Uniform Glivenko-Cantelli Theorems and Concentration of Measure in the Mathematical Modelling of Learning

Martin Anthony
Department of Mathematics
London School of Economics
Houghton Street
London WC2A 2AE, UK
m.anthony@lse.ac.uk,
www.maths.lse.ac.uk/Personal/martin

CDAM Research Report LSE-CDAM-2002-07
May 2002

Abstract

This paper surveys certain developments in the use of probabilistic techniques for the modelling of generalization in machine learning. Building on 'uniform convergence' results in probability theory, a number of approaches to the problem of quantifying generalization have been developed in recent years. Initially these models addressed binary classification, and as such were applicable, for example, to binary-output neural networks. More recently, analysis has been extended to apply to regression problems, and to classification problems in which the classification is achieved by using real-valued functions (in which the concept of a large margin has proven useful). In order to obtain more useful and realistic bounds, and to analyse model selection, another development has been the derivation of data-dependent bounds. Here, we discuss some of the main probabilistic techniques and key results, particularly the use (and derivation of) uniform Glivenko-Cantelli theorems, and the use of concentration of measure results. Many details are omitted, the aim being to give a high-level overview of the types of approaches taken and methods used.

1 Probabilistic Modelling of Learning

We begin by describing a by-now very standard probabilistic model of supervised learning. Suppose that X is a set of *examples*, which in a neural network context would be elements of $\mathbb{R}^n$ if the network has n real inputs. Suppose also that $Y \subseteq \mathbb{R}$ is the set of possible *outputs*. We shall always assume that $Y \subseteq [0,1]$, and in some cases we shall be interested in the situation where $Y = \{0,1\}$. So Y represents, for instance, the output value of a neural net with one output unit. The set $X \times Y$ will be denoted Z, and elements (x,y) of Z will be called *labelled examples*. In the model, we shall assume that a *learning algorithm* $\mathcal{A}$ takes a randomly generated *training sample* of labelled examples, (each called a *training example*) and produces a function $h : X \to [0,1]$, chosen from some class H of functions. The goal is to produce a *hypothesis* h that is a 'good fit' to the process generating the labelled examples. More precisely, we assume that there is some fixed, but unknown, probability measure μ on Z. (There is a fixed σ-algebra Σ on Z, which when $Z \subseteq \mathbb{R}^n$, we shall take to be the Borel σ-algebra. Then, μ denotes a probability measure on (Z, Σ). A number of measurability conditions are implicitly assumed in what follows, but these conditions are reasonable and not particularly stringent. Details may be found in [16, 36] for instance.) We assume that each training example is generated independently according to μ. (So, if the training sample is of length n, then it is generated according to the product probability measure μ^n.) Formally, a learning algorithm is a function $A : \bigcup_{n=1}^{\infty} Z^n \to H$, and we say that the learning algorithm is *successful* if, with high μ^m-probability, it produces an *output hypothesis* $\mathcal{A}(\mathbf{z})$ which is almost as good a fit to the distribution μ as exists in the class H. More precisely, we have in mind some *loss function* $\ell : [0,1] \times Y \to [0,1]$, and what we hope for is that $\mathcal{A}(\mathbf{z})$ has a relatively small *loss*, where, for $h \in H$, the loss of h is the expectation $L(h) = \mathbb{E}\,\ell(h(x), y)$ (where the expectation is with respect to μ). Examples of loss functions are $\ell(r,s) = |r-s|$, $\ell(r,s) = (r-s)^2$, and the discrete loss, given by $\ell(r,s) = 0$ if $r = s$ and $\ell(r,s) = 1$ if $r \neq s$. Since the best loss one could hope to be near is $L^* = \inf_{h \in H} L(h)$, we want $\mathcal{A}(\mathbf{z})$ to have loss close to L^*, with high probability, provided the sample size n is large enough. Since we do not know μ, we must be able to guarantee this success without reference to the particular distribution μ generating the examples. We therefore have the following definition. (Here, and in the rest of the paper, we use the symbol $\mathbb{P}$ to denote probability. This is slightly imprecise, in that the measure is not specified, but this is usually clear in any case. For instance, in the definition that follows, the probability is with respect to μ^n.)

Definition 1.1 *With the above notations, $\mathcal{A}$ is a* successful *learning algorithm for H if for all $\epsilon, \delta \in (0,1)$, there is some $n_0(\epsilon, \delta)$ (depending on ϵ and δ only) such that, if $n > n_0(\epsilon, \delta)$, then with probability at least $1-\delta$, $L(\mathcal{A}(\mathbf{z})) \leq L^* + \epsilon$. The minimal such $n_0(\epsilon, \delta)$ is referred to as the* sample complexity *of $\mathcal{A}$ and is denoted $n_{\mathcal{A}}(\epsilon, \delta)$.*

Note that if $\mathcal{A}$ is successful, then there is some function $\epsilon_0(n, \delta)$ of n and δ, with the property that for all δ, $\lim_{n \to \infty} \epsilon_0(n, \delta) = 0$, and such that for any probability measure

μ on Z, with probability at least $1-\delta$ we have $L(\mathcal{A}(\mathbf{z})) \leq L^* + \epsilon_0(n,\delta)$. The minimal $\epsilon_0(n,\delta)$ is called the *estimation error* of the algorithm.

When H is a set of binary functions, meaning each function in H maps into $\{0,1\}$, if $Y = \{0,1\}$, and if we use the discrete loss function, then we shall say that we have a *binary* (classification) learning problem. If $L^* = 0$ then we say that we have a *realisable* the learning problem. In this situation there is some $t \in H$ such that with probability 1, for all $(x,y) \in Z$, $y = t(x)$. (In particular, this includes the case in which there is a fixed probability distribution μ on X and a *target function* $t : X \to \{0,1\}$.)

For a binary, realisable learning problem, the definition of successful learning is quite simple to understand: for any $h \in H$, $L(h)$ is the probability that on a randomly drawn element (x,y) of Z, h and t agree on x; that is, $h(x) = t(x)$. So what the definition says is that, provided the sample is large enough (of length greater than $n_0(\epsilon,\delta)$, then, with probability at least $1-\delta$, $\mathcal{A}$ produces a hypothesis which agrees with the target function with probability at least $1-\epsilon$ on a further randomly drawn example.

We might want to use real functions for classification. Here, we would have $Y = \{0,1\}$, but $H : X \to [0,1]$. In this case, one appropriate loss function would be given, for $r \in [0,1]$ and $s \in \{0,1\}$, by $\ell(r,s) = 0$ if $r - 1/2$ and $s - 1/2$ have the same sign, and $\ell(r,s) = 1$ otherwise. We call this the *threshold loss*. Thus, with respect to the threshold loss, $\ell(h(x),y) \in \{0,1\}$ is 0 precisely when the thresholded function $T_h : x \mapsto \text{sign}(h(x) - 1/2)$ has value y. This approach is equivalent to using the binary learning problem involving the class $T_H = \{t_h : h \in H\}$. We shall see, however, that there is some advantage in considering the *margin* of classification by these real-valued hypotheses. (This is consistent with the assumption that large margins are good, a fact that has been emphasised for some time in pattern recognition and learning [15, 39], and which is very important in Support Vector Machines [13].)

