On Boolean Combinations of Definitive Classifiers

Martin Anthony*

CDAM Research Report LSE-CDAM-2003-22
December 2003

Abstract

We consider the sample complexity of concept learning when we classify by using a fixed Boolean function of the outputs of a number of different classifiers. Here, we take into account the 'margins' of each of the constituent classifiers. A special case is that in which the constituent classifiers are linear threshold functions (or perceptrons) and the fixed Boolean function is the majority function. This corresponds to a 'committee of perceptrons', an artificial neural network (or circuit) consisting of a single layer of perceptrons (or linear threshold units) in which the output of the network is defined to be the majority output of the perceptrons. Recent work of Auer *et al.* studied the computational properties of such networks (where they were called 'parallel perceptrons'), proposed an incremental learning algorithm for them, and demonstrated empirically that the learning rule is effective. As a corollary of the sample complexity result presented here, generalization error bounds are derived for this special case that provide further motivation for the use of this learning rule.

*Department of Mathematics and Centre for Discrete and Applicable Mathematics, London School of Economics, London WC2A 2AE, UK. m.anthony@lse.ac.uk

1 Introduction

We consider the sample complexity of concept learning when we classify by using a fixed Boolean function of the outputs of a number of different classifiers. Here, we take into account the 'margins' of each of the constituent classifiers. A special case is that in which the constituent classifiers are linear threshold functions (or perceptrons) and the fixed Boolean function is the majority function. This corresponds to a 'committee of perceptrons', an artificial neural network (or circuit) consisting of a single layer of perceptrons (or linear threshold units) in which the output of the network is defined to be the majority output of the perceptrons. Recent work of Auer *et al.* [5, 6] studied the computational properties of such networks (where they are called 'parallel perceptrons'), proposed an incremental learning algorithm for them, and demonstrated empirically that the learning rule is effective. (They also studied a more general model, in which the outputs of the threshold functions are not passed through a majority function, but are instead summed to give a real-valued output, thus enabling such circuits to approximate real-valued functions.) As a corollary of the sample complexity result presented here, generalization error bounds are derived for this type of network that provide further motivation for the use of this learning rule.

2 Boolean combinations of classifiers

If F is a set of functions from $\mathbb{R}^n$ to $\mathbb{R}$, then F can be used to classify data points into two classes (labelled 0 and 1) by considering the sign of $f(x)$ for $x \in \mathbb{R}^n$. Explicitly, denote by $h = \mathrm{sgn}(f)$ the function $h : \mathbb{R}^n \to \{0,1\}$ given by $h(x) = \mathrm{sgn}(f(x))$ where $\mathrm{sgn}(z) = 1$ if $z \geq 0$ and $\mathrm{sgn}(z) = 0$ if $z < 0$. The set $H = \mathrm{sgn}(F) = \{\mathrm{sgn}(f) : f \in F\}$ is then a set of *classifiers*. Suppose now that $n, k \in \mathbb{N}$ and that $F_1, F_2, \ldots, F_k$ are sets of functions mapping $\mathbb{R}^n$ into $\mathbb{R}$. For each i, let $H_i = \mathrm{sgn}(F_i)$. Suppose that $g : \{0,1\}^k \to \{0,1\}$ is a fixed Boolean function and let $\mathcal{F}$ denote the k-tuple $(F_1, F_2, \ldots, F_k)$. Then we denote by $g(\mathcal{F}) = g(F_1, F_2, \ldots, F_k)$ the set of all functions from $\mathbb{R}^n$ to $\{0,1\}$ of the form

$$x \mapsto g\left(h_1(x), h_2(x), \ldots, h_k(x)\right)$$

where $h_1 \in H_1, h_2 \in H_2, \ldots, h_k \in H_k$. We call the functions in F_i the ith *constituent functions* and the corresponding functions in H_i the ith *constituent classifiers*. The functions (or classifiers) in $g(\mathcal{F})$ are thus a fixed Boolean function of the outputs of some constituent classifiers,

where the ith constituent classifier h_i is from H_i. Such classifiers have been considered often and can describe many natural methods of pattern classification: see [2, 12], for instance.

A special case of this construction is that in which, for each i, $F_i = \{\phi_i(a, .) : a \in \mathbb{R}^d\}$, where $\phi_i : \mathbb{R}^d \times \mathbb{R}^n \to \mathbb{R}$, and $a \in \mathbb{R}^d$ is a vector of parameters. When $\phi_1, \phi_2, \ldots, \phi_k$ belong to a particular class Φ of functions, the resulting classifier $g(\mathcal{F})$ has been called a *k-combination of* $\mathrm{sgn}(\Phi)$ [2]. Such classifiers have been studied extensively in [12] in the case where the ϕ_i are polynomial in the parameters a.

Suppose each f_i is an affine function, of the form $x \mapsto \langle w^i, x\rangle - \theta$ for some $w^i \in \mathbb{R}^n$ and $\theta^i \in \mathbb{R}$. (Here, $\langle a, b\rangle = a^T b$ is the standard inner product on $\mathbb{R}^n$.) The corresponding classifier $h_i = \mathrm{sgn}(f_i)$ is then a *linear threshold function*, given by

$$h_i(x) = \mathrm{sgn}\left(\left\langle w^i, x\right\rangle - \theta^i\right).$$

Suppose also that g is the *majority function*, whose output is 1 if and only if at least $\lceil k/2\rceil$ of its inputs are 1. In this case, $g(\mathcal{F})$ is a majority function of the outputs of linear threshold functions. This corresponds to a simple type of artificial neural network known as a 'committee machine', a (type of) 'madaline', or a 'parallel perceptron' [19, 5, 6].

