon^2 m}{128} \right), \end{aligned}$$

where d is the L^1 -metric, given, on \mathbb{R}^{2m} , by $d(\mathbf{r}, \mathbf{s}) = \frac{1}{2m} \sum_{i=1}^{2m} |r_i - s_i|$.

Haussler [10] has improved this result in a number of ways, but the form given here is sufficient for our purposes.

From this result, it can be seen that uniform convergence of empiricals to expectations will occur if the covering numbers can be bounded in such a way that the bound of Theorem 5.1 tends to 0 as $m \rightarrow \infty$. Such bounds have been obtained in terms of ‘dimensions’ which characterise the ‘richness’ of the class \mathcal{F} in much the same way as the VC-dimension does for $\{0, 1\}$ -valued classes.

5.3 The pseudo-dimension

The *pseudo-dimension* (also known as the *Pollard dimension*) was introduced by Pollard [18]. With the notation as above, we say that $\mathbf{z} \in Z^m$ is *P-shattered* by \mathcal{F} if some translate $\mathbf{r} + \mathcal{F}(\mathbf{z})$ of $\mathcal{F}(\mathbf{z})$ intersects all orthants of

¹The set of random variables must satisfy some measurability conditions; see [10, 18] for details. These are not particularly stringent, and we refer to a class with the required properties as a permissible class.

\mathbb{R}^m . The *pseudo-dimension* of H , denoted $\text{Pdim}(H)$, is the largest length of a P -shattered \mathbf{z} (or it is infinite, if there is no bound on the lengths of P -shattered \mathbf{z}). We state the definition formally in a rather more explicit way.

Definition 5.2 (Pseudo-dimension) *With the usual notation, $\mathbf{z} \in Z^m$ is pseudo-shattered by \mathcal{F} if there are $r_1, r_2, \dots, r_m \in \mathbb{R}$ such that for any $\mathbf{b} \in \{0, 1\}^m$, there is $f_{\mathbf{b}} \in \mathcal{F}$ with*

$$f_{\mathbf{b}}(z_i) \geq r_i \iff b_i = 1.$$

The largest d such that some $\mathbf{z} \in Z^d$ is P -shattered is the pseudo-dimension of \mathcal{F} , denoted $\text{Pdim}(\mathcal{F})$. (When this maximum does not exist, the pseudo-dimension is taken to be infinite.)

Note that when the class \mathcal{F} in fact maps into the set $\{0, 1\}$ rather than the interval $[0, 1]$, the definition of pseudo-dimension reduces to the VC-dimension. Furthermore, when \mathcal{F} is a vector space of real functions, the pseudo-dimension of \mathcal{F} is precisely the linear dimension of \mathcal{F} ; see [10].

The following result follows from one due to Pollard [18]; see [10].

Theorem 5.3 *Suppose that d is the L^1 -metric on \mathbb{R}^k , where k is any positive integer, and that $\mathbf{z} \in Z^k$. Suppose also that \mathcal{F} has finite pseudo-dimension. Then, for any $\epsilon \in (0, 1)$,*

$$\mathcal{N}(\epsilon, \mathcal{F}(\mathbf{z}), d) < 2 \left(\frac{2e}{\epsilon} \ln \left(\frac{2e}{\epsilon} \right) \right)^{\text{Pdim}(\mathcal{F})}.$$

Thus, when \mathcal{F} has finite pseudo-dimension, d , we see from Theorem 5.1 that, for any P ,

$$P^m \left(\left\{ \mathbf{z} \in Z^m : \sup_{f \in \mathcal{F}} |\hat{\mathbb{E}}_{\mathbf{z}}(f) - \mathbb{E}(f)| \geq \epsilon \right\} \right) < f(m, \epsilon),$$

where

$$f(m, \epsilon) = 8 \left(\frac{32e}{\epsilon} \ln \left(\frac{32e}{\epsilon} \right) \right)^d \exp \left(-\frac{\epsilon^2 m}{128} \right) \rightarrow 0 \text{ as } m \rightarrow \infty,$$

for each fixed ϵ . One therefore has the following result.