Explicitly, suppose that $\gamma > 0$, and for $r \in [0,1]$, define $\text{mar}(r,1) = r - 1/2$ and $\text{mar}(r,0) = 1/2 - r$. The *margin* of $h \in H$ on $z = (x,y) \in Z \times \{0,1\}$ is defined to be $\text{mar}(f(x),y)$. Now, define the loss function ℓ^γ by $\ell^\gamma(r,s) = 1$ if $\text{mar}(r,s) < \gamma$ and $\ell^\gamma(r,s) = 0$ if $\text{mar}(r,s) \geq \gamma$. If $L^\gamma(h)$ is the corresponding loss of a hypothesis, then $L^\gamma(h)$ is the probability that for a random $z = (x,y)$, $h(x)$ is not within $1/2 - \gamma$ of y. We call $L^\gamma(h)$ the loss of h at margin γ. Clearly $L^\gamma(h)$ is an increasing function of γ and $L^0(h) = L(h)$ where L corresponds to the simple threshold loss. We make the following definition (as in [3]).

Definition 1.2 *We say that* $\mathcal{A} : (0,1) \times \bigcup_{n=1}^{\infty} Z^n \to H$ *is a* successful real-valued classification algorithm *if for all* $\epsilon, \delta \in (0,1)$*, there is some* $n_0(\epsilon,\delta)$ *(depending on* ϵ *and* δ *only) such that, if* $n > n_0(\epsilon,\delta)$*, then with probability at least* $1-\delta$*,*

$$(L(\mathcal{A}(\gamma,\mathbf{z})) \leq \inf_{h \in H} L^\gamma(h) + \epsilon.$$

So the aim here is to produce an output hypothesis whose loss (in the standard sense, with respect to the threshold loss function) can (with high probability) be made as close as we like to the lowest loss achievable when measured at margin γ.

Definition 1.1 has its origins in the work of Vapnik and Chervonenkis [40, 38, 39] and, in the theoretical science community, in the binary realisable case, in work of Valiant [37]. Valiant's paper was concerned also with the computational complexity of producing successful hypotheses, and this has subsequently been much studied in the Computational Learning Theory community, where the term 'Probably Approximately Correct (PAC) algorithm' has been used. A general loss-theoretical model was developed by Haussler [20]. Many other variants of the learning model have also been explored. (See the books [3, 4, 25, 41] and the Proceedings of the Annual COLT conferences, for wide-ranging results on the computational and sample complexity of a number of variants of this standard learning model.)

2 Uniform Convergence Results

2.1 Uniform Glivenko-Cantelli Classes

Much work in proving successful learnability and in quantifying sample complexity and estimation error has used existing or new 'uniform convergence' or Glivenko-Cantelli type theorems from probability theory.

In order to describe what this means, we need a few more notations. Suppose that F is a set of (measurable) functions from Z to $[0,1]$ and that μ is a probability measure on Z. Denote the expectation $\mathbb{E}_\mu f$ by $\mu(f)$ and, for $\mathbf{z} = (z_1, z_2, \ldots, z_n) \in Z^n$, let us denote by $\mu_n(f)$ the empirical measure of f on $\mathbf{z}$, $\mu_n(f) = n^{-1}\sum_{i=1}^n f(z_i)$. (For fixed f, we shall regard $\mu_n(f)$ as a random variable on Z^n. The notation is not ideal in that in does not specify $\mathbf{z}$, but it will do for our purposes.)

Definition 2.1 *We say that F is a* uniform Glivenko-Cantelli *class if it has the following property (also known as uniform convergence of empiricals to expectations):*

$$\forall \epsilon > 0 \lim_{n\to\infty} \sup_\mu \mathbb{P}\left(\sup_{m\geq n} \sup_{f\in F} |\mu(f) - \mu_m(f)| > \epsilon \right) = 0.$$

The strong law of large numbers of classical probability theory tells us that, for each μ and for each fixed f, $\mathbb{P}\left(\sup_{m\geq n} |\mu(f) - \mu_m(f)| > \epsilon\right) \to 0$ as $n \to \infty$. For a class to be a uniform Glivenko-Cantelli class, we must, additionally, be able to bound the rate of convergence uniformly over all $f \in F$, and over all probability measures μ.

If F is finite, then it is a uniform Glivenko-Cantelli class. To see this explicitly, we can use Hoeffding's inequality [22], which tells us that for any μ and for each $f \in F$, $\mathbb{P}\left(|\mu(f) - \mu_n(f)| > \epsilon\right) < 2e^{-2\epsilon^2 n}$. It follows that

$$\begin{aligned}
\mathbb{P}\left(\sup_{f\in F} |\mu(f) - \mu_n(f)| > \epsilon\right) &= \mathbb{P}\left(\bigcup_{f\in F} \{|\mu(f) - \mu_n(f)| > \epsilon\}\right) \\
&\leq \sum_{f\in F} \mathbb{P}\left(|\mu(f) - \mu_n(f)| > \epsilon\right) \\
&\leq 2|F|e^{-2\epsilon^2 n}.
\end{aligned}$$

We now apply the Borell-Cantelli lemma, together with the observation that the bound just given is independent of μ. Since, for each $\epsilon > 0$,

$$\sum_{n=1}^{\infty} \mathbb{P}\left(|\mu(f) - \mu_n(f)| > \epsilon\right) < \sum_{n=1}^{\infty} 2|F|e^{-2\epsilon^2 n} < \infty,$$

we have

$$\lim_{n\to\infty} \sup_{\mu} \mathbb{P}\left(\sup_{m\geq n} \sup_{f\in F} |\mu(f) - \mu_m(f)| > \epsilon\right) = 0$$

and the class has the uniform Glivenko-Cantelli property.

The above derivation was straightforward, but it demonstrates a key technique: we bounded the probability of a union by the sum of the probabilities of the events involved. This is known as using the *union bound*, and is an extremely useful tool. It has been argued by a number of researchers that the union bound is at the root of much of the looseness of many of the rate of convergence results; see, for example [30, 27], and it is not hard to see that this is a reasonable accusation.

2.2 Uniform Glivenko Cantelli Classes and Successful Learning

With the notations above, define the *loss class* (corresponding to ℓ and H) to be $\ell_H = \{\ell_h : h \in H\}$ where, for $z = (x, y)$, $\ell_h(z) = \ell(h(x), y)$. Suppose that ℓ_H is a uniform Glivenko-Cantelli class. For $\mathbf{z} \in Z^n$, the *empirical loss* of $h \in H$ on $\mathbf{z}$ is defined to be $L_{\mathbf{z}}(h) = \mu_n(\ell_h) = n^{-1}\sum_{i=1}^n \ell(h(x_i), y_i)$, where $z_i = (x_i, y_i)$. Let us say that $\mathcal{A}$ is an *approximate empirical loss minimisation* algorithm if for all $\mathbf{z} \in Z^n$,

$$L_{\mathbf{z}}(\mathcal{A}(\mathbf{z})) < \frac{1}{n} + \inf_{h\in H} L_{\mathbf{z}}(h).$$

Then $\mathcal{A}$ is a successful learning algorithm. (In the binary case, the infimum is a minimum, and the $1/n$ is not needed.) To see this, we can give a fairly standard argument

(see [1, 3], for example). Suppose that $\epsilon > 0$ and $\delta > 0$ are given and let $h^* \in H$ be such that $L(h^*) < L^* + \epsilon/4$. Suppose $n > 4/\epsilon$, so that $1/n < \epsilon/4$. By the uniform Glivenko-Cantelli property for ℓ_H, there is $n_0(\epsilon/4, \delta)$ such that for all $n > n_0$, with probability at least $1 - \delta$, $\sup_{h \in H} |L(h) - L_{\mathbf{z}}(h)| < \epsilon/4$. So, with probability at least $1 - \delta$,

$$\begin{aligned} L(\mathcal{A}(\mathbf{z})) &< L_{\mathbf{z}}(\mathcal{A}(\mathbf{z})) + \frac{\epsilon}{4} \\ &< \left(\inf_{h \in H} L_{\mathbf{z}}(h) + \frac{1}{n} \right) + \frac{\epsilon}{4} \\ &< L_{\mathbf{z}}(h^*) + 2\frac{\epsilon}{4} \\ &< \left(L(h^*) + \frac{\epsilon}{4} \right) + \frac{\epsilon}{2} \\ &< \left(L^* + \frac{\epsilon}{4} \right) + \frac{3\epsilon}{4} \\ &= L^* + \epsilon. \end{aligned}$$

So $\mathcal{A}$ is a successful learning algorithm and its sample complexity is no more than $\max(4/\epsilon, n_0(\epsilon/4, \delta))$.

