Structural Risk Minimization over Data-Dependent Hierarchies

John Shawe-Taylor
Department of Computer Science
Royal Holloway and Bedford New College
University of London
Egham, TW20 0EX, UK
jst@dcs.rhbnc.ac.uk

Peter L. Bartlett
Department of Systems Engineering
Australian National University
Canberra 0200 Australia
Peter.Bartlett@anu.edu.au

Robert C. Williamson
Department of Engineering
Australian National University
Canberra 0200 Australia
Bob.Williamson@anu.edu.au

Martin Anthony
Department of Mathematics
London School of Economics
Houghton Street
London WC2A 2AE, UK
M.Anthony@lse.ac.uk

November 28, 1997

Abstract

The paper introduces some generalizations of Vapnik's method of structural risk minimisation (SRM). As well as making explicit some of the details on SRM, it provides a result that allows one to trade off errors on the training sample against improved generalization performance. It then considers the more general case when the hierarchy of classes is chosen in response to the data. A result is presented on the generalization performance of classifiers with a "large margin". This theoretically explains the impressive generalization performance of the maximal margin hyperplane algorithm of Vapnik and co-workers (which is the basis for their support vector machines). The paper concludes with a more general result in terms of "luckiness" functions, which provides a quite general way for exploiting serendipitous simplicity in observed data to obtain better prediction accuracy from small training sets. Four examples are given of such functions, including the VC dimension measured on the sample.

Keywords: Learning Machines, Maximal Margin, Support Vector Machines, Probable Smooth Luckiness, Uniform Convergence, Vapnik-Chervonenkis Dimension, Fat Shattering Dimension, Computational Learning Theory, Probably Approximately Correct Learning.

1 Introduction

The standard Probably Approximately Correct (PAC) model of learning considers a fixed hypothesis class H together with a required accuracy ϵ and confidence $1-\delta$. The theory characterises when a target function from H can be learned from examples in terms of the Vapnik-Chervonenkis dimension, a measure of the flexibility of the class H and specifies sample sizes required to deliver the required accuracy with the allowed confidence.

In many cases of practical interest the precise class containing the target function to be learned may not be known in advance. The learner may only be given a hierarchy of classes

$$H_1 \subseteq H_2 \subseteq \cdots \subseteq H_d \subseteq \cdots$$

and be told that the target will lie in one of the sets H_d.

Structural Risk Minimization (SRM) copes with this problem by minimizing an upper bound on the expected risk, over each of the hypothesis classes. The principle is a curious one in that in order to have an *algorithm* it is necessary to have a good theoretical bound on the generalization performance. A formal statement of the method is given in the next section.

Linial, Mansour and Rivest [29] studied learning in a framework as above by allowing the learner to seek a consistent hypothesis in each subclass H_d in turn, drawing enough extra examples at each stage to ensure the correct level of accuracy and confidence should a consistent hypothesis be found.

This paper[1] addresses two shortcomings of Linial *et al.*'s approach. The first is the requirement to draw extra examples when seeking in a richer class. It may be unrealistic to assume that examples can be obtained cheaply, and at the same time it would be foolish not to use as many examples as are available from the start. Hence, we suppose that a fixed number of examples is allowed and that the aim of the learner is to bound the expected generalization error with high confidence. The second drawback of the Linial *et al.* approach is that it is not clear how it can be adapted to handle the case where errors are allowed on the training set. In this situation there is a need to trade off the number of errors with the complexity of the class, since taking a class which is too complex can result in a worse generalization error (with a fixed number of examples) than allowing some extra errors in a more restricted class.

The model we consider allows a precise bound on the error arising in different classes and hence a reliable way of applying the structural risk minimisation principle introduced by Vapnik [48, 50]. Indeed, the results reported in Sections 2 and 3 of this paper are implicit in the cited references, but our treatment serves to introduce the main results of the paper in later sections, and we make explicit some of the assumptions implicit in the presentations in [48, 50]. A more recent paper by Lugosi and Zeger [38] considers standard SRM and provides bounds for the true error of the hypothesis with lowest empirical error in each class. Whereas our Theorem 2.3 gives an error bound that decreases to twice the empirical error roughly linearly with the ratio

[1] Some of the results of this paper appeared in [43].

of the VC dimension to the number of examples, they give an error bound that decreases to the empirical error itself, but as the square root of this ratio.

