The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network

Peter L. Bartlett*

Member, IEEE

April 29, 1996

Revised: May 24, 1997

Abstract

Sample complexity results from computational learning theory, when applied to neural network learning for pattern classification problems, suggest that for good generalization performance the number of training examples should grow at least linearly with the number of adjustable parameters in the network. Results in this paper show that if a large neural network is used for a pattern classification problem and the learning algorithm finds a network with small weights that has small squared error on the training patterns, then the generalization performance depends on the size of the weights rather than the number of weights. For example, consider a two-layer feed-forward network of sigmoid units, in which the sum of the magnitudes of the weights associated with each unit is bounded by A and the input dimension is n. We show that the misclassification probability is no more than a certain error estimate (that is related to squared error on the training set) plus $A^3\sqrt{(\log n)/m}$ (ignoring $\log A$ and $\log m$ factors), where m is the number of training patterns. This may explain the generalization performance of neural networks, particularly when the number of training

*Department of Systems Engineering, Research School of Information Sciences and Engineering, Australian National University, Canberra, 0200 Australia. Peter.Bartlett@anu.edu.au

examples is considerably smaller than the number of weights. It also supports heuristics (such as weight decay and early stopping) that attempt to keep the weights small during training. The proof techniques appear to be useful for the analysis of other pattern classifiers: when the input domain is a totally bounded metric space, we use the same approach to give upper bounds on misclassification probability for classifiers with decision boundaries that are far from the training examples.

Index Terms — neural networks, scale-sensitive dimensions, weight decay, computational learning theory, pattern recognition.

1 Introduction

Neural networks are commonly used as learning systems to solve pattern classification problems. For these problems, it is important to establish how many training examples ensure that the performance of a network on the training data provides an accurate indication of its performance on subsequent data. Results from statistical learning theory (for example, [40, 10, 8, 19]) give sample size bounds that are linear in the Vapnik-Chervonenkis dimension of the class of functions used by the learning system. (The VC-dimension is a combinatorial complexity measure that is typically at least as large as the number of adjustable network parameters.) These results do not provide a satisfactory explanation of the sample size requirements of neural networks for pattern classification applications, for several reasons. First, neural networks often perform successfully with training sets that are considerably smaller than the number of network parameters (see, for example, [29]). Second, the VC-dimension of the class of functions computed by a network is sensitive to small perturbations of the computation unit transfer functions (to the extent that an arbitrarily small change can make the VC-dimension infinite, see [39]). That this could affect the generalization performance seems unnatural, and has not been observed in practice.

In fact, the sample size bounds in terms of VC-dimension are tight in the sense that, for every learning algorithm that selects hypotheses from some class, there is a probability distribution and a target function for which, if training data is chosen independently from the distribution and labelled according to the target function, the function chosen by the learning algorithm will misclassify a random example with probability at least proportional to the VC-dimension of the class divided by the number of training examples. However, for many neural networks, results in this paper show that these probability distributions and target functions are such that learning algorithms like back-propagation that are used in applications are unlikely to find a network that accurately classifies the training data. That is, these algorithms avoid choosing a network that overfits the data in these cases because they are not powerful enough to find *any* good solution.

The VC-theory deals with classes of $\{-1, 1\}$-valued functions. The algorithms it studies need only find a hypothesis from the class that minimizes the number of mistakes on the training examples. In contrast, neural networks have real-valued outputs. When they are used for classification problems, the sign of the network output is interpreted as the classification of an input example. Instead of minimizing the number of misclassifications of the training examples directly, learning algorithms typically attempt to minimize a smooth cost function, the total squared error of the (real-valued) network output over the training set. As well as encouraging the correct sign of the real-valued network output in response to a training example, this tends to push the output away from zero by some margin. Rather than maximizing the proportion of the training examples that are correctly classified, it approximately maximizes the proportion of the training examples that are "distinctly correct" in this way.

When a learning algorithm maximizes the proportion of distinctly correct training examples, the misclassification probability depends not on the VC-dimension of the function class, but on a scale-sensitive version of this dimension known as the fat-shattering dimension. The first main result of this paper shows that if an algorithm finds a function that performs well on the training data (in the sense that most examples are correctly classified with some margin), then with high confidence the misclassification probability is bounded in terms of the fat-shattering dimension and the number of examples. The second main result gives upper bounds on the fat-shattering dimension for neural networks in terms of the network depth and the magnitudes of the network parameters (and independent of the number of parameters). Together, these results imply the following sample complexity bounds for two-layer sigmoid networks. (Computation units in a sigmoid network calculate an affine combination of their inputs, composed with a fixed, bounded, Lipschitz function.) A more precise statement of these results appears in Theorem 28.

Consider a two-layer layer sigmoid network with an arbitrary number of hidden units, in which the sum of the magnitudes of the weights in the output unit is bounded by A and the input space is $\mathbb{R}^d$. If the training examples are generated independently according to some probability distribution, and the number of training examples increases roughly as $A^2 d/\epsilon^2$ (ignoring log factors), then with high probability every network function that classifies a fraction at least $1-\alpha$ of the training set correctly and with a fixed margin has misclassification probability no more than $\alpha+\epsilon$.

Consider a two-layer sigmoid network as above, for which each hidden unit also has the sum of the magnitudes of its weights bounded by A, and the network input patterns lie in $[-B, B]^d$. Then a similar result applies, provided the number of training examples increases roughly as $A^6 B^2 \log d/\epsilon^2$ (again ignoring log factors).

These results show that, for problems encountered in practice for which neural networks are well-suited (that is, for which gradient descent algorithms are likely to find good parameter values), the magnitude of the parameters may be more important than the number of parameters. Indeed, the number of parameters, and hence the VC-dimension, of both function classes described above is unbounded.

The result gives theoretical support for the use of "weight decay" and "early stopping" (see, for example, [21]), two heuristic techniques that encourage gradient descent algorithms to produce networks with small weights.

Some of the results in this paper were presented at NIPS'96 [5].

1.1 Outline of the paper

The next section gives estimates of the misclassification probability in terms of the proportion of "distinctly correct" examples and the fat-shattering dimension. Section 3 gives some extensions to this result. Results in that section show that it is not necessary to specify in advance the margin by which the examples are distinctly correct. It also gives a lower bound on the misclassification probability in terms of a related scale-sensitive dimension, which shows that the upper bound in Section 2 is tight to within a log factor for a large family of function classes.

Section 4 gives bounds on the fat-shattering dimension for a variety of function classes, which imply misclassification probability estimates for these classes. In particular, Section 4.1 shows that in low dimensional Euclidean domains, any classification procedure that finds a decision boundary that is well separated from the examples will have good generalization performance, irrespective of the hypothesis class used by the procedure. Section 4.2 studies the fat-shattering dimension for neural networks, and Section 5 comments on the implications of this result for neural network learning algorithm design. Section 6 describes some recent related work and open problems.

2 Bounds on Misclassification Probability

We begin with some definitions.

Define the threshold function $\mathrm{sgn} : \mathbb{R} \to \{-1, 1\}$ as

$$\mathrm{sgn}(\alpha) = \begin{cases} -1 & \alpha < 0 \\ 1 & \alpha \geq 0. \end{cases}$$

Suppose X is a set (the input space), h is a real-valued function defined on X, and P is a probability distribution on $X \times \{-1, 1\}$. (Throughout, we ignore issues of measurability, and assume that all sets considered are measurable.) Define the misclassification probability of a hypothesis h as the probability that a random (x, y) pair is mislabelled,

$$\mathrm{er}_P(h) = P\left\{\mathrm{sgn}(h(x)) \neq y\right\}.$$

The training data is a sequence of elements of $X \times \{-1, 1\}$ that are generated independently according to the probability distribution P. For a training data sequence $z = ((x_1, y_1), \ldots, (x_m, y_m))$ of length m and a real number $\gamma > 0$, define the error estimate

$$\hat{\mathrm{er}}_z^\gamma(h) = \frac{1}{m}\left|\{i : y_i h(x_i) < \gamma\}\right|.$$

This estimate counts the proportion of examples that are not correctly classified with a margin of γ.

Let H be a class of real-valued functions defined on X. For $\gamma > 0$, a sequence $(x_1, \ldots, x_m)$ of m points from X is said to be γ-shattered by H if there is an $r = (r_1, \ldots, r_m) \in \mathbb{R}^m$ such that, for all $b = (b_1, \ldots, b_m) \in \{-1, 1\}$ there is an $h \in H$ satisfying $(h(x_i) - r_i)b_i \geq \gamma$. Define the fat-shattering dimension of H as the function

$$\mathrm{fat}_H(\gamma) = \max\left\{m : H \ \gamma\text{-shatters some } x \in X^m\right\}.$$

The fat-shattering dimension was introduced by Kearns and Schapire [26].