In fact, as we have stated the definition of learning, the apparently weaker form of convergence

$$\forall \epsilon > 0 \lim_{n \to \infty} \sup_{\mu} \mathbb{P} \left(\sup_{f \in F} |\mu(f) - \mu_n(f)| > \epsilon \right) = 0$$

, where $F = \ell_H$, suffices for this type of learning algorithm to be successful. We may regard the infinite cartesian product Z^∞ to be equipped with a measure μ^∞ which coincides with the product measure μ^n on the canonical projection onto X^n. (If Σ is the σ algebra on Z, then the appropriate σ algebra on X^∞ is that generated by all cylinders, of the form $\Pi_{i=1}^\infty A_i$, where $A_i \in \Sigma$ for all i, with $A_i = Z$ for all but finitely many i. See [41].) Then, with the appropriate interpretation of μ_n as a random variable on Z^∞ (namely, $\mu_n(f)$ is defined precisely as before, and depends only on the first n components of an element of X^∞), the weaker condition described above is equivalent to uniform convergence *in probability* of $\sup_{f \in F} |\mu(f) - \mu_n(f)|$ to 0, whereas the uniform Glivenko-Cantelli property is equivalent to *almost sure* uniform convergence. Since the uniform convergence in probability is a sufficient condition for learning, it is the rate of this type of convergence that we often bound if we are interested primarily in applications to learning.

For the real-valued classification learning problem, it can be shown (see [3]) that a different kind of convergence result is a sufficient condition for the existence of a successful learning algorithm. Explicitly, we say that $\mathcal{A}$ is a *large-margin approximate loss minimisation algorithm* if $L_{\mathbf{z}}^\gamma(\mathcal{A}(\mathbf{z})) = \min_{h \in H} L_{\mathbf{z}}^\gamma(h)$. It turns out that a sufficient

condition for such an algorithm to be sucessful is that

$$\lim_{n\to\infty} \sup_{\mu} \mathbb{P}\left(\sup_{h\in H} \left(L(h) - L_{\mathbf{z}}^{\gamma}(h)\right) > \epsilon \right) = 0$$

for all $\gamma, \epsilon \in (0,1)$.

3 Probabilistic Techniques

In this section we discuss some of the key methods than have been used to prove the probability theorems used in learning theory (such as Glivenko-Cantelli results, but also the 'data-dependent' results to be discussed later).

3.1 Symmetrization

A key technique is *symmetrization*, in which the probability that $\mu(f)$ and $\mu_n(f)$ differ significantly is bounded (uniformly over F) by the probability of an event involving only empirical measures of f. Symmetrization can also be used to bound the expectation of $\sup_{f\in F} |\mu(f) - \mu_n(f)|$.

A symmetrization result for the tail probabilities [19] may now be obtained as follows. First, it is quite easy to show that

$$\mathbb{P}\left(\sup_{f\in F} |\mu(f) - \mu_n(f)| > \epsilon \right) \leq 2\,\mathbb{P}\left(\sup_{f\in F} |\mu_n'(f) - \mu_n(f)| > \frac{\epsilon}{2} \right),$$

where $\mu_n'(f)$ is the empirical measure of f on a second, independent, $\mathbf{z}' \in Z^n$, and the probability on the right is with respect to the product measure μ^{2n} on Z^{2n}. For $1 \leq i \leq n$, let $\sigma_i \in \{-1, 1\}$ be *Rademacher* random variables, taking value 1 with probability $1/2$ and -1 with probability $1/2$. Then, by symmetry and the definition of the empirical measures,

$$\begin{aligned}
\mathbb{P}\left(\sup_{f\in F} |\mu_n'(f) - \mu_n(f)| > \epsilon/2 \right) &= \mathbb{P}\left(\sup_{f\in F} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i \left(f(z_i') - f(z_i)\right) \right| \geq \epsilon/2 \right) \\
&\leq \mathbb{P}\left(\sup_{f\in F} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i f(z_i') \right| \geq \epsilon/4 \right) + \mathbb{P}\left(\sup_{f\in F} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i f(z_i) \right| \geq \epsilon/4 \right) \\
&= 2\,\mathbb{P}\left(\sup_{f\in F} \frac{1}{n} \left| \sum_{i=1}^{n} \sigma_i f(z_i) \right| \geq \epsilon/4 \right).
\end{aligned}$$

(Here the probability is jointly over the distributions of the samples, and of the σ_i.) Summarising, we have:

Theorem 3.1 *If F is a class of functions mapping from Z to $[0,1]$ and μ is a probability measure on Z, then*

$$\mathbb{P}\left(\sup_{f\in F}|\mu(f)-\mu_n(f)|>\epsilon\right)\leq 4\,\mathbb{P}\left(\sup_{f\in F}\frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i f(z_i)\right|\geq\epsilon/4\right),$$

where the σ_i are idependent Rademacher variables.

In a similar way (see [14]) a symmetrization for expectations is obtainable. We have

$$\begin{aligned}\mathbb{E}\left(\sup_{f\in F}|\mu(f)-\mu_n(f)|\right) &\leq \mathbb{E}\left(\frac{1}{n}\sup_{f\in F}\left|\sum_{i=1}^{n}\sigma_i(f(z_i')-f(z_i))\right|\right)\\ &= \frac{2}{n}\,\mathbb{E}\sup_{f\in F}\left|\sum_{i=1}^{n}\sigma_i f(z_i)\right|.\end{aligned}$$

That is,

Theorem 3.2 *If F maps from Z to $[0,1]$, and μ is a probability measure on Z, then*

$$\mathbb{E}\left(\sup_{f\in F}|\mu(f)-\mu_n(f)|\right)\leq\frac{2}{n}\,\mathbb{E}\sup_{f\in F}\left|\sum_{i=1}^{n}\sigma_i f(z_i)\right|.$$

3.2 Concentration

We now describe a type of *concentration of measure result* that can be used to move from bounds on the expectation to bounds on the tail probability (as in [14]). It can also be used in many other ways, such as to obtain data-dependent bounds on estimation error (as described later). A concentration result of this type states that, under certain conditions, a random variable is sharply concentrated about its expectation, in the sense that the probability of a certain deviation from its expectation is exponentially small in the deviation. The most well-known such result is Hoeffding's inequality.

