

# Mathematical Modelling of Generalization

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

**Abstract.** This paper surveys certain developments in the use of probabilistic techniques for the modelling of generalization. Some of the main methods and key results are discussed. 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

Suppose that  $X$  is a set of *examples* and that  $Y \subseteq [0, 1]$  is a set of possible *outputs*. Elements  $(x, y)$  of  $Z = X \times Y$  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 and produces a function  $h : X \rightarrow [0, 1]$ , chosen from some *hypothesis class*  $H$  of functions. We assume that there is some fixed, but unknown, probability measure<sup>1</sup>  $\mu$  on  $Z$ , and that each training example is generated independently according to  $\mu$ . A learning algorithm is a function  $\mathcal{A} : \bigcup_{n=1}^{\infty} Z^n \rightarrow H$ , where  $H$  is a *hypothesis class* of functions from  $Z$  to  $[0, 1]$ . We have in mind some *loss function*  $\ell : [0, 1] \times Y \rightarrow [0, 1]$ . 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  $r \neq s$ . 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  $\mu$ ). 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. (Here, and in the rest of the paper, we use the symbol  $\mathbb{P}$  to denote probability. In the definition that follows, the probability is with respect to  $\mu^n$ .) This definition has its origins in [35, 33, 32, 19]. (See also the books [3, 4, 21, 36].)

We say that  $\mathcal{A}$  is a *successful* learning algorithm for  $H$  if for all  $\epsilon, \delta \in (0, 1)$ , there is some  $n_0(\epsilon, \delta)$  (depending on  $\epsilon$  and  $\delta$  only) such that, if  $n > n_0(\epsilon, \delta)$ , then with probability at least  $1 - \delta$ ,  $L(\mathcal{A}(\mathbf{z})) \leq L^* + \epsilon$ . Note that if  $\mathcal{A}$  is successful, then there is some function  $\epsilon_0(n, \delta)$  of  $n$  and  $\delta$ , with the property that for all  $\delta$ ,  $\lim_{n \rightarrow \infty} \epsilon_0(n, \delta) = 0$ , and such that for any probability measure  $\mu$  on  $Z$ , with

---

<sup>1</sup> Certain measurability conditions are implicitly assumed in what follows, but these conditions are reasonable and not particularly stringent. Details may be found in [31] for instance.

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* learning problem.

We might want to use real functions for classification. Here, we would have  $Y = \{0, 1\}$ , but  $H : X \rightarrow [0, 1]$ . 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  $y$ . There is some advantage in considering the *margin* of classification by these real-valued hypotheses (a fact that has been emphasised for some time in pattern recognition and learning [34], 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  $\ell^\gamma$  by  $\ell^\gamma(r, s) = 1$  if  $\text{mar}(r, s) < \gamma$  and  $\ell^\gamma(r, s) = 0$  if  $\text{mar}(r, s) \geq \gamma$ . The corresponding loss  $L^\gamma(h)$  of a hypothesis is called the loss of  $h$  at margin  $\gamma$ . We say (as in [3]) that  $\mathcal{A} : (0, 1) \times \bigcup_{n=1}^{\infty} Z^n \rightarrow H$  is a *successful real-valued classification algorithm* if for all  $\epsilon, \delta \in (0, 1)$  there is  $n_0(\epsilon, \delta)$  such that, if  $n > n_0(\epsilon, \delta)$ , then with probability at least  $1 - \delta$ ,  $L(\mathcal{A}(\gamma, \mathbf{z})) \leq \inf_{h \in H} L^\gamma(h) + \epsilon$ .