3 Generalization error

Following a form of the PAC model of computational learning theory (see [4, 17, 9]), we assume that labelled data points (x, b) (where $x \in \mathbb{R}^n$ and $b \in \{0, 1\}$) have been generated randomly according to a fixed probability distribution P on $Z = \mathbb{R}^n \times \{0, 1\}$. Note that this includes as a special case the situation in which x is drawn according to a fixed distribution μ on $\mathbb{R}^n$ and the label b is then given by $b = t(x)$ where t is some fixed function. (More formally, we should say that P is defined on an appropriate σ-algebra of subsets of Z, usually taken to be the Borel algebra. Certain measure-theoretic conditions are required for the analysis that follows, but these are fairly undemanding and will not be addressed here; see [13], for example.) Thus, if there are m data points, we may regard the data set as a *sample* $s = ((x_1, b_1), \ldots, (x_m, b_m)) \in Z^m$, drawn randomly according to the product probability distribution P^m. Suppose that H is a set of functions from $\mathbb{R}^n$ to $\{0, 1\}$. Given any function $h \in H$, we might measure how closely

h matches the classifications given by the sample s through its *sample error*

$$\mathrm{er}_s(h) = \frac{1}{m}|\{i : h(x_i) \neq b_i\}|$$

(the proportion of points in the sample incorrectly classified by h). An appropriate measure of how well h would perform on further examples is its *(generalization) error*,

$$\mathrm{er}_P(h) = P\left(\{(x, b) \in Z : h(x) \neq b\}\right),$$

the probability that a further randomly drawn labelled data point would be incorrectly classified by h.

Much effort has gone into obtaining high-probability bounds on $\mathrm{er}_P(h)$ in terms of the sample error. A typical result would state that, for all $\delta \in (0, 1)$, with probability at least $1-\delta$, $\mathrm{er}_P(h) < \mathrm{er}_s(h) + \epsilon(m, \delta)$, for all $h \in \mathcal{H}$, where $\epsilon(m, \delta)$ is decreasing in m and increasing in δ. Such results can be derived using uniform convergence theorems from probability theory [18, 13, 11], in which case $\epsilon(m, \delta)$ would typically involve the VC-dimension; see [18, 9, 17, 2].

Recently, some emphasis has been placed in practical machine learning techniques, such as Support Vector Machines (see [10], for instance), on 'learning with a large margin'; see [16, 2, 3, 14], for instance. Broadly speaking, the rationale is that if a classifier has managed to achieve a 'wide' separation between (most of) the points of different classification, then this indicates that it is a good classifier, possibly with small generalization error. The classical example of this is linear separation, where the classifier is a linear threshold function. If we have found a linear threshold function that classifies the points of a sample correctly *and*, moreover, the points of opposite classifications are separated by a wide margin (so that the hyperplane achieves not just a correct, but a 'definitely' correct classification), then this function might be a better classifier of future, unseen, points than one which 'merely' separates the points correctly, but with a small margin.

Here, we obtain generalization error bounds for classifiers that are fixed Boolean functions of the outputs of constituent classifiers. The bounds we obtain depend on the classification margins *of the constituent functions*; that is, on how 'definitive' these individual classifications are. This is to be contrasted with generalization error bounds that consider the margin obtained by an aggregation of the constituent classifiers. For example, to be specific, suppose that we have k constituent classifiers that are linear threshold functions, and the combining Boolean function g is the majority function (so that we have the 'parallel perceptron' of [5, 6]). We derive

here bounds on the generalization error that depend on the margins of each of the k individual linear threshold functions. These bounds do *not* involve the margin achieved by the overall majority function g, as measured by the 'size' of the majority (to use an electoral analogy). The problem considered here is therefore similar in nature to that considered by Shawe-Taylor and Cristianini [15] and Bennett *et al.* [8], where generalization bounds for *perceptron decision trees* were obtained, these bounds depending on the margins obtained at each decision node.

The following definition describes the margins that we consider, and the corresponding definition of error.

Definition 3.1 *Suppose, with the above notation, that g is a fixed Boolean function, that F_i maps $\mathbb{R}^n$ to $\mathbb{R}$ (for $i = 1, 2, \ldots, k$), and that $h \in g(\mathcal{F})$. Suppose that*

$$h(x) = g(\mathrm{sgn}(f_1(x)), \mathrm{sgn}(f_2(x)), \ldots, \mathrm{sgn}(f_k(x))),$$

where $f_i \in F_i$. For $(x, b) \in \mathbb{R}^n \times \{0, 1\}$, we say that h classifies the labelled example (x, b) (correctly and) with margin $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k) \in \mathbb{R}^k$, *where all γ_i are positive, if $h(x) = b$ and, for $i = 1, 2, \ldots, k$, $|f_i(x)| \geq \gamma_i$. For a sample $s \in (\mathbb{R}^n \times \{0, 1\})^m$, we define the* empirical (or observed) error at margin Γ, $\mathrm{er}_s^\Gamma(h)$, *to be the proportion of labelled examples in the sample s that are not classified with margin Γ.*

Informally, then, if (x, b) is classified with margin Γ, and if the numbers γ_i are quite large, then the constituent classifications are *definitive* in the sense that the sign of $f_i(x)$ is *either* at least γ_i or is at most $-\gamma_i$, and is not merely positive or negative. (For this interpretation to be valid, we need to make assumptions about the range of the function classes F_i, so that the margins being 'quite large' can sensibly be defined. Often this is accomplished, as we shall see, by restricting the domain of the inputs x that are considered.)