Theorem 3.3 (Hoeffding's inequality) *Suppose X_i, for $i=1,2,\ldots,n$, are independent random variables such that $X_i\in[a_i,b_i]$. Then the random variable $S_n=\sum_{i=1}^{n}X_i$ satisfies*

$$\mathbb{P}\left(|S_n-\mathbb{E}S_n|>\alpha\right)<2\exp\left(-2\alpha^2\Big/\sum_{i=1}^{n}(b_i-a_i)^2\right).$$

An important generalization of Hoeffding's inequality is the following result from [29]. We say that a function $g : Z^n \to \mathbb{R}$ has the *bounded differences* property if for $1 \leq i \leq n$, there are constants c_i such that for any $\mathbf{z}, \mathbf{z}' \in Z^n$ which differ only in the ith coordinate (so $z_i \neq z_i'$ but $z_j = z_j'$ for all $j \neq i$), we have $|g(\mathbf{z}) - g(\mathbf{z}')| \leq c_i$.

Theorem 3.4 (Bounded differences inequality) *Suppose that $z_1, z_2, \ldots, z_n$ are independent, and that the function $g : Z^n \to \mathbb{R}$ has the bounded differences property. Then*

$$\mathbb{P}\left(|g(\mathbf{z}) - \mathbb{E}g(\mathbf{z})| \geq \alpha\right) < 2\exp\left(-2\alpha^2 \Big/ \sum_{i=1}^{n} c_i^2\right),$$

for all α.

In particular, as observed in [14], if we take $g(\mathbf{z}) = \sup_{f \in F} |\mu(f) - \mu_n(f)|$, and note that g has the bounded differences property with $c_i = 1/n$, we obtain that

$$\mathbb{P}\left(\left|\sup_{f \in F} |\mu(f) - \mu_n(f)| - \mathbb{E}\sup_{f \in F} |\mu(f) - \mu_n(f)|\right| > \alpha\right) < 2e^{-2n\alpha^2}.$$

Thus, a bound on the expectation of $\sup_{f \in F} |\mu(f) - \mu_n(f)|$ will yield a good bound on the corresponding tail probability. Explicitly, we have the following.

Theorem 3.5 *Suppose that F is a set of functions from Z to $[0,1]$ and that μ is a probability measure on Z. For $\delta \in (0,1)$, with probability at least $1 - \delta$,*

$$\sup_{f \in F} |\mu(f) - \mu_n(f)| < \sqrt{\frac{1}{2n} \ln\left(\frac{2}{\delta}\right)} + \mathbb{E}\sup_{f \in F} |\mu(f) - \mu_n(f)|.$$

3.3 Using Covering Numbers

Given a (pseudo-)metric space (A, d) and a subset S of A, we say that the set $T \subseteq A$ is an *ϵ-cover* for S (where $\epsilon > 0$) if, for every $s \in S$ there is $t \in T$ such that $d(s,t) < \epsilon$. For a fixed $\epsilon > 0$ we denote by $\mathcal{N}(S, \epsilon, d)$ the cardinality of the smallest ϵ-cover for S. (We define $\mathcal{N}(S, \epsilon, d)$ to be ∞ if there is no such cover.) In our setting, for $\mathbf{z} \in Z^n$, and $f \in F$, let $f|_{\mathbf{z}} = (f(z_1), f(z_2), \ldots, f(z_n))$ and let $F|_{\mathbf{z}} = \{f|_{\mathbf{z}} : f \in F\} \subseteq [0,1]^n$. For $r \geq 1$, let

$$d_r(\mathbf{v}, \mathbf{w}) = \left(\frac{1}{n}\sum_{i=1}^{n} |v_i - w_i|^r\right)^{1/r},$$

and let $d_\infty(\mathbf{v}, \mathbf{w}) = \max_{1 \leq i \leq n} |v_i - w_i|$. Define the *uniform covering number* $\mathcal{N}_r(F, \epsilon, n)$ to be $\sup_{\mathbf{z} \in Z^n} \mathcal{N}(F|_{\mathbf{z}}, \epsilon, d_r)$. Note that if $r > s$ then

$$d_s(\mathbf{v}, \mathbf{w}) \leq d_r(\mathbf{v}, \mathbf{w}) \leq d_\infty(\mathbf{v}, \mathbf{w})$$

and, consequently,

$$\mathcal{N}_s(F,\epsilon,n) \le \mathcal{N}_r(F,\epsilon,n) \le \mathcal{N}_\infty(F,\epsilon,n).$$

The following result is from the excellent survey by Mendelson [30], and uses techniques developed in [40, 32, 20] and elsewhere.

Theorem 3.6 ([30]) *Suppose that F is a set of functions from Z to $[0,1]$. Then for any $\epsilon \in (0,1)$,*

$$\mathbb{P}\left(\sup_{f\in F}|\mu(f)-\mu_n(f)|>\epsilon\right) \le 8\,\mathbb{E}_{\mu^n}\left(\mathcal{N}(F|_{\mathbf{z}},\epsilon/8,d_1)\right)\,e^{-n\epsilon^2/128}.$$

Proof: Using the symmetrization bound of Theorem 3.1, and conditioning on $\mathbf{z}$, the required probability may be bounded by

$$\mathbb{P}\left(\sup_{f\in F}\left|\sum_{i=1}^n \sigma_i f(z_i)\right|>\frac{n\epsilon}{4}\right) = \mathbb{E}\left(\mathbb{P}\left(\sup_{f\in F}\left|\sum_{i=1}^n \sigma_i f(z_i)\right|>\frac{n\epsilon}{4}\,\middle|\,\mathbf{z}\right)\right) = \mathbb{E}\,P(\mathbf{z}),$$

say. So, fix $\mathbf{z} \in Z^n$ and let $C \subseteq [0,1]^n$ be an $\epsilon/8$-cover for $F|_{\mathbf{z}}$, with respect to d_1, of minimum cardinality $\mathcal{N}(F|_{\mathbf{z}},\epsilon/8,d_1)$. It is easy to see that, for $f \in F$, if $|\sum_{i=1}^n \sigma_i f(z_i)| > n\epsilon/4$ then there exists $c \in C$ such that $|\sum_{i=1}^n \sigma_i c_i| > n\epsilon/4$. (Choose c within d_1-distance $\epsilon/8$ of $f|_{\mathbf{z}}$.) Using the union bound,

$$P(\mathbf{z}) \le \mathbb{P}\left(\bigcup_{c\in C}\left\{\left|\sum_{i=1}^n \sigma_i c_i\right|>\frac{n\epsilon}{8}\right\}\right) \le \sum_{c\in C}\mathbb{P}\left(\left|\sum_{i=1}^n \sigma_i c_i\right|>\frac{n\epsilon}{8}\right) \le |C|\,2e^{-n\epsilon^2/128},$$

where we have used Hoeffding's inequality in the final step. The result now follows. □

Of course, μ is not normally known, so the expectation in Theorem 3.6 cannot be determined. Consequently, we usually upper-bound the expected covering number by the uniform covering number $\mathcal{N}_1(F,\epsilon/8,n)$. The theorem then implies the following.

Theorem 3.7 *With probability at least $1-\delta$,*

$$\sup_{f\in F}|\mu(f)-\mu_n(f)| < \sqrt{\frac{64}{n}\left(\ln\mathcal{N}_1(F,\epsilon/8,n)\right)+\ln\left(\frac{8}{\delta}\right)}.$$

In the binary case, better bounds are possible; see [3, 14]. In particular [14, 17], using a technique known as *chaining*, the following can be shown.