The following theorem gives a generalization error bound when the hypothesis makes no mistakes on the training examples and its value is bounded away from zero. The result is essentially the main result in [38], where it was observed that a similar but slightly weaker result follows trivially from the main result in [2]. The proof of this theorem is very similar to the proof in [2], which closely followed the proofs of Vapnik and Chervonenkis [41] and Pollard [35]. In this theorem and in what follows, we assume that X is a set, H is a class of real-valued functions defined on X, P is a probability distribution on $X \times \{-1, 1\}$, $0 < \delta < 1/2$, and $0 < \gamma < 1$.

Theorem 1 ([38]) *Suppose $z = ((x_1, y_1), \ldots, (x_m, y_m))$ is chosen by m independent draws from P. Then with probability at least $1 - \delta$, every h in H with $\hat{\mathrm{er}}_z^\gamma(h) = 0$ has*

$$\mathrm{er}_P(h) < \frac{2}{m}\left(d \log_2(34em/d) \log_2(578m) + \log_2(4/\delta)\right),$$

where $d = \mathrm{fat}_H(\gamma/16)$.

The next theorem is one of the two main technical results of the paper. It gives generalization error bounds when the hypothesis classifies a significant proportion of the training examples correctly, and its value is bounded away from zero for these points. In this case, it may be possible to get a better generalization error bound by excluding examples on which the hypothesis takes a value close to zero, even if these examples are correctly classified.

Theorem 2 *Suppose $z = ((x_1, y_1), \ldots, (x_m, y_m))$ is chosen by m independent draws from P. Then with probability at least $1 - \delta$, every h in H has*

$$\mathrm{er}_P(h) < \hat{\mathrm{er}}_z^\gamma(h) + \sqrt{\frac{2}{m}\left(d \ln(34em/d) \log_2(578) + \ln(4/\delta)\right)},$$

where $d = \mathrm{fat}_H(\gamma/16)$.

The idea of using the magnitudes of the values of $h(x_i)$ to give a more precise estimate of the generalization performance was first proposed in [40], and was further developed in [18] and [11]. There it was used only for the case of linear function classes. Rather than giving bounds on the generalization error, the results in [40] were restricted to bounds on the misclassification probability for a fixed test sample, presented in advance. The problem was further investigated in [37]. That paper gave a proof that Vapnik's result for the linear case could be extended to give bounds on misclassification probability. Theorem 1 generalizes this result to more arbitrary function classes. In [37] and [38] we also gave a more abstract result that provides generalization error bounds in terms of any hypothesis performance estimator ("luckiness function") that satisfies two properties (roughly, it must be consistent, and large values of the function must be unusual). Some applications are described in [38].

Lugosi and Pintér [33] have also obtained bounds on misclassification probability in terms of properties of regression functions. These bounds improve on the VC-bounds by using information about the behaviour of the true regression function (conditional expectation of y given x). Specifically, they show that the error of a skeleton-based estimator depends on certain covering numbers (with respect to an unusual pseudometric) of the class of possible regression functions, rather than the VC-dimension of the corresponding class of Bayes classifiers. Horváth and Lugosi [23] show that these covering numbers can be bounded in terms of a scale-sensitive dimension (which is closely related to the fat-shattering dimension of a squashed version of the function class—see Definition 3 below). However, these results do not extend to the case when the true regression function is not in the class of real-valued functions used by the estimator.

The error estimate $\hat{\mathrm{er}}_z^\gamma$ is related to Glick's smoothed error estimate (see, for example, [12], Chapter 31), which also takes into account the value of the real-valued prediction $h(x)$. The key feature of Glick's estimate is that it varies smoothly with $h(x)$, and hence in many cases provides a low variance (although biassed) estimate of the error.

The proof of Theorem 2 is in two parts. The first lemma uses an ℓ_∞ approximation argument, as well as the standard permutation technique to give sample complexity bounds in terms of ℓ_∞ covering numbers of a certain function class related to the hypothesis class. We then calculate these covering numbers.

Definition 3 *Suppose that (S, ρ) is a pseudometric space. For $A \subseteq S$, a set $T \subseteq S$ is an ϵ-cover of A with respect to ρ if for all a in A there is a t in T with $\rho(t, a) < \epsilon$. We define $\mathcal{N}(A, \epsilon, \rho)$ as the size of the smallest ϵ-cover of A.*

For a class F of functions defined on a set X and a sequence $x = (x_1, \ldots, x_m) \in X^m$, define the pseudometric $d_{\ell_\infty(x)}$ by

$$d_{\ell_\infty(x)}(f, g) = \max_i |f(x_i) - g(x_i)|.$$

Denote $\max_{x \in X^m} \mathcal{N}(A, \epsilon, d_{\ell_\infty(x)})$ *by* $\mathcal{N}_\infty(A, \epsilon, m)$.

For $\gamma > 0$ define $\pi_\gamma : \mathbb{R} \to \mathbb{R}$ as the piecewise-linear squashing function

$$\pi_\gamma(\alpha) = \begin{cases} \gamma & \text{if } \alpha \geq \gamma \\ -\gamma & \text{if } \alpha \leq -\gamma \\ \alpha & \text{otherwise.} \end{cases}$$

For a class H of functions mapping from a set X to $\mathbb{R}$, define

$$\pi_\gamma(H) = \{\pi_\gamma \circ h : h \in H\}.$$

Lemma 4 *Suppose $\gamma > 0$, $0 < \delta < 1/2$, P is a probability distribution on $X \times \{-1, 1\}$, and $z = ((x_1, y_1), \ldots, (x_m, y_m))$ is chosen by m independent draws from P. Then with probability at least $1 - \delta$, every h in H has*

$$\mathrm{er}_P(h) < \hat{\mathrm{er}}_z^\gamma(h) + \sqrt{\frac{2}{m} \ln\left(\frac{2\mathcal{N}_\infty(\pi_\gamma(H), \gamma/2, 2m)}{\delta}\right)}.$$

The proof uses techniques that go back to Vapnik and Chervonenkis [41] and Pollard [35], but using an ℓ_∞ cover as in [2], rather than the ℓ_1 covers used by Pollard.

Proof: Clearly, $\mathrm{er}_P(h) \leq P\{|\pi_\gamma(h(x)) - \gamma y| \geq \gamma\}$. Also, $y_i h(x_i) < \gamma$ if and only if $\pi_\gamma(h(x_i)) \neq \gamma y_i$, so we have

$$\begin{aligned} & \Pr\left(\exists h \in H,\ \mathrm{er}_P(h) \geq \hat{\mathrm{er}}_z^\gamma(h) + \epsilon\right) \\ \leq\ & \Pr\left(\exists h \in H,\ P\{|\pi_\gamma(h(x)) - \gamma y| \geq \gamma\} \geq \frac{1}{m} |\{i : \pi_\gamma(h(x_i)) \neq \gamma y_i\}| + \epsilon\right). \end{aligned}$$

We now relate this probability to a probability involving a second sample,

$$\tilde{z} = ((\tilde{x}_1, \tilde{y}_1), \ldots, (\tilde{x}_m, \tilde{y}_m)),$$

chosen independently according to P. Standard techniques (see, for example, [41]) show that the probability above is no more than

$$2\Pr\left(\exists h \in H,\ \frac{1}{m} |\{i : |\pi_\gamma(h(\tilde{x}_i)) - \gamma\tilde{y}_i| \geq \gamma\}| \geq \frac{1}{m} |\{i : \pi_\gamma(h(x_i)) \neq \gamma y_i\}| + \epsilon/2\right) \qquad (1)$$

(where the probability is over the double sample $(z, \tilde{z})$), provided $\epsilon^2 m \geq 2\ln 4$, and we shall see later that our choice of ϵ always satisfies this inequality. Next, we introduce a random permutation that swaps elements of z and $\tilde{z}$. Let U be the uniform distribution on the set of permutations σ on $\{1,\ldots,2m\}$ that swap some corresponding elements from the first and second half (that is, $\{\sigma(i),\sigma(i+m)\}=\{i,i+m\}$), and let w^σ denote $(w_{\sigma(1)},\ldots,w_{\sigma(2m)})$ for w in Z^{2m}. We denote the permuted elements of z and $\tilde{z}$ as $(z^\sigma,\tilde{z}^\sigma)=(z,\tilde{z})^\sigma$, and define the permuted vectors x^σ, $\tilde{x}^\sigma$, etc, in the obvious way. Then since $(z,\tilde{z})$ is chosen according to a product probability measure, the probability above is not affected by such a permutation, so (1) is no more than

$$2\sup_{(z,\tilde{z})} U\left\{\sigma : \exists h\in H,\ \frac{1}{m}\left|\{i : |\pi_\gamma(h(\tilde{x}_i^\sigma))-\gamma\tilde{y}_i^\sigma|\geq\gamma\}\right| \geq \frac{1}{m}\left|\{i : \pi_\gamma(h(x_i^\sigma))\neq\gamma y_i^\sigma\}\right|+\epsilon/2\right\}. \tag{2}$$