To use an electoral analogy—which at least has some merit if the function g is the majority function—large margin classification in the sense meant here means that the 'voters' (the constituent classifiers) each have a strong opinion about the 'candidate' x. Again, to emphasise a key difference between this and other analyses based on 'margins', we are saying nothing about how definitive the 'aggregation' g is: the 'definitive' classifications are made at the level of the individual constituent classifiers. Thus, in the voting analogy, we assume the voters have strong views, but we assume nothing about the size of the majority that determines the outcome of the election (beyond the fact that it is a majority).

4 Generalization error bounds

A key tool in the derivation of margin-based generalization error bounds is the *covering number* of a class of real functions. Suppose that $F : X \to \mathbb{R}$ is a set of real-valued functions with domain X, which we shall usually take to be a bounded subset of $\mathbb{R}^n$. Suppose $x = (x_1, x_2, \ldots, x_m) \in X^m$. Then, for $\epsilon > 0$, $C \subseteq F$ is an *ϵ-cover of F with respect to the d_∞^x-metric* if for all $f \in F$ there is $\hat{f} \in C$ such that $d_\infty^x(f, \hat{f}) < \epsilon$, where

$$d_\infty^x(f, g) = \max_{1 \le i \le m} |f(x_i) - g(x_i)|.$$

The class F is said to be totally bounded if it has a finite ϵ-cover with respect to the d_∞^x metric for all $\epsilon > 0$ and all $x \in X^m$ (for all m). In this case, given $x \in X^m$, we define the *d_∞^x-covering numbers* $\mathcal{N}_\infty(F, \epsilon, x)$ to be the minimum cardinality of an ϵ-cover of F with respect to the d_∞^x-metric. We then define the *d_∞-covering numbers* $\mathcal{N}_\infty(F, \epsilon, m)$ by

$$\mathcal{N}_\infty(F, \epsilon, m) = \sup\{\mathcal{N}_\infty(F, \epsilon, x) : x \in X^m\}.$$

The following result bounds the generalization error in terms of the empirical margin error and the covering numbers of the constituent function classes.

Theorem 4.1 *Suppose that g is a fixed Boolean function, and that F_i maps $X \subseteq \mathbb{R}^n$ to $\mathbb{R}$ (for $i = 1, 2, \ldots, k$). Let $H = g(\mathcal{F})$ (as defined above) and suppose that each F_i is totally bounded. Let $\gamma_1, \gamma_2, \ldots, \gamma_k \in (0, 1]$ be given and let Z denote $X \times \{0, 1\}$. Then, for any probability measure P on Z, with P^m-probability at least $1 - \delta$, the following hold for $s \in Z^m$:*

- *if $h \in H$ and $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$, then*

$$\mathrm{er}_P(h) < \mathrm{er}_s^\Gamma(h) + \sqrt{\frac{8}{m}\left(\sum_{i=1}^{k} \ln \mathcal{N}_\infty(F_i, \gamma_i/2, 2m) + \ln\left(\frac{2}{\delta}\right)\right)};$$

- *if $h \in H$ classifies s with margin $\Gamma = (\gamma_1, \ldots, \gamma_k)$ (so that $\mathrm{er}_s^\Gamma(h) = 0$), then*

$$\mathrm{er}_P(h) < \frac{2}{m}\left(\sum_{i=1}^{k} \log_2 \mathcal{N}_\infty(F_i, \gamma_i/2, 2m) + \log_2\left(\frac{2}{\delta}\right)\right).$$

One difficulty with Theorem 4.1 is that the margins γ_i are specified *a priori*. A more useful result is the following, which would apply to situations in which we might choose, or observe, these parameters after learning.

Theorem 4.2 *Suppose that g is a fixed Boolean function, and that F_i maps $X \subseteq \mathbb{R}^n$ to $\mathbb{R}$ (for $i = 1, 2, \ldots, k$). Let $H = g(\mathcal{F})$ (as defined above) and suppose that each F_i is totally bounded. Let Z denote $X \times \{0, 1\}$. Then, for any probability measure P on Z^m, with P^m-probability at least $1 - \delta$, the following hold for $s \in Z^m$:*

- *if $h \in H$, then for* any $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$,

$$\mathrm{er}_P(h) < \mathrm{er}_s^{\Gamma}(h) + \sqrt{\frac{8}{m}\left(\sum_{i=1}^{k} \ln \mathcal{N}_\infty(F_i, \gamma_i/4, 2m) + \ln\left(\frac{2}{\delta}\right) + k \ln 2 + \sum_{i=1}^{k} \ln\left(\frac{1}{\gamma_i}\right)\right)};$$