From Section 4 onwards we address a shortcoming of the SRM method which Vapnik [48, page 161] highlights: *according to the SRM principle the structure has to be defined* a priori *before the training data appear.* An algorithm using maximally separating hyperplanes proposed by Vapnik [46] and co-workers [14, 16] violates this principle in that the hierarchy defined depends on the data. In Section 4 we prove a result which shows that if one achieves correct classification of some training data with a class of $\{0,1\}$-valued functions which are thresholded, and if the values of the real-valued functions on the training points are all well away from zero, then there is a bound on the generalization error which can be much better than the one obtained from the VC-dimension of the thresholded class. In Section 5 we apply this to the case considered by Vapnik: separating hyperplanes with a large margin.

In Section 6 we introduce a more general framework which allows a rather large class of methods of measuring the luckiness of a sample, in the sense that the large margin is "lucky". In Section 7 we explictly show how Vapnik's maximum margin hyperplanes fit into this general framework, which then also allows the radius of the set of points to be estimated from the data. In addition, we show that the function which measures the VC dimension of the set of hypotheses on the sample points is a valid (un)luckiness function. This leads to a bound on the generalization performance in terms of this measured dimension rather than the "worst case" bound which involves the VC dimension of the set of hypotheses over the whole input space.

Our approach can be interpreted as a general way of encoding our bias, or prior assumptions, and possibly taking advantage of them if they happen to be correct. In the case of the fixed hierarchy, we expect the target (or a close approximation to it) to be in a class H_d with small d. In the maximal separation case, we expect the target to be consistent with some classifying hyperplane that has a large separation from the examples. This corresponds to a collusion between the probability distribution and the target concept, which would be impossible to exploit in the standard PAC distribution independent framework. If these assumptions happen to be correct for the training data, we can be confident we have an accurate hypothesis from a small data set (at the expense of some small penalty if they are incorrect).

A commonly studied related problem is that of *model order selection* (see for example [34]), and we here briefly make some remarks on the relationship with the work presented in this paper. Assuming the above hierarchy of hypothesis classes, the aim there is to identify the best class index. Often "best" in this literature simply means "correct" in the sense that if in fact the target hypothesis $h \in H_i$, then as the sample size grows to infinity, the selection procedure will (in some probabilistic sense) pick i. Other methods of "complexity regularization" can be seen to also solve similar problems. (See for example [20, 6, 7, 8].) We are not aware of any methods (apart from SRM) for which explicit finite sample size bounds on their performance are available. Furthermore, with the exception of the methods discussed in [8], all such methods take the form of minimizing a cost function comprising an empirical risk plus an additive complexity term which does not depend on the data.

We denote logarithms to base 2 by log, and natural logarithms by ln. If S is a set, $|S|$ denotes its cardinality. We do not explictly state the measurability conditions needed for our arguments to hold. We assume with no further discussion "permissibility" of the function classes involved (see Appendix C of [41] and section 2.3 of [45]).

2 Standard Structural Risk Minimisation

As an initial example we consider a hierarchy of classes

$$H_1 \subseteq H_2 \subseteq \ldots \subseteq H_d \subseteq \ldots$$

where $H_i \subset \{0,1\}^X$ for some input space X, and where we will assume $\mathrm{VCdim}(H_d) = d$ for the rest of this section. (Recall that the VC-dimension of a class of $\{0,1\}$-valued functions is the size of the largest subset of their domain for which the restriction of the class to that subset is the set of all $\{0,1\}$-valued functions; see [49].) Such a hierarchy of classes is called a decomposable concept class by Linial *et al.* [29]. Related work is presented by Benedek and Itai [12]. We will assume that a fixed number m of labelled examples are given as a vector $\boldsymbol{z} = (\boldsymbol{x}, t(\boldsymbol{x}))$ to the learner, where $\boldsymbol{x} = (x_1, \ldots, x_m)$, and $t(\boldsymbol{x}) = (t(x_1), \ldots, t(x_m))$, and that the target function t lies in one of the subclasses H_d. The learner uses an algorithm to find a value of d which contains an hypothesis h that is consistent with the sample $\boldsymbol{z}$. What we require is a function $\epsilon(m, d, \delta)$ which will give the learner an upper bound on the generalization error of h with confidence $1 - \delta$. The following theorem gives a suitable function. We use $\mathrm{Er}_{\boldsymbol{z}}(h) = |\{i : h(x_i) \neq t(x_i)\}|$ to denote the *number* of errors that h makes on $\boldsymbol{z}$, and $\mathrm{er}_P(h) = P\{\boldsymbol{x}\colon h(\boldsymbol{x}) \neq t(\boldsymbol{x})\}$ to denote the *expected error* when $x_1, \ldots, x_m$ are drawn independently according to P. In what follows we will often write $\mathrm{Er}_{\boldsymbol{x}}(h)$ (rather than $\mathrm{Er}_{\boldsymbol{z}}(h)$) when the target t is obvious from the context. The following theorem, which appears in [43], covers the case where there are no errors on the training set. It is a well-known result which we quote for completeness.