Theorem 5.4 *If \mathcal{F} has finite pseudo-dimension then the empiricals of the random variables in \mathcal{F} converge uniformly to their expectations.*

Returning to function generalization, and recalling that what we require for PAC-generalization is uniform convergence of empiricals to expectations for the loss space $l_{\mathcal{H}}$, we obtain the following result of Haussler [10].

Corollary 5.5 *If $l_{\mathcal{H}}$ has finite pseudo-dimension then \mathcal{H} PAC-generalizes.*

For certain loss functions l , the pseudo-dimension of l can be related directly to the pseudo-dimension of \mathcal{H} ; see [10].

5.4 Scale-sensitive pseudo-dimension

Recently, it has been shown that finiteness of the pseudo-dimension of \mathcal{F} is a stronger condition than is needed for uniform convergence of empiricals to expectations. Alon *et al.* [1] have determined a weaker sufficient condition for the uniform convergence. Their characterisation involves a ‘scale-sensitive’ counterpart of the pseudo-dimension, which was introduced by Kearns and Schapire [12] in their work on the learnability of probabilistic concepts, and which, after [7], we shall call the *fat-shattering function*.

Definition 5.6 (fat-shattering) *Suppose that \mathcal{F} is a set of functions from Z to $[0, 1]$ and that $\gamma > 0$. We say that $\mathbf{z} \in Z^m$ is γ -shattered if there is $\mathbf{r} = (r_1, r_2, \dots, r_m) \in \mathbb{R}^m$ such that for every $\mathbf{b} = (b_1, b_2, \dots, b_m) \in \{0, 1\}^m$, there is a function $f_{\mathbf{b}} \in \mathcal{F}$ with $f_{\mathbf{b}}(z_i) \geq r_i + \gamma$ if $b_i = 1$ and $f_{\mathbf{b}}(z_i) \leq r_i - \gamma$ if $b_i = 0$.*

Thus, \mathbf{z} is γ -shattered if it is shattered with a ‘width of shattering’ of at least γ . We define the *fat-shattering function*, $\text{fat}_{\mathcal{F}} : \mathbb{R}^+ \rightarrow \mathbb{N} \cup \{0, \infty\}$, by defining $\text{fat}_{\mathcal{F}}(\gamma)$ to be the largest d such that some $\mathbf{z} \in Z^d$ is γ -shattered. (We define $\text{fat}_{\mathcal{F}}(\gamma) = \infty$ if there is no maximum such d .) It is easy to see that $\text{Pdim}(\mathcal{F}) = \lim_{\gamma \rightarrow 0} \text{fat}_{\mathcal{F}}(\gamma)$. It should be noted, however, that it is possible for the pseudo-dimension to be infinite, even when $\text{fat}_{\mathcal{F}}(\gamma)$ is finite for all positive γ . We shall say that \mathcal{F} has *finite fat-shattering function* whenever it is the case that for all $\gamma \in (0, 1)$, $\text{fat}_{\mathcal{F}}(\gamma)$ is finite.

Alon *et al.* [1] bounded the covering numbers in terms of the fat-shattering function.

Theorem 5.7 (Alon et al. [1]) *Suppose that \mathcal{F} is a set of $[0, 1]$ -valued random variables on Z and that \mathcal{F} has finite fat-shattering function. Let m be*

a positive integer. Suppose $\gamma > 0$ and that $d = \text{fat}_{\mathcal{F}}(\gamma/4)$. Let

$$B = \sum_{i=1}^d \binom{m}{i} \left(\left\lceil \frac{2}{\gamma} \right\rceil \right)^i.$$

Then, provided $m \geq \log B + 1$, for any $\mathbf{z} \in Z^m$,

$$\mathcal{N}(\epsilon, \mathcal{F}(\mathbf{z}), d^\infty) < 2 \left(m \left\lceil \frac{2}{\gamma} \right\rceil^2 \right)^{\log B},$$

where d is the L^∞ -metric on \mathbb{R}^m , given by $d^\infty(\mathbf{r}, \mathbf{s}) = \max_{1 \leq i \leq m} |r_i - s_i|$.