- *for* any $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$, *if $h \in H$ classifies s with margin Γ (so that $\mathrm{er}_s^{\Gamma}(h) = 0$), then*

$$\mathrm{er}_P(h) < \frac{2}{m}\left(\sum_{i=1}^{k} \log_2 \mathcal{N}_\infty(F_i, \gamma_i/4, 2m) + \log_2\left(\frac{2}{\delta}\right) + k + \sum_{i=1}^{k} \log_2\left(\frac{1}{\gamma_i}\right)\right).$$

5 A special case: parallel perceptrons and the p-delta learning rule

We now apply the above results to the special case in which g is the majority function and the constituent classifiers are linear threshold functions (or perceptrons). In this case, the results tell us something about the generalization performance of a learning algorithm from [5, 6] known as the *p-delta learning rule* (for Boolean-valued parallel perceptrons). This algorithm incrementally updates the weight-vectors of the constituent perceptrons in such a way as to try to maintain, for each, a margin γ. In their papers [5, 6], Auer *et al.* write that: "Since our parallel perceptron is an aggregation of many simple perceptrons with large margins [...], one expects that parallel perceptrons [trained with the p-delta algorithm] also exhibit good generalization." They provide empirical evidence of good generalization on standard data-sets. Theorem 4.1 and

Theorem 4.2 help in giving some further justification for this learning paradigm, by providing generalization error bounds that depend on the margins achieved by the constituent perceptrons. Specifically, Theorem 5.1 (below) indicates that one might well expect better generalization when the margins of the constituent linear threshold functions are large; and this is the rationale behind the p-delta learning procedure. (The generalization error bounds given are just upper bounds, and they also involve the empirical margin error $\mathrm{er}_s^\Gamma(h)$, which increases as the margins are increased. Therefore it does not follow that the smallest error is necessarily achieved when the margins are large. Nonetheless, the bounds suggest that maximizing these margins, subject to maintaining a small empirical margin error, is a sensible strategy.)

We shall assume, for simplicity—and because there is no loss of generality in doing so—that the thresholds θ_i are fixed at 0, so that the constituent threshold functions are *homogeneous* threshold functions. (Any threshold function with a non-zero threshold can be realized as a restriction of a homogeneous threshold function in one more variable.) We shall also assume that the domain of interest (the support of the marginal distributions on $\mathbb{R}^n$ that we consider) is the bounded set $B_R = \{x \in \mathbb{R}^n : \|x\| \leq R\}$ for some $R \geq 1$. The threshold functions then take the form $h(x) = \mathrm{sgn}(\langle w, x\rangle)$ and, since scaling w by a positive constant does not change the functionality, we may assume that $\|w\| = 1$. So, we can assume that, for each i, $H_i = \mathrm{sgn}(F)$ where F is the set $\{x \mapsto \langle w, x\rangle : \|w\| = 1\}$, regarded as a set of functions $B_R \to [-R, R]$.

The first part of the following theorem bounds the error when the margins are prescribed in advance, and the second part bounds the error when the margins are not set *a priori*.

Theorem 5.1 *Suppose that g is the majority function of k variables, and that H is the set of all functions $B_R \to \{0,1\}$ of the form*

$$h(x) = g(h_1(x), h_2(x), \ldots, h_k(x))$$

where each h_i is a homogeneous linear threshold function. Let Z denote $B_R \times \{0,1\}$.

1. *Let $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k) \in (0,1]^k$ be given. Then, for any probability measure P on Z, with probability at least $1-\delta$, the following hold for $s \in Z^m$:*

   - *for all $h \in H$,*

$$\mathrm{er}_P(h) < \mathrm{er}_s^\Gamma(h) + \sqrt{\frac{8}{m}\left(216\,\ln(13m)\sum_{i=1}^{k}\frac{R^2}{\gamma_i^2} + \ln\left(\frac{2}{\delta}\right)\right)};$$

- *if $h \in H$ classifies s with margin Γ (so that $\mathrm{er}_s^\Gamma(h) = 0$) then*

$$\mathrm{er}_P(h) < \frac{2}{m}\left(216\,\log_2(13m)\sum_{i=1}^{k}\frac{R^2}{\gamma_i^2} + \log_2\left(\frac{2}{\delta}\right)\right).$$

2. *For any probability measure P on Z^m, with probability at least $1-\delta$, the following hold for $s \in Z^m$:* for all $\gamma_1, \gamma_2, \ldots, \gamma_k \in (0,1]$,

   - *if $h \in H$ and $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k)$, then*

$$\mathrm{er}_P(h) < \mathrm{er}_s^\Gamma(h) + \sqrt{\frac{8}{m}\left(864\,\ln(18m)\sum_{i=1}^{k}\frac{R^2}{\gamma_i^2} + \ln\left(\frac{2}{\delta}\right) + k\ln 2 + \sum_{i=1}^{k}\ln\left(\frac{1}{\gamma_i}\right)\right)};$$

   - *if $h \in H$ classifies s with margin $\Gamma = (\gamma_1, \ldots, \gamma_k)$ (so that $\mathrm{er}_s^\Gamma(h) = 0$) then*

$$\mathrm{er}_P(h) < \frac{2}{m}\left(864\,\log_2(18m)\sum_{i=1}^{k}\frac{R^2}{\gamma_i^2} + \log_2\left(\frac{2}{\delta}\right) + k + \sum_{i=1}^{k}\log_2\left(\frac{1}{\gamma_i}\right)\right).$$

We see from these bounds that a key term controlling the amount by which the true error $\mathrm{er}_P(h)$ can differ from the observed margin error $\mathrm{er}_s^\Gamma(h)$ is the quantity $\sum_{i=1}^{k}(1/\gamma_i^2)$.