For a given $(z,\tilde{z})$, let T be a minimal $\gamma/2$-cover with respect to $d_{\ell_\infty(z,\tilde{z})}$ of the set $\pi_\gamma(H)$. That is, for all h in H, there is a g in T such that for $i=1,2,\ldots,2m$ we have $|\pi_\gamma(h(x_i))-g(x_i)|<\gamma/2$. For that h and g, it is clear that

$$\begin{aligned}\{i : |\pi_\gamma(h(\tilde{x}_i^\sigma))-\gamma\tilde{y}_i^\sigma|\geq\gamma\} &\subseteq \{i : |g(\tilde{x}_i^\sigma)-\gamma\tilde{y}_i^\sigma|\geq\gamma/2\}, \qquad \text{and}\\ \{i : |g(x_i^\sigma)-\gamma y_i^\sigma|\geq\gamma/2\} &\subseteq \{i : \pi_\gamma(h(x_i^\sigma))\neq\gamma y_i^\sigma\}.\end{aligned}$$

Hence the expression (2) is no more than

$$\begin{aligned} & 2\sup_{(z,\tilde{z})} U\left\{\sigma : \exists g\in T,\ \frac{1}{m}\left|\{i : |g(\tilde{x}_i^\sigma)-\gamma\tilde{y}_i^\sigma|\geq\gamma/2\}\right|\geq\frac{1}{m}\left|\{i : |g(x_i^\sigma)-\gamma y_i^\sigma|\geq\gamma/2\}\right|+\epsilon/2\right\}\\ \leq\ & 2\sup_{(z,\tilde{z})}|T|\sup_{g\in T} U\left\{\sigma : \frac{1}{m}\left|\{i : |g(\tilde{x}_i^\sigma)-\gamma\tilde{y}_i^\sigma|\geq\gamma/2\}\right|\geq\frac{1}{m}\left|\{i : |g(x_i^\sigma)-\gamma y_i^\sigma|\geq\gamma/2\}\right|+\epsilon/2\right\}\\ \leq\ & 2\mathcal{N}_\infty(\pi_\gamma(H),\gamma/2,2m)\sup_{\{a_i,b_i\}}\Pr\left(\frac{1}{m}\sum_{i=1}^m(a_i-b_i)\beta_i\geq\epsilon/2\right),\end{aligned}$$

where $a_i,b_i\in\{0,1\}$ satisfy $a_i=1$ iff $|g(\tilde{x}_i)-\gamma\tilde{y}_i|\geq\gamma/2$ and $b_i=1$ iff $|g(x_i)-\gamma y_i|\geq\gamma/2$, and the probability is over the β_i chosen independently and uniformly on $\{-1,1\}$. Hoeffding's inequality [22] implies that this is no more than $2\mathcal{N}_\infty(\pi_\gamma(H),\gamma/2,2m)\exp(-\epsilon^2 m/2)$. Setting this to δ and solving for ϵ gives the desired result. □

The following result of Alon et al [1] is useful to get bounds on these covering numbers.

Theorem 5 ([1]) *Consider a class F of functions that map from $\{1,\ldots,n\}$ to $\{1,\ldots,b\}$ with $\mathrm{fat}_F(1)\leq d$. Then*

$$\log_2\mathcal{N}_\infty(F,2,n)<1+\log_2(nb^2)\log_2\left(\sum_{i=0}^d\binom{n}{i}b^i\right),$$

provided that

$$n \geq 1 + \log_2 \left(\sum_{i=0}^{d} \binom{n}{i} b^i \right).$$

This result, together with a quantization argument, gives bounds on $\mathcal{N}_\infty(F, \gamma, m)$. We use this approach to prove Theorem 2.

Proof: (of Theorem 2) Define the quantization function $Q_\alpha : \mathbb{R} \to \mathbb{R}$ as $Q_\alpha(x) = \lceil (x - \alpha/2)/\alpha \rceil \alpha$. Define the class $F = Q_{\gamma/8}(\pi_\gamma(H))$ of quantized functions. Since $|Q_{\gamma/8}(a) - Q_{\gamma/8}(b)| < |a - b| + \gamma/16$, we have $\mathrm{fat}_F(\gamma/8) \leq \mathrm{fat}_{\pi_\gamma(H)}(\gamma/16)$. Let $\mathcal{M}_\infty(F, \alpha, m)$ denote the maximum over all $x \in X^m$ of the size of the largest subset of F for which all pairs of elements are α-separated with respect to $d_{\ell_\infty(x)}$. It is easy to see that $\mathcal{M}_\infty(\pi_\gamma(H), \gamma/2, 2m) \leq \mathcal{M}_\infty(F, \gamma/2, 2m)$, and it is well known that $\mathcal{N}_\infty(\pi_\gamma(H), \gamma/2, 2m) \leq \mathcal{M}_\infty(\pi_\gamma(H), \gamma/2, 2m)$ and $\mathcal{M}_\infty(F, \gamma/2, 2m) \leq \mathcal{N}_\infty(F, \gamma/4, 2m)$ (see [27]), hence

$$\mathcal{N}_\infty(\pi_\gamma(H), \gamma/2, 2m) \leq \mathcal{N}_\infty(F, \gamma/4, 2m).$$

Applying Theorem 5 with $n = 2m$ and $b = 17$ gives

$$\log_2 \mathcal{N}_\infty(\pi_\gamma(H), \gamma/2, 2m) < 1 + d \log_2(34em/d) \log_2(578m),$$

provided that $m \geq d \log_2(34em/d) + 1$, which can be assumed since the result is trivial otherwise. Substituting into Lemma 4, and observing that $\mathrm{fat}_{\pi_\gamma(H)}(\gamma/16) \leq \mathrm{fat}_H(\gamma/16)$ gives the desired result. $\square$

3 Discussion

Theorems 1 and 2 show that the accuracy of the error estimate $\hat{\mathrm{er}}_z^\gamma$ depends on the fat-shattering dimension rather than the VC-dimension. This can lead to large improvements over the VC bounds; the next section contains examples of function classes that have infinite VC-dimension but small fat-shattering dimension, and we shall see later in this section that for many function classes the fat-shattering dimension is always no more than a constant factor bigger than the VC-dimension. This decrease in estimation error comes at the cost of a possible increase in approximation error. Specifically, for a function class H it is possible to construct distributions P for which some h has small $\mathrm{er}_P(h)$ but with high probability every h in H has $\hat{\mathrm{er}}_z^\gamma(h)$ large. However, in many practical situations this is not relevant. For example, learning algorithms for neural networks typically minimize squared error, and for the distributions described above every h has large squared error (with high probability). So the distributions for which the use of the error estimate $\hat{\mathrm{er}}_z^\gamma(\cdot)$ incurs a large approximation error are those for which the learning algorithm fails in any case.

We can obtain a more general result that implies variants of Theorems 1 and 2. The following result can be proved using the techniques from the proof of Lemma 4, together with the proof of the corresponding result in [40] (or the simpler proof in [3]).

Theorem 6 *Suppose $\gamma > 0$, $0 < \delta < 1/2$, P is a probability distribution on $X \times \{-1, 1\}$, and $z = ((x_1, y_1), \ldots, (x_m, y_m))$ is chosen by m independent draws from P. Then*

$$\Pr\left(\exists h \in H : \frac{\mathrm{er}_P(h) - \hat{\mathrm{er}}_z^\gamma(h)}{\sqrt{\mathrm{er}_P(h)}} > \epsilon\right) \leq 4\mathcal{N}_\infty\left(\pi_\gamma(H), \gamma/2, 2m\right) \exp(-\epsilon^2 m/4).$$

Corollary 7 *Under the conditions of Theorem 6, and for all $\alpha > 0$,*

$$(i) \quad \Pr\left(\exists h \in H : \mathrm{er}_P(h) > \epsilon \text{ and } \hat{\mathrm{er}}_z^\gamma(h) = 0\right) \leq 4\mathcal{N}_\infty\left(\pi_\gamma(H), \gamma/2, 2m\right) \exp(-\epsilon m/4).$$

$$(ii) \quad \Pr\left(\exists h \in H : \mathrm{er}_P(h) > \hat{\mathrm{er}}_z^\gamma(h) + \epsilon\right) \leq 4\mathcal{N}_\infty\left(\pi_\gamma(H), \gamma/2, 2m\right) \exp(-\epsilon^2 m/4).$$

$$(iii) \quad \Pr\left(\exists h \in H : \mathrm{er}_P(h) > (1+\alpha)\hat{\mathrm{er}}_z^\gamma(h) + \epsilon\right) \leq 4\mathcal{N}_\infty\left(\pi_\gamma(H), \gamma/2, 2m\right) \times \exp\left(-\frac{\alpha^2 \epsilon m}{4(1+\alpha)^2}\right).$$

Proof: The proofs of (i) and (ii) are immediate. To see (iii), suppose that $\mathrm{er}_P(h) - \hat{\mathrm{er}}_z^\gamma(h) \leq \epsilon\sqrt{\mathrm{er}_P(h)}$, and consider separately the cases in which $\mathrm{er}_P(h)$ is smaller than and at least $(1 + 1/\alpha)^2\epsilon^2$. In either case, we conclude that

$$\mathrm{er}_P(h) \leq (1+\alpha)\hat{\mathrm{er}}_z^\gamma(h) + (1 + 1/\alpha)^2\epsilon^2.$$

□

Parts (i) and (ii) of this corollary give results essentially identical to Theorems 1 and 2, but with slightly worse constants.