Theorem 2.1 *[43] Let H_i, $i = 1, 2, \ldots$ be a sequence of hypothesis classes mapping X to $\{0,1\}$ such that $\mathrm{VCdim}(H_i) = i$, and let P be a probability distribution on X. Let p_d be any set of positive numbers satisfying $\sum_{d=1}^{\infty} p_d = 1$. With probability $1 - \delta$ over m independent examples drawn according to P, for any d for which a learner finds a consistent hypothesis h in H_d, the generalization error of h is bounded from above by*

$$\epsilon(m, d, \delta) = \frac{4}{m}\left(d \ln\left(\frac{2em}{d}\right) + \ln\left(\frac{1}{p_d}\right) + \ln\left(\frac{4}{\delta}\right)\right),$$

provided $d \leq m$.

The role of the numbers p_d may seem a little counter-intuitive as we appear to be able to bias our estimate by adjusting these parameters. The numbers must, however, be specified in advance and represent some apportionment of our confidence to the different points where failure might occur. In this sense they should be one of the arguments of the function $\epsilon(m, d, \delta)$. We have deliberately omitted this dependence as they have a different status in the learning framework. It is helpful to think of p_d as our prior estimate of the probability that the smallest class containing a consistent hypothesis is H_d. In particular we can set $p_d = 0$ for $d > m$, since we would expect to be able to find a consistent hypothesis in H_m and if we fail the bound will not be useful for such large d in any case.

We also wish to consider the possibility of errors on the training sample. The result presented here is analagous to those obtained by Lugosi and Zeger [37] in the statistical framework.

We will make use of the following result of Vapnik in a slightly improved version due to Anthony and Shawe-Taylor [4]. Note also that the result is expressed in terms of the quantity

$\mathrm{Er}_{\boldsymbol{z}}(h)$ which denotes the number of errors of the hypothesis h on the sample $\boldsymbol{z}$, rather than the usual proportion of errors.

Theorem 2.2 ([4]) *Let $0 < \epsilon < 1$ and $0 < \gamma \leq 1$. Suppose H is an hypothesis space of functions from an input space X to $\{0,1\}$, and let μ be any probability measure on $S = X \times \{0,1\}$. Then the probability (with respect to μ^m) that for $\boldsymbol{z} \in S^m$, there is some $h \in H$ such that*

$$\mathrm{er}_\mu(h) > \epsilon \quad \text{and} \quad \mathrm{Er}_{\boldsymbol{z}}(h) \leq m(1-\gamma)\mathrm{er}_\mu(h)$$

is at most

$$4\,\Pi_H(2m)\exp\left(-\frac{\gamma^2\epsilon m}{4}\right).$$

Our aim will be to use a double stratification of δ; as before by class (via p_d), and also by the number of errors on the sample (via q_{dk}). The generalization error will be given as a function of the size of the sample m, index of the class d, the number of errors on the sample k, and the confidence δ.

Theorem 2.3 *Let H_i, $i = 1, 2, \ldots$, be a sequence of hypothesis classes mapping X to $\{0,1\}$ and having VC-dimension i. Let μ be any probability measure on $S = X \times \{0,1\}$, and let p_d, q_{dk} be any sets of positive numbers satisfying*

$$\sum_{d=1}^{\infty} p_d = 1,$$

and $\sum_{k=0}^{m} q_{dk} = 1$ for all d. Then with probability $1-\delta$ over m independent identically distributed examples $\boldsymbol{x}$, if the learner finds an hypothesis h in H_d with $\mathrm{Er}_{\boldsymbol{x}}(h) = k$, then the generalization error of h is bounded from above by

$$\epsilon(m,d,k,\delta) = \frac{1}{m}\left(2k + 4\ln\left(\frac{4}{p_d q_{dk}\delta}\right) + 4d\ln\left(\frac{2em}{d}\right)\right),$$

provided $d \leq m$.