Now, for any $\mathbf{r}, \mathbf{s} \in \mathbb{R}^m$, the L^1 -distance is bounded by the L^∞ -distance:

$$d(\mathbf{r}, \mathbf{s}) = \frac{1}{m} \sum_{i=1}^m |r_i - s_i| \leq \max_{1 \leq i \leq m} |r_i - s_i| = d^\infty(\mathbf{r}, \mathbf{s}).$$

This means that any ϵ -cover with respect to d^∞ is also an ϵ -cover with respect to d , and hence, for any \mathcal{F} , for all ϵ , and all \mathbf{z} , $\mathcal{N}(\epsilon, \mathcal{F}(\mathbf{z}), d) \leq \mathcal{N}(\epsilon, \mathcal{F}(\mathbf{z}), d^\infty)$. Thus, the bound of Theorem 5.7 is a bound on the required covering numbers. This bound is sub-exponential in m and so, as earlier, it leads to a uniform convergence result: if \mathcal{F} has finite fat-shattering function then the empiricals converge to their expectations uniformly. In fact, as shown in [1], the converse is also true.

Theorem 5.8 (Alon et al. [1]) *One has uniform convergence of empiricals to expectations for the class \mathcal{F} of $[0, 1]$ -random variables if and only if \mathcal{F} has finite fat-shattering function.*

We have the following immediate corollary for generalization.

Theorem 5.9 (Alon et al. [1]) *The class of functions \mathcal{H} from X to $[0, 1]$ PAC-generalizes (with respect to loss function l) if and only if the loss space $l_{\mathcal{H}}$ has finite fat-shattering function.*

6 Generalization from interpolation

6.1 The strong model

The problem of generalization from approximate interpolation may be regarded to some extent as a problem within the loss functions approach to

function learning. To see this, let us fix $\eta \in (0, 1)$ and take l^η to be the loss function given by

$$l^\eta(y, y') = 0 \text{ if } |y - y'| < \eta, \quad l^\eta(y, y') = 1 \text{ if } |y - y'| \geq \eta.$$

Then, for fixed η , generalization from interpolation is equivalent to a restricted form of PAC-generalization with respect to l^η , in which we only consider distributions P on $Z = X \times \mathbb{R}$ which correspond (in the obvious fashion) to pairs (t, μ) where t is a function on X and μ is a probability measure on X . The loss space $l^\eta_{\mathcal{H}}$ is $\{0, 1\}$ -valued, so its pseudo-dimension and fat-shattering function are precisely its VC-dimension. In [3], a scale-sensitive dimension $\text{Bdim}_{\mathcal{H}}$ is defined as follows: for $\gamma \in (0, 1)$,

$$\text{Bdim}_{\mathcal{H}}(\gamma) = \text{VCdim}(l^\gamma_{\mathcal{H}}).$$

We call this the *band dimension* and say that \mathcal{H} has *finite band dimension* if $\text{Bdim}_{\mathcal{H}}(\gamma)$ is finite for all γ . In [3], the following result is obtained.

Theorem 6.1 *The class \mathcal{H} generalizes from approximate interpolation if and only if it has finite band dimension.*

The band dimension is not a well-known measure of dimension, having appeared only rarely in other work on learning theory (such as [17]). However, it can be related to the pseudo-dimension [3], as follows.