In Theorems 1 and 2, the quantity γ (the margin by which hypothesis values are separated from 0) is specified in advance. This seems unnatural, since it is a quantity that will be observed after the examples are seen. It is easy to give a similar result in which the statement is made uniform over all values of this quantity. This follows from the following proposition.

Proposition 8 *Let $(X, \mathcal{F}, P)$ be a probability space, and let $\{E(\alpha_1, \alpha_2, \delta) : 0 < \alpha_1, \alpha_2, \delta \leq 1\}$ be a set of events satisfying*

1. *For all $0 < \alpha \leq 1$ and $0 < \delta \leq 1$, $P(E(\alpha, \alpha, \delta)) \leq \delta$.*

2. *For all $0 < a < 1$ and $0 < \delta \leq 1$, $\bigcup_{\alpha \in (0,1]} E(\alpha a, \alpha, \delta\alpha(1-a))$ is measurable, and*

3. *For all $0 < \alpha_1 \leq \alpha \leq \alpha_2 \leq 1$ and $0 < \delta_1 \leq \delta \leq 1$, $E(\alpha_1, \alpha_2, \delta_1) \subseteq E(\alpha, \alpha, \delta)$.*

Then for $0 < a, \delta < 1$, $P\left(\bigcup_{\alpha \in (0,1]} E(\alpha a, \alpha, \delta\alpha(1-a))\right) \leq \delta$.

Proof:

$$\begin{aligned}
P\left(\bigcup_{\alpha \in (0,1]} E(\alpha a, \alpha, \delta\alpha(1-a))\right) &= P\left(\bigcup_{i=0}^{\infty} \left\{E(\alpha a, \alpha, \delta\alpha(1-a)) : \alpha \in (a^{i+1}, a^i]\right\}\right) \\
&\leq P\left(\bigcup_{i=0}^{\infty} E(a^{i+1}, a^{i+1}, \delta a^i(1-a))\right) \\
&\leq \sum_{i=0}^{\infty} P\left(E(a^{i+1}, a^{i+1}, \delta a^i(1-a))\right) \\
&\leq \delta(1-a)\sum_{i=0}^{\infty} a^i = \delta.
\end{aligned}$$

□

This gives the following corollary of Theorems 1 and 2.

Corollary 9 *Suppose $z = ((x_1, y_1), \ldots, (x_m, y_m))$ is chosen by m independent draws from P.*

1. *With probability at least $1-\delta$, every h in H and every γ in $(0,1]$ with $\hat{\mathrm{er}}_z^\gamma(h) = 0$ have*

$$\mathrm{er}_P(h) < \frac{2}{m}\left(d \log_2(34em/d)\log_2(578m) + \log_2(8/(\gamma\delta))\right),$$

where $d = \mathrm{fat}_H(\gamma/32)$.

2. *With probability at least $1-\delta$, every h in H and every γ in $(0,1]$ have*

$$\mathrm{er}_P(h) < \hat{\mathrm{er}}_z^\gamma(h) + \sqrt{\frac{2}{m}\left(d\ln(34em/d)\log_2(578m) + \ln(8/(\gamma\delta))\right)},$$

where $d = \mathrm{fat}_H(\gamma/32)$.

Proof: For the first inequality, define $E(\gamma_1, \gamma_2, \delta)$ as the set of $z \in Z^m$ for which some h in H has $\hat{\mathrm{er}}_z^{\gamma_2}(h) = 0$ and

$$\mathrm{er}_P(h) \geq \frac{2}{m}\left(d \log_2(34em/d)\log_2(578m) + \log_2(4/\delta)\right),$$

where $d = \mathrm{fat}_H(\gamma_1/32)$. The result follows from the proposition with $a = 1/2$. The second inequality is derived similarly. ∎

Desirable behaviour of the fat-shattering dimension fat_H is clearly not necessary for good generalization performance bounds. It is only the behaviour of elements of the hypothesis class H in some neighbourhood of the origin that is important. As the proof shows, the generalization error bound can be expressed as a function of $\mathrm{fat}_{\pi_\gamma(H)}$. While it is possible to construct function classes for which this complexity measure is considerably smaller than fat_H (see, for example, [23]), the distinction is apparently not useful for applications.

It is possible to obtain generalization error bounds like those of Theorems 1 and 2 in terms of other versions of the fat-shattering dimension.

Definition 10 *For a class H of real-valued functions defined on X and $\gamma > 0$, a sequence $(x_1, \ldots, x_m)$ of m points from X is said to be uniformly γ-shattered by H if there is an $r \in \mathbb{R}$ such that, for all $b = (b_1, \ldots, b_m) \in \{-1, 1\}$ there is an $h \in H$ satisfying $(h(x_i) - r)b_i \geq \gamma$. Define*

$$\mathrm{fatV}_H(\gamma) = \max\{m : H \textit{ uniformly } \gamma\textit{-shatters some } x \in X^m\}.$$

We say that a sequence is uniformly γ-level-shattered by H if it is uniformly γ-shattered and $r = 0$ will suffice. We denote the corresponding dimension LfatV_H.

We use the notation fatV, as in [7], since this is a scale-sensitive version of a dimension introduced by Vapnik in [40]. The dimension LfatV has been used in approximation theory [31]. These complexity measures are closely related. Clearly, $\mathrm{LfatV}_H(\gamma) \leq \mathrm{fatV}_H(\gamma) \leq \mathrm{fat}_H(\gamma)$. If for every real number a and every function h in H we have $h + a \in H$ (that is, the class H has an adjustable output offset), then $\mathrm{LfatV}_H(\gamma) = \mathrm{fatV}_H(\gamma)$. It is also possible to show (by quantizing and then applying the pigeonhole principle—the proof is identical to that of Theorem 5 in [9]) that

$$\mathrm{fat}_{\pi_\gamma(H)}(\gamma/16) \leq c_1 \mathrm{fatV}_{\pi_\gamma(H)}(c_2\gamma)$$

for constants c_1 and c_2. It follows that for a class H with an adjustable output offset,

$$\mathrm{fat}_{\pi_\gamma(H)}(\gamma/16) \leq c_1 \mathrm{LfatV}_H(c_2\gamma),$$

so the bounds of Theorems 1 and 2 can also be expressed in terms of LfatV_H. Notice that $\mathrm{LfatV}_H(\gamma) \leq \mathrm{VCdim}(H)$ for all γ, so for classes with an adjustable output offset the bounds of Theorems 1 and 2 are always within log factors of the corresponding results from the VC-theory. For these classes, Theorem 11 below shows that the upper bounds are nearly optimal. (However, expressing the upper bounds in terms of LfatV_H introduces extra constant factors, and does not appear to be useful for applications.)

Theorem 11 *Suppose that X is a set, $Z = X \times \{-1, 1\}$, H is a class of real-valued functions defined on X, $m \geq 8$, L is a mapping from Z^m to H, $0 < \gamma < 1$, and $0 < \delta < 1/100$. Then there is a probability distribution P on Z such that*

1. Some function h in H satisfies $\mathrm{er}_P(h) = 0$ and $|h(x)| \geq \gamma$ almost surely, but

2. With probability at least δ over $z \in Z^m$ chosen according to P,

$$\mathrm{er}_P(L(z)) \geq \max\left(\frac{\mathrm{LfatV}_H(\gamma) - 1}{32m}, \frac{7\ln(1/\delta)}{8m}\right). \tag{3}$$

The proof makes use of the following lower bound for PAC-learning that is a special case of the main result in [13].