Proof : We bound the required probability of failure

$$\mu^m\{\boldsymbol{z} : \exists d, k,\ \exists h \in H_d, \mathrm{Er}_{\boldsymbol{z}}(h) = k, \mathrm{er}_\mu(h) > \epsilon(m,d,k,\delta)\} < \delta,$$

by showing that for all d and k

$$\mu^m\{\boldsymbol{z} : \exists h \in H_d, \mathrm{Er}_{\boldsymbol{z}}(h) = k, \mathrm{er}_\mu(h) > \epsilon(m,d,k,\delta)\} < \delta p_d q_{dk}.$$

We will apply Theorem 2.2 once for each value of k and d. We must therefore ensure that only one value of $\gamma = \gamma_{dk}$ is used in each case. An appropriate value is

$$\gamma_{dk} = 1 - \frac{k}{m\epsilon(m,d,k,\delta)}.$$

This ensures that if $\mathrm{er}_\mu(h) > \epsilon(m,d,k,\delta)$ and $\mathrm{Er}_{\boldsymbol{z}}(h) = k$, then

$$\mathrm{Er}_{\boldsymbol{z}}(h) = k = m(1-\gamma_{dk})\epsilon(m,d,k,\delta) \leq m(1-\gamma_{dk})\mathrm{er}_\mu(h),$$

as required for an application of the theorem. Hence, if $m \geq d$ Sauer's lemma implies

$$\begin{aligned}
& \mu^m\{\boldsymbol{z} : \exists h \in H_d, \mathrm{Er}_{\boldsymbol{z}}(h) = k, \mathrm{er}_\mu(h) > \epsilon(m,d,k,\delta)\} < \delta p_d q_{dk} \\
\Leftarrow \quad & 4\left(\frac{2em}{d}\right)^d \exp\left(\frac{-\gamma_{dk}^2 \epsilon(m,d,k,\delta)m}{4}\right) < \delta p_d q_{dk} \\
\Leftarrow \quad & d\ln\left(\frac{2em}{d}\right) + \ln\left(\frac{4}{p_d q_{dk}\delta}\right) - \frac{\epsilon(m,d,k,\delta)m}{4} + \frac{2k}{4} = 0 \\
\Leftarrow \quad & \epsilon(m,d,k,\delta) = \frac{1}{m}\left(2k + 4\ln\left(\frac{4}{p_d q_{dk}\delta}\right) + 4d\ln\left(\frac{2em}{d}\right)\right),
\end{aligned}$$

ignoring one term of $k^2/(4m\epsilon)$. The result follows. ∎

The choice of the prior q_{dk} for different k will again affect the resulting trade-off between complexity and accuracy. In view of our expectation that the penalty term for choosing a large class is probably an overestimate, it seems reasonable to give a correspondingly large penalty for a large numbers of errors. One possibility is an exponentially decreasing prior distribution such as

$$q_{dk} = 2^{-(k+1)},$$

though the rate of decrease could also be varied between classes. Assuming the above choice, observe that an incremental search for the optimal value of d would stop when the reduction in the number of classification errors in the next class was less than

$$0.84\ln\left(\frac{2em}{d}\right).$$

Note that the tradeoff between errors on the sample and generalization error is also discussed in [16].

3 Classifiers with a Large Margin

The standard methods of structural risk minimization require that the decomposition of the hypothesis class be chosen in advance of seeing the data. In this section we introduce our first variant of SRM which effectively makes a decomposition after the data has been seen. The main tool we use is the fat-shattering dimension, which was introduced in [26], and has been used for several problems in learning since [1, 11, 2, 10]. We show that if a classifier correctly classifies a training set with a large margin, and if its fat-shattering function at a scale related to this margin is small, then the generalization error will be small. (This is formally stated in Theorem 3.9 below.)

Definition 3.1 *Let $\mathcal{F}$ be a set of real valued functions. We say that a set of points X is* γ-shattered by *$\mathcal{F}$ if there are real numbers r_x indexed by $x \in X$ such that for all binary vectors b*

indexed by X, there is a function $f_b \in \mathcal{F}$ satisfying

$$f_b(x) \begin{cases} \geq r_x + \gamma & \text{if } b_x = 1 \\ \leq r_x - \gamma & \text{otherwise.} \end{cases}$$

The fat shattering dimension $\text{fat}_{\mathcal{F}}$ *of the set $\mathcal{F}$ is a function from the positive real numbers to the integers which maps a value γ to the size of the largest γ-shattered set, if this is finite or infinity otherwise.*

Let T_θ denote the threshold function at θ, $T_\theta : \mathbb{R} \to \{0,1\}$, $T_\theta(\alpha) = 1$ iff $\alpha > \theta$. Fix a class of $[0,1]$-valued functions. We can interpret each function f in the class as a classification function by considering the thresholded version, $T_{1/2} \circ f$. The following result implies that, if a real-valued function in the class maps all training examples to the correct side of $1/2$ by a large margin, the misclassification probability of the thresholded version of the function depends on the fat-shattering dimension of the class, at a scale related to the margin. This result is a special case of Corollary 6 in [2], which applied more generally to arbitrary real-valued target functions. (This application to classification problems was not described in [2].)