Theorem 6.2 *Suppose that \mathcal{H} maps from a domain X into a bounded real interval. Then \mathcal{H} has finite band dimension if and only if it has finite pseudo-dimension. Furthermore, there are constants $c_1, c_2 > 0$ such that for all $\gamma \in (0, 1)$,*

$$c_1 \frac{\text{Pdim}(\mathcal{H})}{\log(1/\eta)} \leq \text{Bdim}_{\mathcal{H}}(\gamma) \leq c_2 \text{Pdim}(\mathcal{H}).$$

Corollary 6.3 *\mathcal{H} generalizes from approximate interpolation if and only if \mathcal{H} has finite pseudo-dimension*

Thus, although it looks like a very difficult definition to satisfy, generalization from interpolation holds for many natural classes of functions. The fact that finite pseudo-dimension of \mathcal{H} is necessary for generalization from approximate interpolation is, in some ways, in contrast to the results for PAC-generalization; there, the finiteness condition on the pseudo-dimension (namely, $\text{Pdim}(l_{\mathcal{H}}) < \infty$) is not necessary and can be replaced by finiteness of the fat-shattering function.

The following result from [3] provides indication of the appropriate size of $m_0(\eta, \epsilon, \delta)$.

Theorem 6.4 *Suppose that \mathcal{H} has finite pseudo-dimension d . Then there is a constant c such that a sufficient value of $m_0(\eta, \epsilon, \delta)$ for generalization from approximate interpolation is*

$$\frac{c}{\epsilon} \left(\text{Pdim}(\mathcal{H}) \ln \left(\frac{1}{\epsilon} \right) + \ln \left(\frac{1}{\delta} \right) \right).$$

The paper [3] contains many other results on generalizing from approximate interpolation, including a characterisation of those measures of dimension whose finiteness is a necessary and sufficient condition for such generalization.

6.2 The weak model

Since finite pseudo-dimension is a sufficient condition for generalization from approximate interpolation, it is also a sufficient condition for weak generalization from approximate interpolation. However, the following result has been obtained.

Theorem 6.5 *\mathcal{H} weakly generalizes from approximate interpolation if and only if \mathcal{H} has finite fat-shattering function.*

Unlike generalization from approximate interpolation, the problem of weak generalization cannot be expressed directly as a problem involving loss functions. The ‘if’ part of Theorem 6.5 follows from the following ‘convergence’ result, which is implicit in [2].

Theorem 6.6 *Suppose that \mathcal{H} is a class of functions mapping from a domain X to the real interval $[0, 1]$ and that \mathcal{H} has finite fat-shattering function. Let t be any function from X to \mathbb{R} and let $\gamma, \eta, \epsilon \in (0, 1)$. Let μ be any probability distribution on X and m any positive integer. Define the subset Q of X^m to be the set of \mathbf{x} for which there exists $h \in \mathcal{H}$ such that*

$$\mu(\{x \in X : |h(x) - t(x)| \geq \eta + \gamma\}) > \epsilon \text{ and } |h(x_i) - t(x_i)| < \eta, (1 \leq i \leq m).$$

Then

$$\mu^m(Q) < 2\mathbb{E}_{\mu^{2m}} \left(\mathcal{N} \left(\frac{\gamma}{2}, \mathcal{H}(\mathbf{z}), d^\infty \right) \right) 2^{-\epsilon m/2},$$

where d^∞ is the L^∞ -metric on \mathbb{R}^{2m} , given by $d^\infty(\mathbf{r}, \mathbf{s}) = \max_{1 \leq i \leq 2m} |r_i - s_i|$.

Combining this with the result of Alon *et al.*, Theorem 5.7 gives the ‘if’ part of Theorem 6.5. The following result from [2] indicates a suitable value of $m_0(\eta, \gamma, \epsilon, \delta)$.