The bounds take simpler forms, of course, when all the γ_i are equal. For example, we see that, for any distribution P, with probability at least $1-\delta$, for all $\Gamma \in (0,1]^k$ where $\gamma_i = \gamma$ for all i, if $h \in H$ classifies s with margin Γ, then

$$\mathrm{er}_P(h) < \frac{2}{m}\left(864\,k\log_2(18m)\frac{R^2}{\gamma^2} + \log_2\left(\frac{2}{\delta}\right) + k\left(1 + \log_2\left(\frac{1}{\gamma}\right)\right)\right).$$

In particular, therefore, if it is observed that each constituent perceptron has achieved a margin $\gamma > 0$ after training on a sample, this gives a bound on the generalization error. (This is one way in which the results where the margins are not pre-specified can be useful: this result probabilistically bounds the generalization error of any classifier in H having zero sample error, in terms of the (observed) margins that are achieved by its constituent classifiers.)

6 Proofs

Proof of Theorem 4.1

The proof follows [1], which extends a technique from [15, 8] (where the case $\mathrm{er}_s^\Gamma(h) = 0$ is considered). It is a modification of proofs in [2, 3, 7, 14], which in turn are based on [18].

Given $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_n)$ and $(s, s') \in Z^m \times Z^m$, we note that if $\mathrm{er}_P(h) \geq \mathrm{er}_s^\Gamma(h) + \epsilon$ and $\mathrm{er}_{s'}(h) \geq \mathrm{er}_P(h) - \epsilon/2$, then $\mathrm{er}_{s'}(h) \geq \mathrm{er}_s^\Gamma(h) - \epsilon/2$. It follows that if

$$Q = \{s \in Z^m : \exists h \in H \text{ with } \mathrm{er}_P(h) \geq \mathrm{er}_s^\Gamma(h) + \epsilon\}$$

and

$$R = \{(s, s') \in Z^m \times Z^m : \exists h \in H \text{ with } \mathrm{er}_{s'}(h) \geq \mathrm{er}_s^\Gamma(h) + \epsilon/2\},$$

then

$$\begin{aligned} P^{2m}(R) &\geq P^{2m}\left(\exists h \in H : \mathrm{er}_P(h) \geq \mathrm{er}_s^\Gamma(h) + \epsilon \text{ and } \mathrm{er}_{s'}(h) \geq \mathrm{er}_P(h) - \epsilon/2\right) \\ &= \int_Q P^m\left(\left\{s' : \exists h \in H, \mathrm{er}_P(h) \geq \mathrm{er}_s^\Gamma(h) + \epsilon \text{ and } \mathrm{er}_{s'}(h) \geq \mathrm{er}_P(h) - \epsilon/2\right\}\right) dP^m(s) \\ &\geq \frac{1}{2} P^m(Q), \end{aligned}$$

for $m \geq 2/\epsilon^2$, where the final inequality follows from $P^m(\mathrm{er}_{s'}(h) \geq \mathrm{er}_P(h) - \epsilon/2) \geq 1/2$, for any $h \in H$, by Chebychev's inequality. Hence, $P^m(Q) \leq 2\, P^{2m}(R)$. Let G be the permutation group (the 'swapping group') on the set $\{1, 2, \ldots, 2m\}$ generated by the transpositions $(i, m+i)$ for $i = 1, 2, \ldots, m$. Then G acts on Z^{2m} by permuting the coordinates: for $\sigma \in G$, $\sigma(z_1, z_2, \ldots, z_{2m}) = (z_{\sigma(1)}, \ldots, z_{\sigma(m)})$. Now, by invariance of P^{2m} under the action of G, $P^{2m}(R) = \mathbb{E}\Pr(\sigma \in R) \leq \max\{\Pr(\sigma z \in R) : z \in Z^{2m}\}$, where $\Pr$ denotes the probability over uniform choice of σ from G, and the expectation is with respect to P^m. (See [18, 2], for instance.) Now, *fix* $z \in Z^{2m}$, where $z_i = (x_i, b_i)$. Let $x = (x_1, x_2, \ldots, x_{2m})$. Suppose that C_i is a $\gamma_i/2$-cover of F_i with respect to d_∞^x, of minimum cardinality, which will be no more than $\mathcal{N}_\infty(F_i, \gamma_i/2, 2m)$. Suppose that $\sigma z = (s, s') \in R$. This means that for some $h \in H = g(F)$, $\mathrm{er}_{s'}(h) \geq \mathrm{er}_s^\Gamma(h) + \epsilon/2$. Now, suppose that