Theorem 3.2 *Let H be a set of $[0,1]$-valued functions defined on a set X. Let $0 < \gamma < 1/2$. There is a positive constant K such that, for any function $t : X \to \{0,1\}$ and any probability distribution P on X, with probability at least $1 - \delta$ over a sequence $x_1, \ldots, x_m$ of examples chosen independently according to P, every h in H that has*

$$|h(x_i) - t(x_i)| < 1/2 - \gamma$$

for $i = 1, \ldots, m$ satisfies

$$\Pr\left(|h(x) - t(x)| \geq 1/2\right) < \epsilon,$$

provided that

$$m \geq \frac{K}{\epsilon}\left(\log\frac{1}{\delta} + d\log^2\left(\frac{d}{\gamma\epsilon}\right)\right),$$

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

Clearly, this implies that the misclassification probability is less than ϵ under the conditions of the theorem, since $T_{1/2}(h(x)) \neq t(x)$ implies $|h(x) - t(x)| \geq 1/2$. In the remainder of this section, we present an improvement of this result. By taking advantage of the fact that the target values fall in the finite set $\{0,1\}$, and the fact that only the behaviour near the threshold of functions in H is important, we can remove the d/ϵ factor from the $\log^2$ factor in the bound. We also improve the constants that would be obtained from the argument used in the proof of Theorem 3.2.

Before we can quote the next lemma, we need another definition.

Definition 3.3 *Let (X,d) be a (pseudo-) metric space, let A be a subset of X and $\epsilon > 0$. A set $B \subseteq X$ is an* ϵ-cover *for A if, for every $a \in A$, there exists $b \in B$ such that $d(a,b) < \epsilon$. The* ϵ-covering number *of A, $\mathcal{N}_d(\epsilon, A)$, is the minimal cardinality of an ϵ-cover for A (if there is no such finite cover then it is defined to be ∞).*

The idea is that B should be finite but approximate all of A with respect to the pseudometric d. As in [2], we will use the l^∞ distance over a finite sample $\boldsymbol{x} = (x_1, \ldots, x_m)$ for the pseudo-metric in the space of functions,

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

We write $\mathcal{N}(\epsilon, \mathcal{F}, \boldsymbol{x})$ for the ϵ-covering number of $\mathcal{F}$ with respect to the pseudo-metric $d_{\boldsymbol{x}}$.

We now quote a lemma from Alon *et al.* [1] which we will use below.

Lemma 3.4 (Alon *et al.* [1]) *Let $\mathcal{F}$ be a class of functions $X \to [0, 1]$ and P a distribution over X. Choose $0 < \epsilon < 1$ and let $d = \text{fat}_{\mathcal{F}}(\epsilon/4)$. Then*

$$E\left(\mathcal{N}(\epsilon, \mathcal{F}, \boldsymbol{x})\right) \leq 2\left(\frac{4m}{\epsilon^2}\right)^{d\log(2em/(d\epsilon))},$$

where the expectation E is taken w.r.t. a sample $\boldsymbol{x} \in X^m$ drawn according to P^m.

Corollary 3.5 *Let $\mathcal{F}$ be a class of functions $X \to [a, b]$ and P a distribution over X. Choose $0 < \epsilon < 1$ and let $d = \text{fat}_{\mathcal{F}}(\epsilon/4)$. Then*

$$E\left(\mathcal{N}(\epsilon, \mathcal{F}, \boldsymbol{x})\right) \leq 2\left(\frac{4m(b-a)^2}{\epsilon^2}\right)^{d\log(2em(b-a)/(d\epsilon))},$$

where the expectation E is over samples $\boldsymbol{x} \in X^m$ drawn according to P^m.

Proof: We first scale all the functions in $\mathcal{F}$ by the affine transformation mapping the interval $[a, b]$ to $[0, 1]$ to create the set of functions $\mathcal{F}'$. Clearly, $\text{fat}_{\mathcal{F}'}(\gamma) = \text{fat}_{\mathcal{F}}(\gamma(b-a))$, while $E(\mathcal{N}(\epsilon, \mathcal{F}, \boldsymbol{x})) = E(\mathcal{N}(\epsilon/(b-a), \mathcal{F}', \boldsymbol{x}))$. The result follows. ∎

In order to motivate the next lemma we first introduce some notation we will use when we come to apply it. The aim is to transform the problem of observing a large margin into one of observing the maximal value taken by a set of functions. We do this by 'folding' over the functions at the threshold. The following hat operator implements the folding.