Theorem 3.8 *If F is a set of functions from Z to $\{0,1\}$ and μ is a probability measure on Z, then*

$$\mathbb{E} \sup_{f \in F} |\mu(f) - \mu_n(f)| \leq \frac{24}{\sqrt{n}} \max_{\mathbf{z} \in Z^n} \int_0^1 \sqrt{\ln \left(2\mathcal{N}(F|_{\mathbf{z}}, r, d_2)\right)}\, dr.$$

Furthermore, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\sup_{f \in F} |\mu(f) - \mu_n(f)| < \sqrt{\frac{1}{2n} \ln\left(\frac{2}{\delta}\right)} + \frac{24}{\sqrt{n}} \max_{\mathbf{z} \in Z^n} \int_0^1 \sqrt{\ln \left(2\mathcal{N}(F|_{\mathbf{z}}, r, d_2)\right)}\, dr.$$

The second statement in the Theorem follows from the first on applying Theorem 3.5.

Next, we have the following result [6, 3] which concerns real-valued classification. (See [6, 35] for similar results.) First, for $\gamma > 0$, define $\pi_\gamma : [0,1] \to [1/2-\gamma, 1/2+\gamma]$ to be the function that truncates at $1/2-\gamma$ and $1/2+\gamma$; that is, for $x \in (1/2-\gamma, 1/2+\gamma)$, $\pi_\gamma(x) = x$, but for $x \geq 1/2+\gamma$, $\pi_\gamma(x) = 1/2+\gamma$ and for $x \leq 1/2-\gamma$, $\pi_\gamma(x) = 1/2-\gamma$. Let $\pi_\gamma(H)$ be the set of functions $\pi_\gamma \circ h$, obtained by composing H with π_γ. Then we have the following theorem.

Theorem 3.9 *Suppose that H is a set of functions from Z to $[0,1]$. Then for any $\epsilon, \gamma \in (0,1)$,*

$$\mathbb{P}\left(\sup_{h \in H} \left(L(h) - L_{\mathbf{z}}^{\gamma}(h)\right) > \epsilon\right) \leq 2\, \mathbb{E}_{\mu^{2n}} \left(\mathcal{N}(F|_{\mathbf{w}}, \gamma/2, d_\infty)\right) e^{-n\epsilon^2/8},$$

where $F = \pi_\gamma(H)$.

Proof: We only sketch the proof; for details, see [6, 3]. The first step is to establish that

$$\mathbb{P}\left(\sup_{h \in H} \left(L(h) - L_{\mathbf{z}}^{\gamma}(h)\right) > \epsilon\right) \leq \mathbb{P}\left(\sup_{h \in H} \left(L_{\mathbf{z}'}(h) - L_{\mathbf{z}}^{\gamma}(h)\right) > \frac{\epsilon}{2}\right),$$

where, on the right, $\mathbf{z}'$ is a second independent sample of length n. A key observation in the remainder of the proof is that if C is a $\gamma/2$-cover of $F|_{\mathbf{z}\mathbf{z}'}$ with respect to d_∞, then the existence of $h \in H$ such that $L_{\mathbf{z}'}(h) > L_{\mathbf{z}}^{\gamma}(h) + \epsilon/2$ implies the existence of $c \in C$ with the property that $m(c, \mathbf{z}') - m(c, \mathbf{z}) > n\epsilon/2$, where, if $z_i = (x_i, y_i)$, then $m(c, \mathbf{z})$ is the number of indices i such that $|c_i - y_i| > 1/2 - \gamma/2$. The use of Rademacher variables and Hoeffding's inequality completes the proof.

Using the uniform d_∞-covering numbers in place of the expectation, we obtain the following.

Theorem 3.10 *For any $\gamma > 0$, for any $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$\forall h \in H, \quad L(h) < L_{\mathbf{z}}^{\gamma}(h) + \sqrt{\frac{8}{n}\left(\ln \mathcal{N}_{\infty}(\pi_{\gamma}(H), \gamma/2, 2n)\right) + \ln\left(\frac{2}{\delta}\right)}.$$

3.4 Rademacher Complexity

Define, for $\mathbf{z} \in Z^n$,

$$R_n(F, \mathbf{z}) = \frac{2}{n}\mathbb{E}\sup_{f \in F}\left|\sum_{i=1}^{n} \sigma_i f(z_i)\right|,$$

where the expectation is over the joint distribution of the σ_i, and define the *Rademacher complexity* (or Rademacher average) of F to be $R_n(F) = \mathbb{E}R_n(F, \mathbf{z})$ (where here the expectation is over Z^n, with respect to μ^n). (See, for example [30, 9, 26, 36].) As Bartlett and Mendelson have observed [9], the Rademacher complexity gives an indication of how well some function in F can be correlated with random noise, and so provides an indication of the complexity of F. By Theorem 3.2, we see directly that $\mathbb{E}\sup_{f \in F}|\mu(f) - \mu_n(f)|$ is bounded above by $R_n(F)$. By Theorem 3.5, with probability at least $1-\delta$, for $\mathbf{z} \in Z^n$ we obtain

$$\sup_{f \in F}|\mu(f) - \mu_n(f)| < R_n(F) + \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\delta}\right)}.$$

In particular, in the context of learning, and with the usual notation, we have (as in [9]) the following result.

Theorem 3.11 *With probability at least $1-\delta$,*

$$\forall h \in H, \;\; L(h) < L_{\mathbf{z}}(h) + R_n(F) + \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\delta}\right)}.$$

The Rademacher complexity possesses some useful structural properties; for example, the Rademacher complexities of a function class and its symmetric convex hull are the same [9]. Estimates of the Rademacher complexity for a number of function classes, including neural networks, can be found in [9].

More recently, attention has turned to *localized* Rademacher complexities, in which the supremum is taken not over the whole of F, but over a subset of those f with small variance. For details, see [30, 7, 12].

3.5 Combinatorial Measures of Function Class Complexity

We have seen that the covering numbers and Rademacher complexity can be used to bound the probabilities of chief interest. These can in turn be bounded by using certain combinatorial measures of function class complexity. We shall focus here on the bounding of covering numbers, but see [30] for results relating Rademacher complexities to combinatorial parameters.

Let's start with the binary case, in which H maps into $\{0,1\}$. Vapnik and Chervonenkis [40] established that what has subsequently been known as the Vapnik-Chervonenkis dimension (or VC-dimension) is a key measure of function class complexity. (The importance for learning theory was highlighted in [10], and expositions may be found in the books [4, 3, 25, 41], and elsewhere.) In this case, for $\mathbf{z} \in Z^n$, the set $F|_{\mathbf{z}}$ is finite, of cardinality at most 2^n, and we may define the *growth function* $\Pi_F : \mathbb{N} \to \mathbb{N}$ by

$$\Pi_F(n) = \max_{\mathbf{z} \in Z^n} |F|_{\mathbf{z}}| \,.$$

It is clear that $\mathcal{N}(F|_{\mathbf{z}}, \epsilon, d_r) = |F|_{\mathbf{z}}|$ and $\mathcal{N}_r(F, \epsilon, n) = \Pi_F(n)$, for all r and for $\epsilon \in (0,1)$. The *VC-dimension* $\mathrm{VCdim}(F)$ is then defined to be (infinity, or) the largest d such that $\Pi_F(d) = 2^d$. The Sauer-Shelah lemma [33, 34] asserts that if $\mathrm{VCdim}(F) = d < \infty$ then for all $n \geq d$,

$$\Pi_F(n) \leq \sum_{i=0}^{d} \binom{n}{i},$$

showing that the growth function is polynomial in this case. For another description of VC-dimension, we may say that a subset S of Z is *shattered* by F if for any $T \subseteq S$ there is $f_T \in F$ with $f_T(z) = 1$ for $z \in T$ and $f_T(z) = 0$ for $z \in S \setminus T$. Then the VC-dimension is the largest cardinality of a shattered set. Note that, with the discrete loss, it is easy to see that if $F = \ell_H$ then $\Pi_F = \Pi_H$ and so $\mathrm{VCdim}(F) = \mathrm{VCdim}(H)$. Now, Theorem 3.6 has the following consequence.