$$h(x) = g\left(\mathrm{sgn}(f_1(x)), \mathrm{sgn}(f_2(x)), \ldots, \mathrm{sgn}(f_k(x))\right),$$

where $f_i \in F_i$, and let $\hat{f}_i \in C_i$ be such that $d^x_\infty(\hat{f}_i, f_i) < \gamma_i/2$. Then, for all i, j, $|f_i(x_j) - \hat{f}(x_j)| < \gamma_i/2$. Let $\hat{h}$ be the function in $g(F)$ defined by

$$\hat{h}(x) = g\left(\mathrm{sgn}(\hat{f}_1(x)), \mathrm{sgn}(\hat{f}_2(x)), \ldots, \mathrm{sgn}(\hat{f}_k(x))\right)$$

and let $\hat{H}$ denote the set of all such $\hat{h}$. Now, $\sigma z = (s, s') \in R$, so $\mathrm{er}_{s'}(h) \geq \mathrm{er}^\Gamma_s(h) + \epsilon/2$. This implies that $\mathrm{er}^{\Gamma/2}_{s'}(h) \geq \mathrm{er}^{\Gamma/2}_s(h) + \epsilon/2$, where $\Gamma/2 = (\gamma_1/2, \ldots, \gamma_k/2)$. This claim follows from two observations: (i) $\mathrm{er}^{\Gamma/2}_{s'}(\hat{h}) \geq \mathrm{er}_{s'}(h)$, and (ii) $\mathrm{er}^\Gamma_s(h) \geq \mathrm{er}^{\Gamma/2}_s(\hat{h})$. To see (i), suppose that h does not classify (x, b) correctly. For each i, either $\mathrm{sgn}(\hat{f}_i(x)) = \mathrm{sgn}(f_i(x))$; or, $\mathrm{sgn}(\hat{f}_i(x)) \neq \mathrm{sgn}(\hat{f}_i(x))$ *but* $|\hat{f}_i(x)| < \gamma_i/2$. This is because replacing the ith constituent function f_i by $\hat{f}_i$ changes the output of that constituent function on x by at most $\gamma_i/2$; this may be enough of a change to change the sign of this output, but it is not enough to do so in such a way as to achieve a margin of $\gamma_i/2$. So, it follows that either $\hat{h}$ classifies x in the same way as h does (that is, wrongly), or it classifies x correctly, but not with margin Γ. To see (ii), suppose that $\hat{h}$ does not classify (x, b) with margin $\Gamma/2$. Then, either $\hat{h}$ does not classify (x, b) correctly, or it does classify the example correctly, but at least one of its constituent functions $\hat{f}_i$ is such that $|\hat{f}_i(x)| < \gamma_i/2$. It is possible that h will classify (x, b) correctly even though $\hat{h}$ does not: this would mean that for at least one i, $\mathrm{sgn}(f_i(x)) \neq \mathrm{sgn}(\hat{f}_i(x))$. However, in this case, since $|f_i(x) - \hat{f}_i(x)| < \gamma_i/2$, we would have $|f_i(x)| < \gamma/2 < \gamma$. In the second case, since $|\hat{f}_i(x)| < \gamma_i/2$, we have

$$|f_i(x)| < \gamma_i/2 + |f_i(x) - \hat{f}_i(x)| < \gamma_i$$

and so h does not classify the example with margin Γ. It now follows that $\sigma z \in R$ if and only if $\sigma z \in S$, where

$$S = \{(s, s') \in Z^{2m} : \exists \hat{h} \in \hat{H} \text{ with } \mathrm{er}^{\Gamma/2}_{s'}(\hat{h}) \geq \mathrm{er}^{\Gamma/2}_s(\hat{h}) + \epsilon/2\} = \bigcup_{\hat{h} \in \hat{H}} S(\hat{h}),$$

and where

$$S(\hat{h}) = \{(s, s') \in Z^{2m} : \mathrm{er}^{\Gamma/2}_{s'}(\hat{h}) \geq \mathrm{er}^{\Gamma/2}_s(\hat{h}) + \epsilon/2\}.$$

Hence,

$$\Pr(\sigma z \in R) \leq \Pr(\sigma z \in S) \leq \sum_{\hat{h} \in \hat{H}} \Pr\left(\sigma z \in S(\hat{h})\right).$$

Now, fix $\hat{h} \in \hat{H}$ and let $\psi_i = 0$ if $\hat{h}$ classifies z_i with margin at least $\Gamma/2$, and 1 otherwise. Then

$$\Pr\left(\sigma z \in S(\hat{h})\right) = \Pr\left(\frac{1}{m}\sum_{i=1}^{m}(\psi_{m+i} - \psi_i) \geq \epsilon/2\right) = \Pr\left(\frac{1}{m}\sum_{i=1}^{m} r_i|\psi_i - \psi_{m+i}| \geq \epsilon/2\right),$$

where the r_i are independent (Rademacher) $\{-1, 1\}$ random variables, each taking value 1 with probability $1/2$, and where the last probability is over the joint distribution of the r_i. Hoeffding's inequality bounds this probability by $e^{-\epsilon^2 m/8}$. (See [2], for instance, for details.) We therefore have (for $m \geq 2/\epsilon^2$)

$$\Pr(\sigma z \in R) \leq |\hat{H}| e^{-\epsilon^2 m/8},$$

which gives

$$P^m(Q) \leq 2\, P^{2m}(R) \leq 2 \prod_{i=1}^{k} |C_i| \exp(-\epsilon^2 m/8) \leq 2 \prod_{i=1}^{k} \mathcal{N}_\infty(F_i, \gamma_i/2, 2m) e^{-\epsilon^2 m/8}.$$

This quantity is at most δ if

$$\epsilon \geq \sqrt{\frac{8}{m}\left(\sum_{i=1}^{k} \ln \mathcal{N}_\infty(F_i, \gamma_i/2, 2m) + \ln\left(\frac{2}{\delta}\right)\right)}$$

(which also implies $m \geq 2/\epsilon^2$). The first statement of the Theorem follows.