We define the mapping $\hat{}: \mathbb{R}^X \to \mathbb{R}^{X\times\{0,1\}}$ by

$$\hat{}: f \mapsto \hat{f}(x, c) = f(x)(1 - c) + (2\theta - f(x))c,$$

for some fixed real θ. For a set of functions $\mathcal{F}$, we define $\hat{\mathcal{F}} = \hat{\mathcal{F}}_\theta = \{\hat{f} : f \in \mathcal{F}\}$. The idea behind this mapping is that for a function f, the corresponding $\hat{f}$ maps the input x and it classification c to an output value, which will be less than θ provided the classification obtained by thresholding $f(x)$ at θ is correct.

Lemma 3.6 *Suppose $\mathcal{F}$ is a set of functions that map from X to $\mathbb{R}$ with finite fat-shattering dimension bounded by the function* afat : $\mathbb{R} \to \mathbb{N}$ *which is continuous from the right. Then for any distribution P on X, and any $k \in \mathbb{N}$ and any $\theta \in \mathbb{R}$*

$$P^{2m}\left\{\boldsymbol{xy}: \exists f \in \mathcal{F}, r = \max_j\{f(x_j)\}, 2\gamma = \theta - r, \text{afat}(\gamma/4) = k, \right.$$
$$\left. \frac{1}{m}\left|\{i : f(y_i) \geq r + 2\gamma\}\right| > \epsilon(m, k, \delta)\right\} < \delta,$$

where $\epsilon(m, k, \delta) = \frac{1}{m}(k \log \frac{8em}{k} \log(32m) + \log \frac{2}{\delta})$.

Proof: Using the standard permutation argument (as in [49]), we may fix a sequence $\boldsymbol{xy}$ and bound the probability under the uniform distribution on swapping permutations that the permuted sequence satisfies the condition stated. Let $\gamma_k = \min\{\gamma' : \text{afat}(\gamma'/4) \leq k\}$. Notice that the minimum is defined since afat is continuous from the right, and also that $\text{afat}(\gamma_k/4) = \text{afat}(\gamma/4)$. For any γ satisfying $\text{afat}(\gamma/4) = k$, we have $\gamma_k \leq \gamma$, so the probability above is no greater than

$$P^{2m}\left\{\boldsymbol{xy} : \exists \gamma \in \mathbb{R}^+, \text{afat}(\gamma/4) = k, \exists f \in \mathcal{F}, A_f(2\gamma_k)\right\},$$

where $A_f(\gamma)$ is the event that $f(y_i) \geq \max_j\{f(x_j)\} + \gamma$ for at least $m\epsilon(m,k,\delta)$ points y_i in $\boldsymbol{y}$. Note that $r + 2\gamma = \theta$. Let

$$\pi(\alpha) := \begin{cases} \theta & \text{if } \alpha > \theta \\ \theta - 2\gamma_k & \text{if } \alpha < \theta - 2\gamma_k \\ \alpha & \text{otherwise,} \end{cases}$$

and let $\pi(\mathcal{F}) = \{\pi(f) : f \in \mathcal{F}\}$. Consider a minimal γ_k-cover $B_{\boldsymbol{xy}}$ of $\pi(\mathcal{F})$ in the pseudo-metric $d_{\boldsymbol{xy}}$. We have that for any $f \in \mathcal{F}$, there exists $\tilde{f} \in B_{\boldsymbol{xy}}$, with $|\pi(f)(x) - \pi(\tilde{f})(x)| < \gamma_k$ for all $x \in \boldsymbol{xy}$. Thus since for all $x \in \boldsymbol{x}$, by the definition of r, $f(x) \leq r = \theta - 2\gamma$, $\pi(f)(x) = \theta - 2\gamma_k = r + 2(\gamma - \gamma_k)$, and so $\pi(\tilde{f})(x) < r + 2\gamma - \gamma_k$. However there are at least $m\,\epsilon(m,k,\delta)$ points $y \in \boldsymbol{y}$ such that $f(y) \geq \theta = r + 2\gamma$, so $\pi(\tilde{f})(y) > r + 2\gamma - \gamma_k > \max_j\{\pi(\tilde{f})(x_j)\}$. Since π only reduces separation between output values, we conclude that the event $A_{\tilde{f}}(0)$ occurs. By the permutation argument, for fixed $\tilde{f}$ at most $2^{-\epsilon(m,k,\delta)m}$ of the sequences obtained by swapping corresponding points satisfy the conditions, since the ϵm points with the largest $\tilde{f}$ values must remain on the right hand side for $A_{\tilde{f}}(0)$ to occur. Thus by the union bound