Theorem 6.7 *Suppose that \mathcal{H} maps from a domain X into $[0, 1]$ and that \mathcal{H} has finite fat-shattering function. There is a constant K such that a sufficient sample length for weak generalization from approximate interpolation is*

$$m_0(\gamma, \eta, \epsilon, \delta) = \frac{K}{\epsilon} \left(\ln \left(\frac{1}{\delta} \right) + \text{fat}_{\mathcal{H}}(\gamma/8) \ln^2 \left(\frac{\text{fat}_{\mathcal{H}}(\gamma/8)}{\gamma\epsilon} \right) \right).$$

The proof that finite fat-shattering function is necessary can be found in [2].

6.3 Some applications to artificial neural networks

In this section, we explain how the results on generalization from interpolation can be used to obtain results for certain types of neural network. (See, for example, [11], for the basic definitions of neural networks.) We consider here artificial neural networks \mathcal{N} having one hidden layer, where each hidden node has linear threshold activation function, and in which the activation function of the single output node is the identity function (so that it outputs the weighted sum of its inputs). We do not assume that the number of hidden units is known. We do, however, make the assumption that the weights from the hidden layer to the output node are bounded in such a way that the sum of the absolute values of all these weights is bounded by a fixed, known, constant B . (If bounds are given on the number k of hidden nodes and on the absolute value of each weight from the hidden layer to the output node, then we certainly have such a bound B : thus the restriction we impose on the weights is weaker than imposing a restriction on the number of hidden units and on the magnitude of the weights.) Each *state* of such a neural network \mathbf{N} is described by weight vectors $\mathbf{b} \in \mathbb{R}^k$ and $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k \in \mathbb{R}^n$. The assumption on the weights is that $\sum_{i=1}^k |b_i| \leq B$. Let $\text{sign}(x) : \mathbb{R} \rightarrow \{0, 1\}$ be given by $\text{sign}(x) = 1 \iff x \geq 0$. Then we may describe the set of functions computable by the neural network explicitly, as follows. Let $\omega = (\mathbf{b}, \mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k)$ denote a typical state of the network. Then the function computed by the network in state ω is $h_\omega : \mathbb{R}^n \rightarrow \mathbb{R}$ given by

$$h_\omega(x_1, x_2, \dots, x_n) = \sum_{j=1}^k b_j \text{sign} \left(\sum_{i=1}^n w_{ji} x_i \right).$$

The set \mathcal{H} of all functions computable by such a neural network architecture is then

$$\mathcal{H} = \left\{ h_\omega : \omega = (\mathbf{b}, \mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k), \text{ where } k \in \mathbb{N} \text{ and } \sum_{j=1}^k |b_j| \leq B \right\}.$$

Results from [9] show that

$$\text{fat}_{\mathcal{H}}(\alpha) \leq K \frac{B^2 n^2}{\alpha^2} \ln \left(\frac{Bn}{\alpha} \right),$$

where K is some fixed constant.

The following result is obtained by using this bound together with Theorem 6.5 and Theorem 6.7 (or, rather, their simple modifications for classes of functions mapping into the interval $[0, B]$). For the sake of simplicity, we have not explicitly determined the constants involved.

Theorem 6.8 *There are constants c and ϵ_0 and γ_0 such that the following holds. Let ϵ , δ and η be fixed positive numbers less than 1, with $\epsilon < \epsilon_0$ and $\gamma < \gamma_0$. Suppose that \mathbf{N} is any one-hidden-layer network of the type described above and let μ be any probability distribution on X , the set of all inputs. Suppose that the target function t is computable by the network. Suppose also that L is a learning algorithm for \mathbf{N} with the property that if $\mathbf{s} = ((x_1, y_1), (x_2, y_2), \dots, (x_m, y_m))$ is any training sample for t of length m , then $h_\omega = L(\mathbf{s})$ satisfies $|h_\omega(x_i) - y_i| < \eta$ for $1 \leq i \leq m$. Then, if a training sample \mathbf{s} is generated by a μ^m -random choice of (x_1, x_2, \dots, x_m) , where $m > (16/\epsilon) \ln(4/\delta)$, then, with probability at least $1 - \delta$, $L(\mathbf{s}) = h_\omega$ is such that*

$$\mu \left(\left\{ x : |h_\omega(x) - t(x)| \geq \eta + c \frac{Bn (\ln m)^2}{\sqrt{\epsilon} \sqrt{m}} \right\} \right) < \epsilon.$$