Theorem 3.12 *For a binary class of finite VC-dimension d, with probability at least $1-\delta$,*

$$\sup_{f \in F} |\mu(f) - \mu_n(f)| < k\sqrt{\frac{1}{n}\left(d \ln\left(\frac{n}{d}\right) + \ln\left(\frac{1}{\delta}\right)\right)},$$

for a fixed constant k.

In this Booolean case, tighter bounds can be obtained (see [28]). Dudley [17] proved (in a result later improved by Haussler [21]), that if $F : Z \to \{0,1\}$ has VC-dimension d then $\mathcal{N}(F|_{\mathbf{z}}, r, d_2) \leq (4e/r^2)^{d/(1-e^{-1})}$, which in combination with Theorem 3.8, establishes the following.

Theorem 3.13 *Suppose F is a binary class of finite VC-dimension d. Then, with probability at least $1-\delta$,*

$$\sup_{f\in F}|\mu(f)-\mu_n(f)| < \sqrt{\frac{1}{2n}\ln\left(\frac{2}{\delta}\right)} + c\sqrt{\frac{d}{n}},$$

for a fixed constant c.

Lower bounds on the sample complexity of learning algorithms can also be obtained in terms of the VC-dimension [18, 10, 3]. The VC-dimensions of many different types of neural network have been estimated; see [3, 23, 2], for example.

Suppose, more generally, that $F : Z \to [0,1]$. We say that $S \subseteq Z$ is *shattered* by F if there are numbers $r_z \in [0,1]$ for $z \in S$ such that for every $T \subseteq S$ there is some $f_T \in F$ with the property that $f_T(z) \geq r_z$ if $z \in T$ and $f_T(z) < r_z$ if $z \in S \setminus T$. We say that F has finite *pseudo-dimension* $d(F) = d$ if d is the maximum cardinality of a shattered set. Another interpretation of this dimension can be given. Let $S(F) = \{\{(z,y) \in Z \times \mathbb{R} : y \leq f(z)\} : f \in F\}$ be the set of *subgraphs* of functions in F. Then $d(F)$ is the VC-dimension of $S(F)$ (or, of the set of indicator functions of the sets in $S(F)$). For this reason, a class F of finite pseudo-dimension is often called a *VC-subgraph* class [16, 36]. Pollard [32] bounded the d_1-covering numbers in terms of the pseudo-dimension, as follows:

$$\mathcal{N}_1(F,\epsilon,n) < 2\left(\frac{2e}{\epsilon}\ln\left(\frac{2e}{\epsilon}\right)\right)^{d(F)},$$

for all n. One can therefore use this in conjunction with Theorem 3.6 to obtain bounds on the rate of convergence, in probability, of $\sup_{f\in F}|\mu(f)-\mu_n(f)|$ to zero. Furthermore, the Borel-Cantelli lemma can be applied to show that F will be a uniform Glivenko-Cantelli class if it has finite pseudo-dimension.

However, finite pseudo-dimension is a stronger condition than is needed for a class to have the uniform Glivenko-Cantelli property [1]. A *scale-sensitive* version of the pseudo-dimension (originally used in [24]) is defined as follows. For $\gamma > 0$, we say that $S \subseteq Z$ is *γ-shattered* by F if there are numbers $r_z \in [0,1]$ for $z \in S$ such that for every $T \subseteq S$ there is some $f_T \in F$ with the property that $f_T(z) \geq r_z + \gamma$ if $z \in T$ and $f_T(z) < r_z - \gamma$ if $z \in S \setminus T$. We say that F has finite *fat-shattering dimension* d at scale γ, and we write $\mathrm{fat}_\gamma(F) = d$, if d is the maximum cardinality of a γ-shattered set. We say simply that F has finite fat-shattering dimension if it has finite fat-shattering dimension at every scale $\gamma > 0$. It is quite possible for a class to have finite fat-shattering dimension but infinite pseudo-dimension. Alon *et al.* [1] obtained an upper bound on the d_∞ covering numbers in terms of the fat-shattering dimension, as a consequence of which if F has range which is a sub-interval of $[0,1]$ of length B, then

$$\mathcal{N}_1(\epsilon,F,n) \leq \mathcal{N}_\infty(\epsilon,F,n) < 2\left(\frac{4nB^2}{\epsilon^2}\right)^{d\log_2(4eBn/(d\epsilon))},$$

where $d = \mathrm{fat}_{\epsilon/4}(F)$. Theorem 3.6 and the Borel-Cantelli lemma then establish that F is a uniform Glivenko-Cantelli class.

For the standard loss functions, the fat-shattering dimensions of the loss class ℓ_H and H itself are often simply related; see [1, 3].

We can apply the covering number bound to real classification learning by using Theorem 3.10, leading [3] to the following.

Theorem 3.14 *With probability at least* $1 - \delta$*,*

$$\forall h \in H, \quad L(h) < L_{\mathbf{z}}^{\gamma}(h) + \sqrt{\frac{8}{n}\left(d \log_2\left(\frac{32en}{d}\right)\ln(128n) + \ln\left(\frac{4}{\delta}\right)\right)},$$

where $d = \mathrm{fat}_{\gamma/8}(H)$.

Alon *et al.* also showed that finite fat-shattering dimension is a necessary condition for a class to have the uniform Glivenko-Cantelli property. In fact, they prove something stronger, a consequence of which is that finite fat-shattering dimension is a necessary condition for a class to have the convergence property

$$\forall \epsilon > 0 \lim_{n\to\infty} \sup_{\mu} \mathbb{P}\left(\sup_{f\in F} |\mu(f) - \mu_n(f)| > \epsilon\right) = 0.$$

Thus the uniform convergence, in probability, of $\sup_{f\in F} |\mu(f) - \mu_n(f)|$ to 0 is equivalent to the uniform Glivenko-Cantelli property (as noted above).

For more on the fat-shattering dimension, including estimates for neural network classes, see [1, 3, 6]. See [3, 8, 30, 31] for improved bounds on covering numbers in terms of the fat-shattering dimension, particularly with respect to the metrics d_p for $p \neq \infty$. The fat-shattering dimension can also be used to provide lower bounds on the sample complexity of learning algorithms; see [3] for instance.