$$P^{2m}\left\{\boldsymbol{xy} : \exists \gamma \in \mathbb{R}^+, \text{afat}(\gamma/4) = k, \exists f \in \mathcal{F}, A_f(2\gamma_k)\right\} \leq E(|B_{\boldsymbol{xy}}|)2^{-\epsilon(m,k,\delta)m},$$

where the expectation is over $\boldsymbol{xy}$ drawn according to P^{2m}. Now for all $\gamma > 0$, $\text{fat}_{\pi(\mathcal{F})}(\gamma) \leq \text{fat}_{\mathcal{F}}(\gamma)$ since every set of points γ-shattered by $\pi(\mathcal{F})$ can be γ-shattered by $\mathcal{F}$. Furthermore, $\pi(\mathcal{F})$ is a class of functions mapping a set X to the interval $[\theta - 2\gamma_k, \theta]$. Hence, by Corollary 3.5 (setting $[a,b]$ to $[\theta - 2\gamma_k, \theta]$, ϵ to γ_k, and m to $2m$),

$$E(|B_{\boldsymbol{xy}}|) = E(\mathcal{N}(\gamma_k, \pi(\mathcal{F}), \boldsymbol{xy})) \leq 2\left(\frac{8m(\theta - \theta + 2\gamma_k)^2}{\gamma_k^2}\right)^{d\log(4em(2\gamma_k)/(d\gamma_k))},$$

where $d = \text{fat}_{\pi(\mathcal{F})}(\gamma_k/4) \leq \text{fat}_{\mathcal{F}}(\gamma_k/4) \leq k$. Thus

$$E(|B_{\boldsymbol{xy}}|) \leq 2(32m)^{k\log(8em/k)},$$

and so $E(|B_{\boldsymbol{xy}}|)2^{-\epsilon(m,k,\delta)m} < \delta$ provided

$$\epsilon(m,k,\delta) \geq \tfrac{1}{m}\left(k\log(8em/k)\log(32m) + \log\tfrac{2}{\delta}\right),$$

as required. ∎

The function afat(γ) is used in this theorem rather than fat$_{\mathcal{F}}(\gamma)$ since we used the continuity property to ensure that afat$(\gamma_k/4) = k$ for every k, while we cannot assume that fat$_{\mathcal{F}}(\gamma)$ is continuous from the right. We could avoid this requirement and give an error estimate directly in terms of fat$_{\mathcal{F}}$ instead of afat, but this would introduce a worse constant in the argument of fat$_{\mathcal{F}}$. Since in practice one works with continuous upper bounds on fat$_{\mathcal{F}}(\gamma)$ (e.g. c/γ^2) by taking the floor of the value, the critical question becomes whether the bound is strict rather than less than or equal. Provided fat$_{\mathcal{F}}$ is strictly less than the continuous bound the corresponding floor function is continuous from the right. If not addition of an arbitrarily small constant to the continuous function will allow substitution of a strict inequality.

Lemma 3.7 *Let $\mathcal{F}$ be a set of real valued functions from X to $\mathbb{R}$. Then for all $\gamma \geq 0$,*

$$\text{fat}_{\hat{\mathcal{F}}}(\gamma) = \text{fat}_{\mathcal{F}}(\gamma).$$

Proof: For any $c \in \{0,1\}^m$, we have that f_b realises dichotomy b on $\boldsymbol{x} = (x_1, \ldots, x_m)$ with margin γ about output values r_i if and only if $\hat{f}_b$ realises dichotomy $b \oplus c$ on

$$\hat{\boldsymbol{x}} = ((x_1, c_1), \ldots, (x_m, c_m)),$$

with margin γ about output values

$$\hat{r_i} = r_i(1 - c_i) + (2\theta - r_i)c_i.$$

■

We will make use of the following lemma, which in the form below is due to Vapnik [46, page 168].