## 2 Techniques

**Uniform Glivenko-Cantelli classes** Suppose that  $F$  is a set of (measurable) functions from  $Z$  to  $[0, 1]$  and that  $\mu$  is a probability measure on  $Z$ . Denote the expectation  $\mathbb{E}_\mu f$  by  $\mu(f)$  and, for  $\mathbf{z} = (z_1, z_2, \dots, z_n) \in Z^n$ , let us denote by  $\mu_n(f)$  the empirical measure of  $f$  on  $\mathbf{z}$ ,  $\mu_n(f) = n^{-1} \sum_{i=1}^n f(z_i)$ .  $F$  is a *uniform Glivenko-Cantelli* class if it has the following property (also known as uniform convergence of empiricals to expectations) if for all  $\epsilon > 0$ ,  $\lim_{n \rightarrow \infty} \sup_\mu \mathbb{P}(\sup_{m \geq n} \sup_{f \in F} |\mu(f) - \mu_m(f)| > \epsilon) = 0$ .

Define the *loss class* (corresponding to  $\ell$  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  $\ell_H$  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  $\mathbf{z} \in Z^n$ ,  $L_{\mathbf{z}}(\mathcal{A}(\mathbf{z})) < 1/n + \inf_{h \in H} L_{\mathbf{z}}(h)$ . Then it is fairly easy to see [1, 3], that  $\mathcal{A}$  is a successful learning algorithm.

**Symmetrization** A key technique is *symmetrization*. The following symmetrization result for expectations is obtainable(see [14]):

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

where, for  $1 \leq i \leq n$ ,  $\sigma_i \in \{-1, 1\}$  are *Rademacher* random variables, taking value 1 with probability 1/2 and  $-1$  with probability 1/2. (Here the expectation is jointly over the distributions of the samples, and of the  $\sigma_i$ .) Symmetrization results for the tail probabilities  $\mathbb{P}(\sup_{f \in F} |\mu(f) - \mu_n(f)| > \epsilon)$  can also be obtained; see [18, 25].

**Concentration** We now describe a type of *concentration result* (see [24]), a generalisation of Hoeffding's inequality. We shall call it the *bounded differences inequality*. Suppose that a function  $g : Z^n \rightarrow \mathbb{R}$  has the following *bounded differences* property: 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  $i$ th coordinate (so  $z_i \neq z'_i$  but  $z_j = z'_j$  for all  $j \neq i$ ), we have  $|g(\mathbf{z}) - g(\mathbf{z}')| \leq c_i$ . Then, if  $z_1, z_2, \dots, z_n$  are independent, we have

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

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

$$\sup_{f \in F} |\mu(f) - \mu_n(f)| < \sqrt{\frac{1}{2n} \ln\left(\frac{2}{\delta}\right)} + \mathbb{E} \sup_{f \in F} |\mu(f) - \mu_n(f)|. \quad (3)$$

**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  $\epsilon$ -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  $\epsilon$ -cover for  $S$ . (We define  $\mathcal{N}(S, \epsilon, d)$  to be  $\infty$  if there is no such cover.) In our setting, for  $\mathbf{z} \in Z^n$ , and  $f \in F$ , let  $f|_{\mathbf{z}} = (f(z_1), f(z_2), \dots, f(z_n))$  and let  $F|_{\mathbf{z}} = \{f|_{\mathbf{z}} : f \in F\} \subseteq [0, 1]^n$ . For  $r \geq 1$ , let  $d_r(\mathbf{v}, \mathbf{w}) = (n^{-1} \sum_{i=1}^n |v_i - w_i|^r)^{1/r}$ , 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,  $\mathcal{N}_s(F, \epsilon, n) \leq \mathcal{N}_r(F, \epsilon, n) \leq \mathcal{N}_\infty(F, \epsilon, n)$ .

As shown in [25] (using techniques developed in [35, 27, 19] and elsewhere), if  $F$  is a set of functions from  $Z$  to  $[0, 1]$ , then for any  $\epsilon \in (0, 1)$ ,

$$\mathbb{P}\left(\sup_{f \in F} |\mu(f) - \mu_n(f)| > \epsilon\right) \leq 8 \mathbb{E}_{\mu^n} (\mathcal{N}(F|_{\mathbf{z}}, \epsilon/8, d_1)) e^{-n\epsilon^2/128}. \quad (4)$$

This implies that, with probability at least  $1 - \delta$ ,

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

In the binary case, better bounds are possible; see [3, 14] (in particular [14, 16] by using a technique known as *chaining*).