Lemma 12 ([13]) *If $X = \{1, 2, \ldots, d\}$, $Z = X \times \{-1, 1\}$, $m \geq 8$, $0 < \delta < 1/100$, and L is a mapping from Z^m to the class $\{-1, 1\}^X$ of all $\{-1, 1\}$-valued functions defined on X, then there is a distribution P on Z for which some $h : X \to \{-1, 1\}$ has $\mathrm{er}_P(h) = 0$ but with probability at least δ*

$$\mathrm{er}_P(L(z)) \geq \max\left(\frac{d-1}{32m}, \frac{7\ln(1/\delta)}{8m}\right). \tag{4}$$

Proof: (of Theorem 11.) Choose P so that its marginal distribution on X, P_X, has support on a uniformly γ-level-shattered set $X_0 \subseteq X$ of cardinality $d = \mathrm{LfatV}_H(\gamma)$. Then define $H_0 \subseteq H$ as the set of h in H for which $|h(x)| \geq \gamma$ for all x in X. Notice that P can be chosen so that the conditional distribution is concentrated on $\mathrm{sgn}(h(x))$ for some h in H_0. Clearly, for any such P the corresponding h satisfies the condition of the theorem. Without loss of generality, we can assume that L maps to H_0. Fix $z \in Z^m$. If L does not satisfy (3), then the corresponding mapping from Z^m to $\{-1,1\}^X$ does not satisfy (4). The result follows from the lemma. $\square$

The standard PAC learning results (see [40, 10]) show that, if the learning algorithm and error estimates are constrained to make use of the sample only through the function $\mathrm{er}_z : H \to [0,1]$ that maps from hypotheses to the proportion of training examples that they misclassify, there is no distribution independent error bound any better than $O(\mathrm{VCdim}(H)/m)$. Theorem 1 shows that if the learning algorithm also makes use of the sample through the functions $\hat{\mathrm{er}}_z^\gamma$, the bound can be better — as good as $O(\mathrm{fat}_H(\gamma)/m)$, ignoring log terms. (In the next section, we study function classes for which $\mathrm{fat}_H(\gamma)$ is finite when $\mathrm{VCdim}(H)$ is infinite.) Theorem 11 shows that there is no better distribution independent error bound if we only have access to the sample through these functions that the sample induces on H.

4 Bounds on fat_H

4.1 Lipschitz classes

This section considers classes of functions that are defined on a metric space and do not vary quickly. It turns out that for "small" metric spaces, such as low dimensional euclidean space, these function classes have small fat_H.

Theorem 13 *Let X be a totally bounded metric space with metric ρ. Suppose that H is a class of real-valued functions defined on X so that every h in H satisfies the Lipschitz condition*

$$|h(x) - h(y)| \leq L\rho(x,y).$$

Then $\mathrm{fat}_H(\gamma) \leq \mathcal{N}(X, \gamma/(2L), \rho)$.

Proof: Any two points in a γ-shattered set must be γ/L apart. It is well known (see, for example, [27]) that every γ/L-separated set in a totally bounded metric space (X,ρ) has cardinality no more than $\mathcal{N}(X,\gamma/(2L),\rho)$. $\square$

It is possible to use this result to give generalization error bounds for any binary-valued function class defined on a sufficiently small metric space, in terms of the number of points that are misclassified or close to the decision boundary. For a metric space (X, ρ) and a function $g : X \to \{-1, 1\}$, define $\mathrm{dist}(g, x)$ as the distance from x to the boundary of g,

$$\mathrm{dist}(g, x) = \inf \{\rho(x, x') : x' \in X,\ g(x) \neq g(x')\}.$$

Corollary 14 *Suppose that X is a totally bounded metric space with metric ρ and $\gamma > 0$, and define $d = \mathcal{N}(X, \gamma/32, \rho)$.*

1. *With probability at least $1 - \delta$ over $z \in Z^m$ chosen according to P, every measurable $\{-1, 1\}$-valued function g defined on X with $g(x_i) = y_i$ and $\mathrm{dist}(g, x_i) \geq \gamma$ for all i has*

$$\mathrm{er}_P(g) \leq \frac{2}{m} \left(d \log_2(34em/d) \log_2(578m) + \log_2(4/\delta)\right).$$

2. *With probability at least $1 - \delta$ over $z \in Z^m$ chosen according to P, every measurable $\{-1, 1\}$-valued function g defined on X satisfies*

$$\mathrm{er}_P(g) \leq \frac{1}{m} \left|\{i : \mathrm{dist}(g, x_i) < \gamma \text{ or } g(x_i) \neq y_i\}\right| + \sqrt{\frac{2}{m} \left(d \ln(34em/d) \log_2(578m) + \ln(4/\delta)\right)}.$$

Proof: Fix γ. Set $H_\gamma = \{x \mapsto \mathrm{sgn}(g(x))\mathrm{dist}(g, x) : g \in G\}$, where G is the set of measurable $\{-1, 1\}$-valued functions defined on X. For h in H_γ, if $\mathrm{sgn}(h(x)) \neq \mathrm{sgn}(h(x'))$ then $|h(x) - h(x')| \leq \rho(x, x')$. Also, if $\mathrm{sgn}(h(x)) = \mathrm{sgn}(h(x'))$, the triangle inequality for ρ implies that $|h(x) - h(x')| \leq \rho(x, x')$. So H_γ satisfies a Lipschitz condition, with constant $L = 1$. Theorem 13 implies that $\mathrm{fat}_{H_\gamma}(\gamma) \leq \mathcal{N}(X, \gamma/2, \rho)$. Theorems 1 and 2 give the result. ∎

So if the metric space is small, in the sense that $\mathcal{N}(X, \gamma, \rho)$ is small, any classification scheme producing a decision boundary that is far from the training examples and correctly classifies them (or the majority of them) will generalize well. In particular, if $X = [0, 1]^n$, $\mathcal{N}(X, \gamma, \rho) \sim \gamma^{-n}$. If the dimension n is small, this can give good generalization error bounds. For example, the "two spirals" problem was a popular test of the generalization performance of neural network learning algorithms [28]. In this case, X is a bounded subset of $\mathbb{R}^2$ and the nature of the problem means there is a large margin solution. So the result above shows that any classifier that gives a margin of at least some fixed value γ will have its error decreasing as a constant over m. This is true even if the classifier chooses these functions from a class with infinite VC-dimension.

4.2 Neural Networks

Neural networks are typically used as real-valued function classes, and are trained by minimizing squared error on the training examples, $\sum_{i=1}^m (y_i - h(x_i))^2$. The following observation shows that this procedure can work to approximately maximize the minimum value of $y_i h(x_i)$.

Proposition 15 *For a function h that maps from a set X to $\mathbb{R}$ and a sequence of examples $(x_1, y_1), \ldots, (x_m, y_m)$ from $X \times \{-1, 1\}$, if*

$$\frac{1}{m}\sum_{i=1}^m (h(x_i) - y_i)^2 < \epsilon$$

then

$$\frac{1}{m}\left|\{i : y_i h(x_i) < \alpha\}\right| < \epsilon/(1-\alpha)^2.$$

The remainder of this section derives bounds on the fat-shattering dimension for classes of functions computed by neural networks. Bounds on the pseudodimension of various neural network classes have been established (see, for example, [8, 19, 34, 25]), but these are all at least linear in the number of parameters. (The pseudodimension is equal to $\lim_{\gamma\to 0} \mathrm{fat}_H(\gamma)$, and hence gives an upper bound on $\mathrm{fat}_H(\gamma)$.) Gurvits and Koiran [17] have obtained an upper bound on the fat-shattering dimension for two-layer networks with bounded output weights and hidden units chosen from a class of binary-valued functions with finite VC-dimension. They obtain the following result for the case of linear threshold hidden units.

Proposition 16 ([17]) *Let F be the class of functions $f : x \mapsto \mathrm{sgn}(w \cdot x)$ defined on $\mathbb{R}^n$ for $n \geq 1$. Let H be the class of two-layer threshold networks with an arbitrary number of hidden units chosen from F and a bound A on the output weights,*

$$H = \left\{ \sum_{i=1}^N \alpha_i f_i : N \in \mathbb{N}, f_i \in F, \sum_{i=1}^N |\alpha_i| \leq A \right\}.$$

Then

$$\mathrm{fat}_H(\gamma) = O\left(\frac{A^2 n^2}{\gamma^2}\log(n/\gamma)\right).$$

The following result is the second key technical result in this paper. It gives a bound on the fat-shattering dimension for networks with real-valued hidden units (including sigmoid networks). In the special case of linear threshold functions, it gives a better bound (for large values of n/γ) than Proposition 16.

Theorem 17 *Let F be a nonempty class of functions that map from X to $[-M/2, M/2]$. For $A > 0$, define the class H of two-layer networks with hidden units chosen from F as*

$$H = \left\{ \sum_{i=1}^{N} w_i f_i : N \in \mathbb{N},\ f_i \in F,\ \sum_{i=1}^{N} |w_i| \leq A \right\}.$$

Suppose $\gamma > 0$ is such that $d = \mathrm{fat}_F(\gamma/(32A)) \geq 1$. Then

$$\mathrm{fat}_H(\gamma) \leq \frac{cM^2A^2d}{\gamma^2} \log^2(MAd/\gamma),$$

for some universal constant c.

The proof requires the introduction of two more pseudometrics and covering numbers.

Definition 18 *For real-valued functions defined on a set X, define the pseudometric $d_{\ell_1(x)}$ for $x = (x_1, \ldots, x_m) \in X^m$ by*

$$d_{\ell_1(x)}(f, g) = \frac{1}{m} \sum_{i=1}^{m} |f(x_i) - g(x_i)|.$$

Similarly, define $d_{\ell_2(x)}$ by

$$d_{\ell_2(x)}(f, g) = \left(\frac{1}{m} \sum_{i=1}^{m} (f(x_i) - g(x_i))^2 \right)^{1/2}.$$

If F is a set of functions defined on X, denote $\max_{x \in X^m} \mathcal{N}(F, \gamma, d_{\ell_1(x)})$ by $\mathcal{N}_1(F, \gamma, m)$, and similarly for $\mathcal{N}_2(F, \gamma, m)$.