The second part of the Theorem is proved similarly. It uses, first, the fact (see [18, 2]) that if

$$Q = \{s \in Z^m : \exists h \in H \text{ with } \mathrm{er}_s^\Gamma(h) = 0,\ \mathrm{er}_P(h) \geq \epsilon\}$$

and

$$R = \{(s, s') \in Z^m \times Z^m : \exists h \in H \text{ with } \mathrm{er}_s^\Gamma(h) = 0,\ \mathrm{er}_{s'}(h) \geq \epsilon/2\},$$

then $P^m(Q) \leq 2\, P^{2m}(R)$. As before, $P^{2m}(R) \leq \max_{z \in Z^{2m}} \Pr(\sigma z \in R)$, where $\Pr$ denotes the probability over uniform choice of σ from the 'swapping group' G. It can be shown that for any $z \in Z^{2m}$,

$$\Pr\left(\sigma z \in R\right) \leq \Pr\left(\sigma z \in \bigcup_{\hat{h} \in \hat{H}} S(\hat{h})\right),$$

where

$$S(\hat{h}) = \{(s, s') \in Z^{2m} : \exists h \in H \text{ with } \mathrm{er}_s^\Gamma(h) = 0,\ \mathrm{er}_{s'}(h) \geq \epsilon/2\}.$$

It can then be seen (by an easy counting argument) that, for each fixed $\hat{h} \in \hat{H}$,

$$\Pr\left(\sigma z \in S(\hat{h})\right) \leq \frac{2^{m(1-\epsilon/2)}}{|G|} = 2^{-\epsilon m/2}.$$

The argument continues as above.

Proof of Theorem 4.2

We use a result from [1], which is a k-dimensional version of a 'sieve' result from [7]. This states that if $\mathbb{P}$ is any probability measure, $k \in \mathbb{N}$, and

$$\{E(\Gamma_1, \Gamma_2, \delta) : \Gamma_1, \Gamma_2 \in (0,1]^k, \delta \leq 1\}$$

is a set of events such that:

- for all $\Gamma \in (0,1]^k$, $\mathbb{P}(E(\Gamma, \Gamma, \delta)) \leq \delta$,
- $\Gamma_1 \leq \Gamma \leq \Gamma_2$ (component-wise) and $0 < \delta_1 \leq \delta \leq 1$ imply $E(\Gamma_1, \Gamma_2, \delta_1) \subseteq E(\Gamma, \Gamma, \delta)$,

then

$$\mathbb{P}\left(\bigcup_{\Gamma \in (0,1]^k} E\left((1/2)\Gamma, \Gamma, \delta\, 2^{-k} \prod_{i=1}^{k} \gamma_i \right) \right) \leq \delta$$

for $0 < \delta < 1$. If $\Gamma_1 = (\gamma_1^{(1)}, \ldots, \gamma_k^{(1)})$ and $\Gamma_2 = (\gamma_1^{(2)}, \ldots, \gamma_k^{(2)})$, let

$$E(\Gamma_1, \Gamma_2, \delta) = \left\{ (s, s') : \exists h \in H \text{ with } \mathrm{er}_P(h) \geq \mathrm{er}_s^{\Gamma_2}(h) + \epsilon(\Gamma_1, m, \delta) \right\},$$

where

$$\epsilon(\Gamma_1, m, \delta) = \sqrt{\frac{8}{m}\left(\sum_{i=1}^{k} \ln \mathcal{N}_\infty(F_i, \gamma_i^{(1)}/2, 2m) + \ln\left(\frac{2}{\delta}\right) \right)}.$$

Then Theorem 4.1 states that $P^m(E(\Gamma, \Gamma, \delta)) \leq \delta$ for any probability measure P on Z. It is also easily seen that if $\Gamma_1 \leq \Gamma \leq \Gamma_2$ and $0 < \delta_1 \leq \delta \leq 1$, then $E(\Gamma_1, \Gamma_2, \delta_1) \subseteq E(\Gamma, \Gamma, \delta)$: this is because $\mathrm{er}_s^{\Gamma_2}(h) \geq \mathrm{er}_s^{\Gamma}(h)$ and $\epsilon(\Gamma_1, m, \delta_1) \geq \epsilon(\Gamma, m, \delta)$. It follows that

$$P^m\left(\bigcup_{\Gamma \in (0,1]^k} E\left((1/2)\Gamma, \Gamma, \delta\, 2^{-k} \prod_{i=1}^{k} \gamma_i \right) \right) \leq \delta.$$

That is, with P^m-probability at least $1-\delta$, for any $h \in H$ and any $\Gamma = (\gamma_1, \gamma_2, \ldots, \gamma_k) \in (0,1]^k$,

$$\mathrm{er}_P(h) < \mathrm{er}_s^{\Gamma}(h) + \sqrt{\frac{8}{m}\left(\sum_{i=1}^{k} \ln \mathcal{N}_\infty(F_i, \gamma_i/4, 2m) + \ln\left(\frac{2\, 2^k}{\delta \prod_{i=1}^{k} \gamma_i} \right) \right)},$$

from which the first part of Theorem 4.2 follows. The second part is obtained similarly.