Next, we have the following result [6, 3] which concerns real-valued classification. (See [6, 30] for similar results.) With probability at least  $1 - \delta$ ,

$$\forall h \in H, \quad L(h) < L_{\mathbf{z}}^{\gamma}(h) + \sqrt{\frac{8}{n} (\ln \mathcal{N}_{\infty}(H, \gamma/2, 2n)) + \ln \left( \frac{2}{\delta} \right)}. \quad (5)$$

**Rademacher Complexity** For  $\mathbf{z} \in Z^n$ ,  $\mathbb{E} \sup_{f \in F} |2n^{-1} \sum_{i=1}^n \sigma_i f(z_i)|$  is denoted  $R_n(F, \mathbf{z})$ , where the expectation is over the joint distribution of the  $\sigma_i$ , and the *Rademacher complexity* (or Rademacher average) of  $F$  is defined to be  $R_n(F) = \mathbb{E} R_n(F, \mathbf{z})$  (where here the expectation is over  $Z^n$ , with respect to  $\mu^n$ ). (See, for example [25, 9, 22, 31].) By equation (1), we see directly that  $\mathbb{E} \sup_{f \in F} |\mu(f) - \mu_n(f)|$  is bounded above by  $R_n(F)$ . By (3), with probability at least  $1 - \delta$ , for  $\mathbf{z} \in Z^n$ ,

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

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 [25, 7, 12].

**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 [25] for results relating Rademacher complexities to combinatorial parameters.

For the binary case, Vapnik and Chervonenkis [35] 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, 21, 36], 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*  $\Pi_F : \mathbb{N} \rightarrow \mathbb{N}$  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*  $\text{VCdim}(F)$  is then defined to be (infinity, or) the largest  $d$  such that  $\Pi_F(d) = 2^d$ . The Sauer-Shelah lemma [28, 29] asserts that if  $\text{VCdim}(F) = d < \infty$  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  $\text{VCdim}(F) = \text{VCdim}(H)$ . Now, equation (4) has the following consequence 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 the Boolean case, tighter bounds can be obtained (see [23]).)

Lower bounds on the sample complexity of learning algorithms can also be obtained in terms of the VC-dimension [17, 10, 3]. The VC-dimensions of many different types of neural network have been estimated; see [3, 20, 2], for example.

Suppose, more generally, that  $F : Z \rightarrow [0, 1]$ . For  $\gamma > 0$ , we say that  $S \subseteq Z$  is  $\gamma$ -*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  $\gamma$ , and we write  $\text{fat}_\gamma(F) = d$ , if  $d$  is the maximum cardinality of a  $\gamma$ -shattered set. We say simply that  $F$  has finite fat-shattering dimension if it has finite fat-shattering dimension at every scale  $\gamma > 0$ . Alon *et al.* [1] obtained an upper bound on the  $d_\infty$  covering numbers in terms of the fat-shattering dimension, establishing that  $F$  is a uniform Glivenko-Cantelli class if and only if it has finite fat-shattering dimension. We can apply their results to real classification learning by using (5). This leads [3] to the fact that, with probability at least  $1 - \delta$ ,

$$\forall h \in H, \quad L(h) < L_z^\gamma(h) + \sqrt{\frac{8}{n} \left( d \log_2 \left( \frac{32en}{d} \right) \ln(128n) + \ln \left( \frac{4}{\delta} \right) \right)}, \quad (6)$$

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

For more on the fat-shattering dimension, including estimates for neural network classes, see [1, 3, 6]. See [3, 8, 25, 26] for improved bounds on covering numbers in terms of the fat-shattering dimension, particularly with respect to the metrics  $d_p$  for  $p \neq \infty$ .