The idea of the proof of Theorem 17 is to first derive a general upper bound on an ℓ_1 covering number of the class H, and then apply the following result (which is implicit in the proof of [6], Theorem 2) to give a bound on the fat-shattering dimension.

Lemma 19 ([6]) *If F is a class of $[0,1]$-valued functions defined on a set X with $\mathrm{fat}_F(4\gamma) \geq d$, then $\log_2 \mathcal{N}_1(F, \gamma, d) \geq d/32$.*

To derive an upper bound on $\mathcal{N}_1(H, \gamma, m)$, we start with the bound of Theorem 5 on the ℓ_∞ cover of the class F of hidden unit functions. This implies a bound on the ℓ_2 covering number. Then we use a result on approximation in ℓ_2 to give a bound on the ℓ_2 covering number of the network class H. This implies an upper bound on the ℓ_1 covering number, and comparing this with the lower bound of Lemma 19 gives Theorem 17.

Lemma 20 *Suppose that F is a class of $[-M/2, M/2]$-valued functions defined on a set X. If $d = \mathrm{fat}_F(\gamma/4)$ and $m \geq 2 + 2d\log_2(32M/\gamma)$, then*

$$\log_2 \mathcal{N}_2(F, \gamma, m) < 1 + d\log_2\left(\frac{4emM}{d\gamma}\right)\log_2\left(\frac{9mM^2}{\gamma^2}\right). \tag{5}$$

Proof: Using the same quantization argument as in the proof of Theorem 2, Theorem 5 shows that $m \geq 1 + d\log_2(4emM/(d\gamma))$ implies that $\log_2 \mathcal{N}_\infty(F, \gamma, m)$ is no more than the expression on the right hand side of (5). Since $d_{\ell_2}(f, g) \leq d_{l_\infty}(f, g)$, this implies the same bound for $\mathcal{N}_2(F, \gamma, m)$.

Now, since $\ln(xy) \leq xy$ for all $x, y > 0$, it suffices if

$$m \geq 1 + \frac{d}{\ln 2}\left(\frac{4emM}{d\gamma}y + \ln(1/y)\right).$$

Setting $y = \gamma \ln 2/(8eM)$ and solving for m gives the desired result. ∎

We will make use of the following result on approximation in Hilbert spaces, which has been attributed to Maurey (see [4, 24]).

Lemma 21 *Suppose G is a Hilbert space and $F \subseteq G$ has $\|f\| \leq b$ for all f in F. Let h be an element from the convex closure of F. Then for all $k \geq 1$ and all $c > b^2 - \|h\|^2$, there are functions $\{f_1, \ldots, f_k\} \subseteq F$ such that*

$$\left\|h - \frac{1}{k}\sum_{i=1}^{k} f_i\right\|^2 \leq \frac{c}{k}.$$

The following result uses this approximation lemma to relate ℓ_2 covering numbers of the classes F and H. This technique is due to Lee *et al* [30].

Lemma 22 *For the classes F and H of Theorem 17,*

$$\log_2 \mathcal{N}_2(H, \gamma, m) \leq \frac{2M^2A^2}{\gamma^2}\log_2\left(2\mathcal{N}_2\left(F, \frac{\gamma}{2A}, m\right) + 1\right).$$

Proof: Let $F_0 = F \cup -F \cup \{0\}$, where $-F = \{-f : f \in F\}$ and 0 is the identically zero function. Then the class H is the convex hull of F_0, scaled by A. Fix $x \in X^m$. With the norm $\|h\| = \left(\frac{1}{m}\sum_{i=1}^m h^2(x_i)\right)^{1/2}$, Lemma 21 shows that, for any h in H and any k there are functions $f_1, \ldots, f_k$ in F_0 for which

$$d_{\ell_2(x)}\left(h, \frac{A}{k}\sum_{i=1}^{k} f_i\right) \leq \frac{AM}{2\sqrt{k}}.$$

If, instead of $f_i \in F_0$ we choose $\hat{f}_i$ in a $\gamma/(2A)$-cover of F_0, such that $d_{\ell_2(x)}(f_i, \hat{f}_i) < \gamma/(2A)$, then the triangle inequality implies that

$$d_{\ell_2(x)}\left(h, \frac{A}{k}\sum_{i=1}^{k} \hat{f}_i\right) < \frac{AM}{2\sqrt{k}} + \frac{\gamma}{2}.$$

It follows that we can construct a γ-cover of H from a $\gamma/(2A)$-cover of F_0, by selecting all subsets of the cover of size $k = \lceil M^2A^2/\gamma^2 \rceil$. Some $k \leq 2M^2A^2/\gamma^2$ will suffice, since if $\gamma > MA$ the lemma is trivially true. Hence,

$$\mathcal{N}_2(H, \gamma, m) \leq (2\mathcal{N}_2(F, \gamma/(2A), m) + 1)^{2M^2A^2/\gamma^2}.$$

□

We can now combine this result with Lemmas 19 and 20 to prove the theorem.

Proof: (of Theorem 17) From Lemmas 19, 20, and 22, if $m = \mathrm{fat}_H(4\gamma) \geq 2 + 2d\log_2(64MA/\gamma)$ with $d = \mathrm{fat}_F(\gamma/(8A))$, then

$$m \leq \frac{64M^2A^2}{\gamma^2}\left(3 + d\log_2\left(\frac{8emMA}{\gamma}\right)\log_2\left(\frac{36mM^2A^2}{\gamma^2}\right)\right). \qquad (6)$$

Since we may assume that $\gamma \leq MA$, if $m \geq 2$ then $2 + 2d\log_2(64MA/\gamma)$ is no more than the expression on the right of (6). So either $m = \mathrm{fat}_H(4\gamma) \leq 1$, or

$$m \leq \frac{64M^2A^2}{\gamma^2}\left(3 + d\log_2^2\left(\frac{36mM^2A^2}{\gamma^2}\right)\right).$$

Now, for all $x, y > 0$, $\ln\left(\sqrt{xy}\right) < \sqrt{xy}$, so

$$\ln^2 x \leq (2\sqrt{xy} + \ln(1/y))^2 \leq 4x(y + \sqrt{y}\ln(1/y)) + \ln^2(1/y),$$

provided $y < 1$ and $x > 1$. It follows that, for $b, c > 0$, $b\ln^2(cx) \leq x/2 + b\ln^2(1/y)$, provided that $4bc(y + \sqrt{y}\ln(1/y)) \leq 1/2$. It is easy to see that $y \leq (16bc)^{-4}$ will suffice. That is, $b\ln^2(cx) \leq x/2 + 16b\ln^2(16bc)$. Applying this to inequality (6) with $x = m$, and replacing γ by $\gamma/4$ gives the result. □

We can apply these techniques to give bounds on the fat-shattering dimension of many function classes. In this context, it is useful to consider the pseudodimension of a function class. Recall that the pseudodimension of a class F can be defined as $\dim_P(F) =$

$\lim_{\gamma\to 0}\mathrm{fat}_F(\gamma)$, and that this provides an upper bound on $\mathrm{fat}_F(\gamma)$ for all γ, since fat_F is a non-increasing function. We can use such a bound for a class F, together with Theorem 17, to give bounds on the fat-shattering dimension of the class of bounded linear combinations of functions from F. However, we can obtain better bounds using the following result, due to Haussler and Long [20].

Lemma 23 ([20]) *Let F be a class of functions that take values in $[-M/2, M/2]$ with $d = \dim_P(F)$ finite. For any $m \geq d$ and any $\gamma > 0$,*

$$\log \mathcal{N}_\infty(F, \gamma, m) < d \log\left(\frac{emM}{\gamma d}\right).$$

A simple example of the application of these techniques is to the class of two-layer neural networks.

Corollary 24 *Let $\sigma : \mathbb{R} \to [-M/2, M/2]$ be a non-decreasing function. Define the class F of functions on $\mathbb{R}^n$ as $F = \{x \mapsto \sigma(w \cdot x + w_0) : w \in \mathbb{R}^n,\ w_0 \in \mathbb{R}\}$, and define*

$$H = \left\{\sum_{i=1}^{N} \alpha_i f_i : N \in \mathbb{N},\ f_i \in F,\ \sum_{i=1}^{N} |\alpha_i| \leq A\right\}$$

for $A \geq 1$. Then for $\gamma \leq MA$,

$$\mathrm{fat}_H(\gamma) \leq \frac{cM^2A^2n}{\gamma^2}\log\left(\frac{MA}{\gamma}\right),$$

for some universal constant c.