Proof It follows from Theorem 6.6 together with Theorem 5.7 (see [2]) that a sufficient sample length $m_0(\eta, \gamma, \epsilon, \delta)$ for (the class of functions computable by) \mathbf{N} to approximate from interpolated examples is

$$m_0(\eta, \gamma, \epsilon, \delta) = \frac{8}{\epsilon} \left(3d \ln^2 \left(\frac{1200dB^2}{\epsilon\gamma^2 \ln^2 2} \right) + \ln \left(\frac{4}{\delta} \right) \right)$$

where $d = \text{fat}_{\mathcal{H}}(\gamma/8)$. Let $m/2 > (8/\epsilon) \ln(4/\delta)$. Then m will be at least m_0 if

$$\frac{m}{2} > \frac{24d}{\epsilon \ln^2 2} \ln^2 \left(\frac{1200dB^2}{\epsilon\gamma^2 \ln^2 2} \right).$$

Recalling that $\text{fat}_{\mathcal{H}}(\alpha)$ is bounded by $K(B^2 n^2 / \alpha^2) \ln(Bn/\alpha)$, there is a constant c_1 such that this will be true if $m > c_1 \beta^2 (\ln \beta)^3$, where $\beta = Bn / (\gamma \sqrt{\epsilon})$. There is a constant c_2 such that this inequality holds if $\beta < c_2 (\sqrt{m} / (\ln m)^2)$; for then, $\beta^2 (\ln \beta)^3$ is of order no more than

$$\frac{m}{(\ln m)^4} (\ln m)^3 = \frac{m}{\ln m}.$$

(Clearly, this same argument works provided we take $\beta < c_2(\sqrt{m}/(\ln m)^{3/2+x})$, where x is any fixed positive number, but the present choice ($x = 1/2$) suffices for our purpose.) Now,

$$\frac{Bn}{\gamma\sqrt{\epsilon}} = \beta < c_2 \frac{\sqrt{m}}{(\ln m)^2}$$

means we may take $\gamma = c(Bn(\ln m)^2/(\sqrt{\epsilon}\sqrt{m}))$ for some constant c , as required. \square

Thus, if we fix in advance the accuracy and confidence, ϵ and δ , and if we have a learning procedure which will η -interpolate on any training sample of length m for t , we can be confident (with probability at least $1 - \delta$) that the final state of the network will estimate $t(x)$ to within an error margin of $cBn(\ln m)^2/(\sqrt{\epsilon}\sqrt{m})$ on most inputs. (Formally, it will compute within this error margin on all inputs but for those in some set having probability less than ϵ .)

Consider now the subclass of these neural networks in which the number of hidden units is fixed at some number k . We can bound the pseudo-dimension of the class of functions computable by the network \mathbf{N} in terms of k [9]: specifically, the pseudo-dimension of order $kn \ln(kn)$. (In fact, the same bound also holds without the restriction on the weights into the output node.) From Theorem 6.4, we have the following result.

Theorem 6.9 *Let ϵ , δ and η be fixed positive numbers. Let \mathbf{N} be a network of the type just described, having n input nodes and k hidden nodes. Suppose that μ is a probability distribution on X , the set of all inputs and that the target function t is computable by the network. Suppose also that the learning algorithm L is such that if $\mathbf{s} = ((x_1, y_1), (x_2, y_2), \dots, (x_m, y_m))$ is any training sample for t of length m , then $h_\omega = L(\mathbf{s})$ satisfies $|h_\omega(x_i) - y_i| < \eta$ for $1 \leq i \leq m$. Then there is a constant c , depending only on ϵ and δ , such that the following holds for $m > ckn \ln(kn)$: if a training sample \mathbf{s} is generated by a μ^m -random choice of (x_1, x_2, \dots, x_m) , then, with probability at least $1 - \delta$, $L(\mathbf{s}) = h_\omega$ satisfies*

$$P(\{x \in X : |h_\omega(x) - t(x)| \geq \eta\}) < \epsilon.$$

Acknowledgements

The author's work is supported in part by the European Union through the ESPRIT project 'Neurocolt'.