4 Data-Dependent Analysis

4.1 Data-Dependent Bounds

We have seen that for many learning problems the loss of hypotheses may be bounded uniformly in terms of the empirical losses and the expectation of the empirical covering numbers. In most applications, since we do not know the distribution, we bound the expectation of the covering number by the corresponding uniform covering number, and

then perhaps use combinatorial dimensions to bound these. We have also seen that the losses may be bounded using the Rademacher complexity of the loss class. Generally, the upper bounds on $L(h)$ presented so far consist of two terms; one is the empirical loss, and the other is what might be called a complexity term. Notably, although the empirical loss clearly depends on $\mathbf{z}$, the complexity term depends on the loss class, and does not depend explicitly on $\mathbf{z}$. In this section we present some data-dependent results, in which the class-dependent complexity term is replaced by a complexity term dependent not only on the class, but on the sample $\mathbf{z}$ itself. This has been the subject of much active research in recent years. This has been motivated, at least in part, by the observation that learning algorithms tend to return hypotheses that use the training data in a fairly sophisticated manner, rather than simply return, for instance, any hypothesis with near-minimal empirical loss. Data-dependent bounds have been obtained in a number of ways, in particular through deploying a general 'luckiness' framework developed in [35, 42], and, more recently, through the application of concentration inequalities, as in [11, 5].

4.2 Data-Dependent Learning Results

Suppose that H is a binary function class mapping from X to $\{0,1\}$. By proving a new concentration inequality, Boucheron, Lugosi and Massart [11], established that the *VC-entropy* $H_n(\mathbf{x}) = \log_2 |H|_{\mathbf{x}}|$ (for $\mathbf{x} \in X^n$) is concentrated around its expectation. With this, they were able to establish the following data-dependent result (in which the loss function is the discrete loss).

Theorem 4.1 *With probability at least* $1-\delta$*, for* $\mathbf{z} \in Z^n = (X \times \{0,1\})^n$*,*

$$\forall h \in H, \quad L(h) < L_{\mathbf{z}}(h) + \sqrt{\frac{6 \ln |H|_{\mathbf{x}}|}{n}} + 4\sqrt{\frac{\ln(2/\delta)}{n}}.$$

This should be compared with the bounds that would follow from the results presented earlier: such bounds would involve $\mathbb{E}\,|H|_{\mathbf{x}}|$ or, since μ is not known, the growth function $\Pi_H(n) = \max_{\mathbf{x} \in X^n} |H|_{\mathbf{x}}|$, and therefore would not depend explicitly on the data. It is certainly possible that $|H|_{\mathbf{x}}|$ is much less than $\Pi_H(n)$, and so the data-dependent bound could have significant advantage. This result can also be expressed in terms of the *empirical VC-dimension*. For $\mathbf{x} \in X^n$ let $\mathrm{VCdim}(H|\mathbf{x})$ denote the VC-dimension of the set of functions obtained by restricting H to domain consisting of the elements of $\mathbf{x}$.

Theorem 4.2 *With probability at least* $1-\delta$*, for* $\mathbf{z} \in Z^n = (X \times \{0,1\})^n$*,*

$$\forall h \in H, \quad L(h) < L_{\mathbf{z}}(h) + \sqrt{\frac{6d(\mathbf{x})}{n} \ln\left(\frac{en}{d(\mathbf{x})}\right)} + 4\sqrt{\frac{\log(2/\delta)}{n}},$$

where $d(\mathbf{x}) = \mathrm{VCdim}(H|\mathbf{x})$.

(See also [35] for related results involving empirical VC-dimension, for the case in which the empirical loss is zero.)

There are also data-dependent results for real-valued classification [42, 5]. Using the concentration inequality from [11], Antos, Kégl, Linder and Lugosi [5] have obtained bounds involving the *empirical fat-shattering dimension*. For $\mathbf{x} \in X^n$, and $\gamma > 0$, let $\mathrm{fat}_\gamma(H|\mathbf{x})$ be the fat-shattering dimension of the set of functions obtained by restricting H to the set consisting of the elements of the sample $\mathbf{x}$. Then, in [5], the following theorem is obtained.

Theorem 4.3 *For $\gamma > 0$, with probability at least $1 - \delta$,*

$$\forall h \in H, \quad L(h) < L_{\mathbf{z}}^{\gamma}(h) + \sqrt{\frac{1}{n}\left(9d(\mathbf{x}) + 12.5\ln\left(\frac{8}{\delta}\right)\right)\ln\left(\frac{32en}{d(\mathbf{x})}\right)\ln(128n)},$$

where $d(\mathbf{x}) = \mathrm{fat}_{\gamma/8}(H|\mathbf{x})$.

This should be compared with Theorem 3.14. The former might look better, but the empirical fat-shattering dimension can be significantly less than the fat-shattering dimension, so in some cases the data-dependent bound is better. Moreover, the empirical fat-shattering dimension can be calculated reasonably easily in some cases. (See [5].)

We can also obtain a version of the above result in which the margin γ is not specified beforehand, and could depend on both the data and the chosen hypothesis.

Theorem 4.4 *With probability at least $1 - \delta$, for all $h \in H$ and for all $\gamma \in (0, 1]$,*

$$L(h) < L_{\mathbf{z}}^{\gamma}(h) + \sqrt{\frac{1}{n}\left(9d_1(\mathbf{x}) + 12.5\ln\left(\frac{16}{\delta\gamma}\right)\right)\ln\left(\frac{32en}{d_2(\mathbf{x})}\right)\ln(128n)},$$

where $d_1(\mathbf{x}) = \mathrm{fat}_{\gamma/16}(H|\mathbf{x})$ and $d_2(\mathbf{x}) = \mathrm{fat}_{\gamma/8}(H|\mathbf{x})$.

Proof: We use the 'method of sieves' (see [6, 3]). In [6], the following is shown. Suppose $\mathbb{P}$ is any probability measure and that $\{E(\alpha_1, \alpha_2, \delta) : 0 < \alpha_1, \alpha_2, \delta \leq 1\}$ is a set of events such that:

- for all α, $\mathbb{P}(E(\alpha, \alpha, \delta)) \leq \delta$,
- $0 < \alpha_1 \leq \alpha \leq \alpha_2 < 1$ and $0 < \delta_1 \leq \delta \leq 1$ imply $E(\alpha_1, \alpha_2, \delta_1) \subseteq E(\alpha, \alpha, \delta)$.

Then

$$\mathbb{P}\left(\bigcup_{\alpha\in(0,1]} E(\alpha c, \alpha, \delta\alpha(1-c))\right) \leq \delta$$

for $0 < c, \delta < 1$. We take $E(\alpha_1, \alpha_2, \delta)$ to be the set of $\mathbf{z} \in Z^n$ such that there exists $h \in H$ with

$$L(h) \geq L_{\mathbf{z}}^{\alpha_2}(h) + \sqrt{\frac{1}{n}\left(9\mathrm{fat}_{\alpha_1/8}(H|\mathbf{x}) + 12.5\ln\left(\frac{8}{\delta}\right)\right)\ln\left(\frac{32en}{\mathrm{fat}_{\alpha_2/8}(H|\mathbf{x})}\right)\ln(128n)}.$$

Theorem 4.4 states that $\mathbb{P}(E(\alpha, \alpha, \delta)) \leq \delta$. It is easy to see that $0 < \alpha_1 \leq \alpha \leq \alpha_2 < 1$ and $0 < \delta_1 \leq \delta \leq 1$ imply $E(\alpha_1, \alpha_2, \delta_1) \subseteq E(\alpha, \alpha, \delta)$. The result now follows by using the sieve method, taking $c = 1/2$. □

Turning attention now to the Rademacher complexity, Bartlett and Mendelson [9] have observed that the empirical Rademacher complexity $R_n(F, \mathbf{z})$ is concentrated about its expectation, which is $R_n(F)$. For, it is easy to see that $g(\mathbf{z}) = R_n(F, \mathbf{z})$ satisfies the bounded differences property with each c_i equal to $2/n$, so that with probability at least $1 - \delta$, $R_n(F)$ is at most $R_n(F, \mathbf{z}) + \sqrt{2n^{-1}\ln(2/\delta)}$. Hence, by Theorem 3.5, with probability at least $1 - \delta$,

$$\sup_{f\in F} |\mu(f) - \mu_n(f)| < R_n(F, \mathbf{z}) + 3\sqrt{\frac{1}{n}\ln\left(\frac{2}{\delta}\right)}.$$

In particular, with probability at least $1 - \delta$,

$$\forall h \in H, \quad L(h) < L_{\mathbf{z}}(h) + R_n(F, \mathbf{z}) + 3\sqrt{\frac{1}{n}\ln\left(\frac{2}{\delta}\right)}.$$

Acknowledgements

I am grateful to the organisers of the WIRN 2002 conference. I thank Bob Williamson for useful discussion at a recent Neurocolt workshop, and I acknowledge the support of European Union funding through the Neurocolt 2 Working Group project.