Proof: Given the conditions on σ, F has $\dim_P(F) \leq n+1$ (see, for example, [19]). Applying Lemmas 23, 19 and 22 and solving for $\mathrm{fat}_H(\gamma)$ gives the result. □

Similar bounds can be obtained if the first layer function class is replaced by the class of functions computed by a multilayer network with a bounded number of parameters, and computation units with either a threshold transfer function, a piecewise polynomial transfer function, or the standard sigmoid, $\sigma(\alpha) = (1 - e^{-\alpha})/(1 + e^{-\alpha})$. Bounds for $\dim_P(F)$ are known in these cases (see [8], [16], and [25] respectively).

Composing these functions with a smooth squashing function does not greatly affect these bounds. For the remainder of this section, we fix a squashing function $\sigma : \mathbb{R} \to [-M/2, M/2]$, and assume that it satisfies the following Lipschitz condition: for some $L > 0$ and all $x_1, x_2 \in \mathbb{R}$,

$$|\sigma(x_1) - \sigma(x_2)| \leq L|x_1 - x_2|.$$

For a class F of real-valued functions, let $\sigma(F) = \{\sigma \circ f : f \in F\}$. The proof of the following result is trivial.

Proposition 25 *For a class F of real-valued functions and for all $\gamma > 0$,*

1. $\mathrm{fat}_{\sigma(F)}(\gamma) \leq \mathrm{fat}_F(\gamma/L)$, *and*
2. $\mathcal{N}_p(\sigma(F), \gamma, m) \leq \mathcal{N}_p(F, \gamma/L, m)$ *for all $m \in \mathbb{N}$ and $p \in \{1, 2, \infty\}$.*

Using this result, we can apply the techniques described above to obtain bounds on the fat-shattering dimension for deeper networks.

Let F be a nonempty class of $[-M/2, M/2]$-valued functions defined on a set X. Let $H_0 = F$, and for $\ell \geq 1$, let

$$H_\ell = \left\{ \sigma\left(\sum_{i=1}^{N} w_i f_i \right) : N \in \mathbb{N},\ f_i \in \bigcup_{i=0}^{\ell-1} H_i,\ \sum_{i=1}^{N} |w_i| \leq A \right\},$$

with σ defined as above.

Lemma 26 *For any $m, \ell \geq 1$, and $0 < \gamma < 2MAL$, we have*

$$\log_2 \mathcal{N}_2(H_\ell, \gamma, m) \leq \left(\frac{2MAL}{\gamma} \right)^{2\ell} (2AL)^{\ell(\ell-1)} \log_2 \left(3^\ell \ell!\ \mathcal{N}_2(F, \gamma/(2AL)^\ell, m) \right).$$

Proof: The result is clearly true for $H_0 = F$. Suppose that it is true for H_ℓ. Then by Lemma 22 and Proposition 25,

$$\begin{aligned}
\log_2 \mathcal{N}_2(H_{\ell+1}, \gamma, m) &\leq \left(\frac{2MAL}{\gamma} \right)^2 \log_2 \left(2 \sum_{i=0}^{\ell} \mathcal{N}_2(H_i, \gamma/(2AL), m) + 1 \right) \\
&\leq \left(\frac{2MAL}{\gamma} \right)^2 \Bigg(\log_2(3(\ell+1)) + \\
&\qquad \left(\frac{2MAL}{\gamma} \right)^{2\ell} (2AL)^{\ell(\ell+1)} \log_2 \left(3^\ell \ell!\ \mathcal{N}_2(F, \gamma/(2AL)^{\ell+1}, m) \right) \Bigg) \\
&\leq \left(\frac{2MAL}{\gamma} \right)^{2(\ell+1)} (2AL)^{\ell(\ell+1)} \times \\
&\qquad \log_2 \left(3^{\ell+1} (\ell+1)!\ \mathcal{N}_2(F, \gamma/(2AL)^{\ell+1}, m) \right).
\end{aligned}$$

$\Box$

The following corollary gives a bound on the fat-shattering dimension for multi-layer sigmoid networks with a bound A on the ℓ_1 norm of the parameters in each computational unit.

Corollary 27 *Let $X = \{x \in \mathbb{R}^n : \|x\|_\infty \leq B\}$. Let F be the class of functions on X defined by $F = \{(x_1, \ldots, x_n) \mapsto x_i : 1 \leq i \leq n\}$. For this class, define as above the classes $H_1, H_2, \ldots, H_\ell$ of functions computed by a sigmoid network with $1, \ldots, \ell$ layers. Then for $\ell \geq 2$,*

$$\mathrm{fat}_{H_\ell}(\gamma) \leq \frac{4B^2}{\gamma^2}\left(\frac{M}{\gamma}\right)^{2(\ell-1)}(2AL)^{\ell(\ell+1)}\log_2\left(3^{\ell-1}(\ell-1)!\,(2n+1)\right),$$

and in particular

$$\mathrm{fat}_{H_2}(\gamma) \leq \frac{cM^2B^2(AL)^6}{\gamma^4}\log n,$$

for some universal constant c.

It is also easy to derive analogous results for radial basis function networks. In fact, these techniques give bounds on the fat-shattering dimension for any function class that contains compositions of elements of a class with finite fat-shattering dimension with a bounded number of compositions of bounded-weight linear combinations or scalar Lipschitz functions.

5 Discussion

Together, Theorem 2 and Corollaries 24 and 27 give the following result.

Theorem 28 *Suppose P is a probability distribution on $X = \mathbb{R}^n$, $0 < \gamma \leq 1$, and $0 < \delta < 1/2$.*

1. *Let $\sigma : \mathbb{R} \to [-1, 1]$ be a non-decreasing function. Define the class F of functions on $\mathbb{R}^n$ as $F = \{x \mapsto \sigma(w \cdot x + w_0) : w \in \mathbb{R}^n,\ w_0 \in \mathbb{R}\}$, and define*

$$H = \left\{\sum_{i=1}^N \alpha_i f_i : N \in \mathbb{N},\ f_i \in F,\ \sum_{i=1}^N |\alpha_i| \leq A\right\}$$

for $A \geq 1$. Then with probability at least $1-\delta$ over a training sample $z \in Z^m$ chosen according to P, every h in H has

$$\mathrm{er}_P(h) < \hat{\mathrm{er}}_z^\gamma(h) + \sqrt{\frac{c}{m}\left(\frac{A^2 n}{\gamma^2}\log\left(\frac{A}{\gamma}\right)\log^2 m + \log(1/\delta)\right)},$$

for some universal constant c.

2. *Let $\sigma : \mathbb{R} \to [-1,1]$ satisfy $|\sigma(x_1) - \sigma(x_2)| \le L|x_1 - x_2|$ for all $x_1, x_2 \in \mathbb{R}$. Let $X = \{x \in \mathbb{R}^n : \|x\|_\infty \le B\}$. Let H_0 be the class of functions on X defined by*

$$H_0 = \{(x_1, \ldots, x_n) \mapsto x_i : 1 \le i \le n\},$$

and for $\ell \ge 1$, let

$$H_\ell = \left\{ \sigma\left(\sum_{i=1}^{N} w_i f_i \right) : N \in \mathbb{N},\ f_i \in \bigcup_{i=0}^{\ell-1} H_i,\ \sum_{i=1}^{N} |w_i| \le A \right\},$$

for $A \ge 1$. Then for any fixed depth $\ell \ge 1$, with probability at least $1 - \delta$ over a training sample $z \in Z^m$ chosen according to P, every h in H_ℓ has

$$\mathrm{er}_P(h) < \hat{\mathrm{er}}_z^\gamma(h) + \sqrt{\frac{c}{m}\left(\frac{B^2 (AL)^{\ell(\ell+1)}}{\gamma^{2\ell}} \log n \log^2 m + \log(1/\delta) \right)},$$

for some constant c that depends only on ℓ.

Notice that these networks have infinite VC-dimension. This result provides a plausible explanation for the generalization performance of neural networks: if, in applications, there are networks with many small weights but small squared error on the training examples, then the VC-dimension (and hence number of parameters) is irrelevant to the generalization performance. Instead, the magnitude of the weights in the network is more important.

These results are not sensitive to the form of the squashing function σ. Part 1 of Theorem 28 requires only that it be non-decreasing and have bounded range, and Part 2 (for deeper nets) requires that it satisfies a Lipschitz condition. This is in contrast to the VC-dimension results, which are sensitive to small changes in the function σ.

Applying Corollary 9 gives a similar result in which we can choose γ after seeing the data, in order to optimize the bound on $\mathrm{er}_P(h)$. Choosing a small value of γ corresponds to examining the behaviour of the network on a fine scale, and leads to a large complexity penalty. A larger value of γ gives a smaller complexity penalty, perhaps with some increase in the error estimate $\hat{\mathrm{er}}_z^\gamma(h)$.