References

- [1] Alon, N., S. Ben-David, N. Cesa-Bianchi, and D. Haussler (1993). Scale-sensitive dimensions, uniform convergence, and learnability. In *Proceedings of the Symposium on Foundations of Computer Science*. IEEE Press, 1993.
- [2] Anthony, M. and P. Bartlett (1995). Function learning from interpolation. submitted. (An extended abstract appears in *Proceedings Eurocolt'95*, Springer-Verlag 1995.)
- [3] Anthony, M., P. Bartlett, Y. Ishai, and J. Shawe-Taylor (1994). Valid generalisation from approximate interpolation. To appear, *Combinatorics, Probability and Computing*.
- [4] Anthony, M. and N. Biggs (1992). *Computational Learning Theory: An Introduction*. Cambridge Tracts in Theoretical Computer Science (30). Cambridge University Press, Cambridge, UK, 1992.
- [5] Anthony, M. and J. Shawe-Taylor (1994). A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207–217, 1994.
- [6] Anthony, M. and J. Shawe-Taylor (1994). Valid generalisation of functions from close approximations on a sample. In *Proceedings of EuroCOLT'93*. Oxford University Press, 1994.
- [7] Bartlett, P.L. , P.M. Long, and R.C. Williamson (1994). Fat-shattering and the learnability of real-valued functions. In *Proceedings of the Seventh Annual ACM Conference on Computational Learning Theory*, ACM Press, New York.
- [8] Blumer, A, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth (1989). Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4):929–965, 1989.
- [9] Gurvits, L. and P. Koiran (1995). Approximation and learning of convex superpositions. In *Proceedings of Eurocolt'95*, Springer-Verlag Lecture Notes in Artificial Intelligence, pages 222–236.
- [10] Haussler, D. (1992). Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100(1):78–150, Sept. 1992.
- [11] Hertz, J., A. Krogh, and R. Palmer (1991). *Introduction to the Theory of Neural Computation*. Addison-Wesley, Redwood City, 1991.
- [12] Kearns, M.J. and R.E. Schapire (1990). Efficient distribution-free learning of probabilistic concepts, in *Proceedings of the 1990 IEEE Symposium on Foundations of Computer Science*, IEEE Press.

- [13] Kearns, M.J., R. E. Schapire, and L. M. Sellie (1992). Toward efficient agnostic learning. In *Proceedings of the 5th Annual Workshop on Computational Learning Theory*, pages 341–352. ACM Press, New York, NY, 1992.
- [14] Kearns, M.J. and U. Vazirani (1995). *Introduction to Computational Learning Theory*, MIT Press 1995.
- [15] Maass, W. (1993). Agnostic PAC-learning of functions on analog neural nets (extended abstract). In *Advances in Neural Information Processing Systems, 6*. Morgan Kaufmann, 1993.
- [16] Natarajan, B.K. (1991). *Machine Learning: A Theoretical Approach*. Morgan Kaufmann, San Mateo, California, 1991.
- [17] Natarajan, B.K. (1993). Occam’s razor for functions. In *Proceedings of the Sixth ACM Workshop on Computational Learning Theory, July 1993*, ACM Press.
- [18] Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, 1984.
- [19] Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory (A)*, 13:145–147, 1972.
- [20] Valiant, L.G. (1984). A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, Nov. 1984.
- [21] Vapnik, V.N. (1982). *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.
- [22] Vapnik, V.N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New York, 1995.
- [23] Vapnik, V.N. and A. Y. Chervonenkis (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.