References

[1] Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler: Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM* **44**(5): 616–631. [Extended abstract appears in *Proceedings of the Symposium on Foundations of Computer Science*. IEEE Computer Society Press, Los Alamitos, CA, pp. 292–301, 1993.]

[2] Martin Anthony: Probabilistic analysis of learning in artificial neural networks: the PAC model and its variants. *Neural Computing Surveys*, **1**, 1997.

[3] Martin Anthony and Peter L. Bartlett: *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge UK, 1999.

[4] Martin Anthony and Norman L. Biggs: *Computational Learning Theory: An Introduction*. Cambridge Tracts in Theoretical Computer Science, 30, 1992. Cambridge University Press, Cambridge, UK.

[5] András Antos, Balázs Kégl, Tamás Linder and Gábor Lugosi: Data-dependent margin-based generalization bounds for classification. Preprint, Queen's University at Kingston, Canada. `magenta.mast.queensu.ca/ linder/preprints.html`.

[6] Peter Bartlett: The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE Transactions on Information Theory* **44**(2): 525–536.

[7] Peter L. Bartlett, Olivier Bousquet and Shahar Mendelson: Localized Rademacher complexities. To appear, *Proceedings of the* 15*th Annual Conference on Computational Learning Theory*, ACM Press, New York, NY, 2002.

[8] Peter L. Bartlett and Philip M. Long: More theorems about scale-sensitive dimensions and learning. In *Proceedings of the* 8*th Annual Conference on Computational Learning Theory*, ACM Press, New York, NY, 1995, pp. 392–401.

[9] Peter Bartlett and Shahar Mendelson: Rademacher and Guassian complexities: risk bounds and structural results. In *Proceedings of the* 14*th Annual Conference on Computational Learning Theory*, Lecture Notes in Artificial Intelligence, Springer pp. 224-240, 2001.

[10] Anselm Blumer, Andrzej Ehrenfeucht, David Haussler, and Manfred K. Warmuth: Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, **36**(4): 929–965, 1989.

[11] Stéphane Boucheron, Gábor Lugosi and Pascal Massart: A sharp concentration inequality with applications. *Random Structures and Algorithms*, **16**: 277–292, 2000.

[12] Olivier Bousquet, Vladimir Koltchinskii and Dmitriy Panchenko: Some local measures of complexity on convex hulls and generalization bounds. To appear, *Proceedings of the* 15*th Annual Conference on Computational Learning Theory*, ACM Press, New York, NY, 2002.

[13] Nello Cristianini and John Shawe-Taylor: *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000.

[14] Luc Devroye and Gábor Lugosi: *Combinatorial Methods in Density Estimation*, Springer Series in Statistics, Springer-Verlag, New York, NY, 2001.

[15] Richard O. Duda and Peter E. Hart: *Pattern Classification and Scene Analysis*, John Wiley, 1973.

[16] Richard M. Dudley: *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics, 63, Cambridge University Press, Cambridge, UK, 1999.

[17] Richard M. Dudley: Central limit theorems for empirical measures. *Annals of Probability*, **6**(6): 899–929, 1978.

[18] Andrzej Ehrenfeucht, David Haussler, Michael Kearns, and Leslie Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, **82**: 247–261, 1989.

[19] E. Giné and J. Zinn: Some limit theorems for empirical processes. *Annals of Probability* **12**(4): 929–989, 1984.

[20] David Haussler: Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, **100**(1): 78–150, 1992.

[21] David Haussler: Sphere packing numbers for subsets of the Boolean n-cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory (A)*, **69**(2): 217–232, 1995.

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

[23] Marek Karpinski and Angus MacIntyre: Polynomial bounds for VC dimension of sigmoidal and general Pfaffian neural networks. *Journal of Computer and System Sciences*, **54**: 169–176, 1997.

[24] Michael J. Kearns and Robert E. Schapire: Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, **48**(3): 464–497, 1994. [Extended abstract appear in *Proceedings of the 1990 IEEE Symposium on Foundations of Computer Science*, IEEE Press.

[25] Michael J. Kearns and Umesh Vazirani: *Introduction to Computational Learning Theory*, MIT Press, Cambridge, MA, 1995.

[26] Vladimir Koltchinskii and Dmitry Panchenko: Rademacher processes and bounding the risk of function learning. Technical report, Department of Mathematics and Statistics, University of New Mexico, 2000.

[27] John Langford: *Quantitatively Tight Sample Complexity Bounds*. PhD thesis, Department of Computer Science, Carnegie Mellon University, Pittsburgh, PA.

[28] Gábor Lugosi: *Lectures on Statistical Learning Theory*, presented at the Garchy Seminar on Mathematical Statistics and Applications, August 27-September 1, 2000. (Available from `www.econ.upf.es/ lugosi`.)

[29] Colin McDiarmid: On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics, 1989*, London Mathematical Society Lecture Note Series (141). Cambridge University Press, Cambridge, UK, 1989.

[30] Shahar Mendelson: A few notes on Statistical Learning Theory. Technical Report, Australian National University Computer Science Laboratory.

[31] S. Mendelson and R. Vershynin: Entropy, dimension and the Elton-Pajor theorem. Preprint, Australian National University.

[32] David Pollard: *Convergence of Stochastic Processes*. Springer-Verlag, 1984.

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

[34] S. Shelah: A combinatorial problem: Stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, **41**: 247–261, 1972.

[35] John Shawe-Taylor, Peter Bartlett, Bob Williamson and Martin Anthony: Structural risk minimisation over data-dependent hierarchies. *IEEE Transactions on Information Theory*, **44**(5): 1926–1940, 1998.

[36] Aad W. van der Vaart and Jon A. Wellner: *Weak Convergence and Empirical Processes*, Springer Series in Statistics, Springer-Verlag, New York, NY, 1996.

[37] Leslie G. Valiant: A theory of the learnable. *Communications of the ACM*, **27**(11): 1134–1142, Nov. 1984.

[38] Vladimir N. Vapnik: *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.

[39] Vladimir N. Vapnik: *Statistical Learning Theory*, Wiley, 1998.

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

[41] M. Vidyasagar: *A Theory of Learning and Generalization*, Springer-Verlag, 1996.

[42] Robert Williamson, John Shawe-Taylor, Bernhard Scholkopf, and Alex Smola: *Sample Based Generalization Bounds*, NeuroCOLT Technical Report, NC-TR-99-055, 1999.