We can also use Proposition 8 to give the following result, in which we can choose both γ and A (the bound on the parameter magnitudes) after seeing the data. Define the class of two-layer networks with output weights bounded by A as

$$H_A = \left\{ \sum_{i=1}^{N} \alpha_i f_i : N \in \mathbb{N},\ f_i \in F, \sum_{i=1}^{N} |\alpha_i| \leq A \right\},$$

where F is the class of hidden unit functions defined in Theorem 28 part 1.

Corollary 29 *Suppose P is a probability distribution on $Z = X \times \{-1, 1\}$ and $0 < \delta < 1/2$. With probability at least $1 - \delta$ over $z \in Z^m$ chosen according to P, for every $0 < \gamma \leq 1$, $A \geq 1$, and $h \in H_A$,*

$$\mathrm{er}_P(h) < \hat{\mathrm{er}}_z^{\gamma}(h) + \sqrt{\frac{c}{m}\left(\frac{A^2 n}{\gamma^2}\log\left(\frac{nA}{\gamma}\right)\log^2 m + \log\left(\frac{A}{\delta\gamma}\right)\right)}, \tag{7}$$

for some universal constant c.

The corollary follows from Theorem 28 part 1, on applying Proposition 8 with α representing $1/A$. A similar corollary of Theorem 28 part 2 follows in the same way.

This complexity regularization result suggests the use of an algorithm that chooses an h from $\bigcup_{A\geq 1} H_A$ to minimize the right hand side of (7), in order to give the best bound on misclassification probability. This is qualitatively similar to popular heuristics (such as "weight decay" and "early stopping" — see, for example, [21]) that attempt to find neural network functions that have small error and small weights. In the weight decay heuristic, a penalty term involving the magnitudes of the network weights is added to the cost function, so that the learning algorithm aims to trade squared error for weight magnitudes. The early stopping heuristic restricts a gradient descent algorithm to take only a small number of steps in weight space in a direction that reduces the training sample error. For a fixed step size and small initial weight values, this ensures that the magnitudes of the weights cannot be large after training.

One approach to the problem of minimizing squared error while maintaining small weights is described in [30]. The algorithm analyzed in that paper solves the problem for two-layer networks with linear threshold hidden units. If these units have fan-in bounded by a constant, the algorithm runs in polynomial time. It follows that, if there is a network with small total squared error on the training examples, this algorithm will quickly find a network with small misclassification probability.

Results in this paper also have implications for regression using neural networks. The algorithm described in [30] finds a two-layer network that estimates a real-valued quantity with near minimal squared error. For that algorithm, the estimation error (the difference between the expected squared error of the network and the error of the optimal network) is bounded above by a quantity that increases with the size of the parameters, but is independent of the number of parameters. The bound on the fat-shattering dimension (and covering numbers) given in Corollary 27 immediately imply similar results (but with a slower rate of convergence) for regression using deeper networks. Again, the bounds on estimation error depend on the parameter magnitudes but not on the number of parameters.

6 Further Work

No serious effort has been made to optimize the constants in the results in this paper. Recent work [36] using a more direct proof technique gives a log factor improvement on the estimation rate in Theorem 28. Further improvements might be possible.

It would also be worthwhile to determine how well the generalization performance estimates provided by Theorem 28 coincide with the actual performance of neural networks in pattern classification problems. A preliminary investigation in [32] for an artificial pattern classification problem reveals that the relationship between misclassification probability and the parameter magnitudes is qualitatively very similar to the estimates given here. It would be interesting to determine if this is also true in real pattern classification problems.

Related techniques have recently been used [36] to explain the generalization performance of boosting algorithms [15, 14], which use composite hypotheses that are convex combinations of hypotheses produced by weak learning algorithms.

Acknowledgements

Thanks to Andrew Barron, Jonathan Baxter, Mostefa Golea, Mike Jordan, Adam Kowalczyk, Wee Sun Lee, Phil Long, Gabor Lugosi, Llew Mason, John Shawe-Taylor, Robert Schapire, and Robert Slaviero for helpful discussions and comments. Thanks also to the reviewers for

many helpful suggestions.

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. To appear in *Journal of the ACM*, 1997.

[2] M. Anthony and P. Bartlett. Function learning from interpolation. In *Computational Learning Theory: EUROCOLT'95*, 1995.

[3] M. Anthony and J. Shawe-Taylor. A result of Vapnik with applications. *Discrete Applied Mathematics*, 47:207–217, 1993.

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

[5] P. L. Bartlett. For valid generalization, the size of the weights is more important than the size of the network. To appear in *Neural Information Processing Systems 9*, 1997.

[6] P. L. Bartlett, S. R. Kulkarni, and S. E. Posner. Covering numbers for real-valued function classes. Technical report, Australian National University and Princeton University, 1996.

[7] P. L. Bartlett and P. M. Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *Journal of Computer and System Sciences*, to appear, 1997.

[8] E. Baum and D. Haussler. What size net gives valid generalization? *Neural Computation*, 1(1):151–160, 1989.

[9] S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P. M. Long. Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. *Journal of Computer and System Sciences*, 50(1):74–86, 1995.

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

[11] B. Boser, I. Guyon, and V. Vapnik. A training algorithm for optimal margin classifiers. In *Proceedings of the Fifth Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press, 1992.

[12] L. Devroye, L. Györfi, and G. Lugosi. *A probabilistic theory of pattern recognition.* Springer-Verlag, New York, 1996.

[13] A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82:247–261, 1989.

[14] Y. Freund and R. E. Schapire. A decision-theoretic generalization of online learning and an application to boosting. In *Computational Learning Theory: Second European Conference, EUROCOLT'95*, pages 23–37. Springer-Verlag, 1995.

[15] Y. Freund and R. E. Schapire. Experiments with a new boosting algorithm. In *Machine Learning: Proceedings of the Thirteenth International Conference*, 1996.

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

[17] L. Gurvits and P. Koiran. Approximation and learning of convex superpositions. In *Computational Learning Theory: EUROCOLT'95*, 1995.

[18] I. Guyon, B. Boser, and V. Vapnik. Automatic capacity tuning of very large vc-dimension classifiers. In *NIPS 5*, pages 147–155. Morgan Kaufmann, 1993.

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

[20] D. Haussler and P.M. Long. A generalization of Sauer's Lemma. *Journal of Combinatorial Theory, Series A*, 71(2):219–240, 1995.

[21] J. Hertz, A. Krogh, and R. G. Palmer. *Introduction to the Theory of Neural Computation.* Addison-Wesley, 1991.

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

[23] M. Horvath and G. Lugosi. A data-dependent skeleton estimate and a scale-sensitive dimension for classification. Technical report, Pompeu Fabra University, December 1996.

[24] L. K. Jones. A simple lemma on greedy approximation in hilbert space and convergence rates for projection pursuit regression and neural network training. *The Annals of Statistics*, 20(1):608–613, 1992.

[25] M. Karpinski and A. Macintyre. Quadratic VC-dimension bounds for sigmoidal networks. In *Proceedings of the 27th Annual Symposium on the Theory of Computing*, 1995.

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

[27] A. N. Kolmogorov and V. M. Tihomirov. ϵ-entropy and ϵ-capacity of sets in function spaces. *American Mathematics Society Translations (2)*, 17:277–364, 1961.

[28] K. J. Lang and M. Witbrock. Learning to tell two spirals apart. In *Proceedings of the 1988 Connectionist Models Summer School.* Morgan Kaufmann, 1988.

[29] S. Lawrence, C. L. Giles, and A. C. Tsoi. What size neural network gives optimal generalization? Convergence properties of backpropagation. Technical Report UMIACS-TR-96-22 and CS-TR-3617, Institute for Advanced Computer Studies, University of Maryland, April 1996.

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

[31] G. G. Lorentz. *Approximation of functions.* Holt, Rinehart and Winston, New York, 1966.

[32] G. Loy and P. L. Bartlett. Generalization and the size of the weights: an experimental study. To appear in *Proceedings of the Eighth Australian Conference on Neural Networks*, 1997.

[33] G. Lugosi and M. Pintér. A data-dependent skeleton estimate for learning. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pages 51–56. ACM Press, New York, NY, 1996.

[34] W. Maass. Vapnik-Chervonenkis dimension of neural nets. In M. A. Arbib, editor, *The Handbook of Brain Theory and Neural Networks*, pages 1000–1003. MIT Press, Cambridge, 1995.

[35] D. Pollard. *Convergence of Stochastic Processes.* Springer, New York, 1984.

[36] R. E. Schapire, Y. Freund, P. L. Bartlett, and W. S. Lee. Boosting the margin: a new explanation for the effectiveness of voting methods. To appear in *Proceedings of the International Conference on Machine Learning*, 1997.

[37] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. A framework for structural risk minimisation. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pages 68–76. ACM Press, New York, NY, 1996.

[38] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies, September 1996. Technical Report.

[39] E. D. Sontag. Feedforward nets for interpolation and classification. *Journal of Computer and System Sciences*, 45:20–48, 1992.

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

[41] 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.