### 3 Data-Dependent Learning Bounds

In this section we present some data-dependent results, in which bounds on estimation error are given that depended not only on the hypothesis class, but on the sample  $\mathbf{z}$  itself. Data-dependent bounds have been obtained in a number of ways, in particular through deploying a general ‘luckiness’ framework developed in [30, 37], and, more recently, through the application of concentration inequalities, as in [11, 5].

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): 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  $\mu$  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*.)

There are also data-dependent results for real-valued classification [37, 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  $\text{fat}_{\gamma}(H|\mathbf{x})$  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], it is shown that, 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  $d(\mathbf{x}) = \text{fat}_{\gamma/8}(H|\mathbf{x})$ .

This should be compared with (6). 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  $\gamma$  is not specified beforehand, and could depend on both the data and the chosen hypothesis. Using the ‘method of sieves’ (see [6, 3]), it can be shown that, with probability at least  $1 - \delta$ , the following holds, 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  $d_1(\mathbf{x}) = \text{fat}_{\gamma/16}(H|\mathbf{x})$  and  $d_2(\mathbf{x}) = \text{fat}_{\gamma/8}(H|\mathbf{x})$ .

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

$2/n$ . Hence, by (3), with probability at least  $1 - \delta$ ,

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

## 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.
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](http://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 15th 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 8th Annual Conference on Computational Learning Theory*, ACM Press, New York, NY, 1995, pp. 392–401.
9. Peter Bartlett and Shahar Mendelson: Rademacher and Gaussian complexities: risk bounds and structural results. In *Proceedings of the 14th 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 15th 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 M. Dudley: *Uniform Central Limit Theorems*, Cambridge Studies in Advanced Mathematics, 63, Cambridge University Press, Cambridge, UK, 1999.

16. Richard M. Dudley: Central limit theorems for empirical measures. *Annals of Probability*, **6**(6): 899–929, 1978.
17. 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.
18. E. Giné and J. Zinn: Some limit theorems for empirical processes. *Annals of Probability* **12**(4): 929–989, 1984.
19. David Haussler: Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, **100**(1): 78–150, 1992.
20. 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.
21. Michael J. Kearns and Umesh Vazirani: *Introduction to Computational Learning Theory*, MIT Press, Cambridge, MA, 1995.
22. 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.
23. 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](http://www.econ.upf.es/~lugosi).)
24. 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.
25. Shahar Mendelson: A few notes on Statistical Learning Theory. Technical Report, Australian National University Computer Science Laboratory.
26. S. Mendelson and R. Vershynin: Entropy, dimension and the Elton-Pajor theorem. Preprint, Australian National University.
27. David Pollard: *Convergence of Stochastic Processes*. Springer-Verlag, 1984.
28. N. Sauer: On the density of families of sets. *Journal of Combinatorial Theory (A)*, **13**: 145–147, 1972.
29. S. Shelah: A combinatorial problem: Stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, **41**: 247–261, 1972.
30. 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.
31. Aad W. van der Vaart and Jon A. Wellner: *Weak Convergence and Empirical Processes*, Springer Series in Statistics, Springer-Verlag, New York, NY, 1996.
32. Leslie G. Valiant: A theory of the learnable. *Communications of the ACM*, **27**(11): 1134–1142, Nov. 1984.
33. Vladimir N. Vapnik: *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, New York, 1982.
34. Vladimir N. Vapnik: *Statistical Learning Theory*, Wiley, 1998.
35. 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.
36. M. Vidyasagar: *A Theory of Learning and Generalization*, Springer-Verlag, 1996.
37. Robert Williamson, John Shawe-Taylor, Bernhard Scholkopf, and Alex Smola: *Sample Based Generalization Bounds*, NeuroCOLT Technical Report, NC-TR-99-055, 1999.