Proof of Theorem 5.1

We use a recent bound of Zhang [20] for the d_∞-covering numbers of sets bounded linear mappings. This shows that if F is the set of functions $\{x \mapsto \langle w, x\rangle : \|w\| = 1\}$, regarded as mapping from B_R to $[-R, R]$, then

$$\log_2 \mathcal{N}_\infty(F, \epsilon, m) \leq 36 \frac{R^2}{\epsilon^2} \log_2 \left(2 \left\lceil 4R/\epsilon + 2 \right\rceil m + 1\right).$$

We prove the first of the four stated bounds: the others are very similarly derived from Theorem 4.1 and Theorem 4.2. It follows from the bound of Zhang that

$$\ln \mathcal{N}_\infty(F_i, \gamma_i/2, 2m) \leq \frac{144R^2}{\gamma_i^2} \ln\left(\frac{42Rm}{\gamma_i}\right),$$

so by the first part of Theorem 4.1, for a given Γ, with probability at least $1-\delta$, for all $h \in H$, we have

$$\mathrm{er}_P(h) < \mathrm{er}_s^\Gamma(h) + \sqrt{\frac{8}{m}\left(144R^2 \sum_{i=1}^{k} \frac{1}{\gamma_i^2} \ln\left(\frac{42Rm}{\gamma_i}\right) + \ln\left(\frac{2}{\delta}\right)\right)}.$$

The bound we require is

$$\mathrm{er}_P(h) < \mathrm{er}_s^\Gamma(h) + \sqrt{\frac{8}{m}\left(216 \ln(13m) \sum_{i=1}^{k} \frac{R^2}{\gamma_i^2} + \ln\left(\frac{2}{\delta}\right)\right)}.$$

This bound is trivially true if, for some i, $m \leq R^2/\gamma_i^2$ (since the term under the square root is larger than 1 in this case). If $m > R^2/\gamma_i^2$ for all i, then

$$\ln\left(\frac{42Rm}{\gamma_i}\right) < \frac{3}{2}\ln(13m),$$

and so the required bound follows from the one obtained.

References

[1] M. Anthony. *Margin-based Generalization Error Bounds for Threshold Decision Lists.* CDAM Research Report LSE-CDAM-2003-09, Centre for Discrete and Applicable Mathematics, London School of Economics, 2003.

[2] M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge UK, 1999.

[3] M. Anthony and P. L. Bartlett. Function learning from interpolation. *Combinatorics, Probability and Computing*, 9, 2000: 213–225.

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

[5] P. Auer, H. M. Burgsteiner and W. Maass. The p-Delta learning rule for parallel perceptrons. preprint.

[6] P. Auer, H. Burgsteiner, and W. Maass. Reducing communication for distributed learning in neural networks. In Jos R. Dorronsoro (ed.), *Proceedings of the International Conference on Artificial Neural Networks, ICANN 2002*, volume 2415, Springer Lecture Notes in Computer Science, pp. 123-128. Springer, 2002.

[7] P. L. 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), 1998: 525–536.

[8] K. Bennett, N. Cristianini, J. Shawe-Taylor and D. Wu. Enlarging the Margins in Perceptron Decision Trees. *Machine Learning*, 41, 2000: 295–313.

[9] A. Blumer, A. Ehrenfeucht, D. Haussler, and M. K. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the ACM*, 36(4), 1989: 929–965.

[10] N. Cristianini and J. Shawe-Taylor. *An Introduction to Support Vector Machines*, Cambridge University Press, Cambridge, UK, 2000.

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

[12] P.W. Goldberg and M. R. Jerrum. Bounding the Vapnik-Chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning* 18(2/3), 1995: 131–148.

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

[14] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5), 1996: 1926–1940.

[15] J. Shawe-Taylor and N. Cristianini. Data-Dependent Structural Risk Minimisation for Perceptron Decision Trees. Neurocolt Technical Report NC2-TR-1998-003, May 1998.

[16] A. J. Smola, P. L. Bartlett, B. Schölkopf and D. Schuurmans (editors). *Advances in Large-Margin Classifiers (Neural Information Processing)*, MIT Press 2000.

[17] V. N. Vapnik: *Statistical Learning Theory*, Wiley, 1998.

[18] 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), 1971: 264–280.

[19] B. Widrow and M. A. Lehr. 30 years of adaptive neural networks: Perceptron, madaline, and backpropagation. *Proceedings of the IEEE* 78 (9): pp. 1415–1442.

[20] T. Zhang. Covering number bounds of certain regularized linear function classes. *Journal of Machine Learning Research*, 2, 2002: 527–550.