Lemma 3.8 *Let X be a set and S a system of sets on X, and P a probability measure on X. For $\boldsymbol{x} \in X^m$ and $A \in S$, define $\nu_{\boldsymbol{x}}(A) := |\boldsymbol{x} \cap A|/m$. If $m > 2/\epsilon$, then*

$$P^m\left\{\boldsymbol{x}: \sup_{A\in S} |\nu_{\boldsymbol{x}}(A) - P(A)| > \epsilon\right\} \leq 2P^{2m}\left\{\boldsymbol{xy}: \sup_{A\in S} |\nu_{\boldsymbol{x}}(A) - \nu_{\boldsymbol{y}}(A)| > \epsilon/2\right\}.$$

Let T_θ denote the threshold function at θ: $T_\theta: \mathbb{R} \to \{0,1\}$, $T_\theta(\alpha) = 1$ iff $\alpha > \theta$. For a class of functions $\mathcal{F}$, $T_\theta(\mathcal{F}) = \{T_\theta(f): f \in \mathcal{F}\}$.

Theorem 3.9 *Consider a real valued function class $\mathcal{F}$ having fat shattering function bounded above by the function* afat $: \mathbb{R} \to \mathbb{N}$ *which is continuous from the right. Fix $\theta \in \mathbb{R}$. If a learner correctly classifies m independently generated examples $\boldsymbol{z}$ with $h = T_\theta(f) \in T_\theta(\mathcal{F})$ such that* er$_{\boldsymbol{z}}(h) = 0$ *and $\gamma = \min|f(x_i) - \theta|$, then with confidence $1 - \delta$ the expected error of h is bounded from above by*

$$\epsilon(m, k, \delta) = \frac{2}{m}\left(k \log\left(\frac{8em}{k}\right)\log(32m) + \log\left(\frac{8m}{\delta}\right)\right),$$

where $k =$ afat$(\gamma/8)$.

Proof: The proof will make use of lemma 3.8. First we will move to a double sample and stratify by k. By the union bound, it thus suffices to show that

$$P^{2m}(\bigcup_{k=1}^{2m} J_k) \leq \sum_{k=1}^{2m} P^{2m}(J_k) < \delta/2,$$

where

$$\begin{aligned} J_k = \{\boldsymbol{xy} \quad : \quad & \exists h = T_\theta(f) \in T_\theta(\mathcal{F}), \mathrm{Er}_{\boldsymbol{x}}(h) = 0, k = \mathrm{afat}(\gamma/8), \\ & \gamma = \min |f(x_i) - \theta|, \mathrm{Er}_{\boldsymbol{y}}(h) \geq m\epsilon(m,k,\delta)/2\}. \end{aligned}$$

(The largest value of k we need consider is $2m$, since we cannot shatter a greater number of points from $\boldsymbol{xy}$.) It is sufficient if

$$P^{2m}(J_k) \leq \frac{\delta}{4m} = \delta'.$$

Consider $\hat{\mathcal{F}} = \hat{\mathcal{F}}_\theta$ and note that by Lemma 3.7 the function $\mathrm{afat}(\gamma)$ also bounds $\mathrm{fat}_{\hat{\mathcal{F}}}(\gamma)$. The probability distribution on $\hat{X} = X \times \{0,1\}$ is given by P on X with the second component determined by the target value of the first component. Note that for a point $y \in \boldsymbol{y}$ to be misclassified, it must have

$$\hat{f}(\hat{y}) \geq \max\{\hat{f}(\hat{x}) \colon \hat{x} \in \hat{\boldsymbol{x}}\} + \gamma = \theta,$$

so that

$$\begin{aligned} J_k \subseteq \Big\{\hat{\boldsymbol{x}}\hat{\boldsymbol{y}} \in (X \times \{0,1\})^{2m} \quad : \quad & \exists \hat{f} \in \hat{\mathcal{F}}, r = \max\{\hat{f}(\hat{x}) \colon \hat{x} \in \hat{\boldsymbol{x}}\}, \gamma = \theta - r, \\ & k = \mathrm{afat}(\gamma/8), \left|\{\hat{y} \in \hat{\boldsymbol{y}} \colon \hat{f}(\hat{y}) \geq \theta\}\right| \geq m\epsilon(m,k,\delta)/2\Big\} \end{aligned}$$

Replacing γ by $\gamma/2$ in Lemma 3.6 we obtain

$$P^{2m}(J_k) \leq \delta', \quad \text{for}$$

$$\epsilon(m,k,\delta) = \frac{2}{m}\left(k\log(8em/k)\log(32m) + \log(2/\delta')\right).$$

With this linking of ϵ and m, the condition of Lemma 3.8 is satisfied. Appealling to this and noting that the union bound gives $P^{2m}(\bigcup_{k=1}^{2m} J_k) \leq \sum_{k=1}^{2m} P^{2m}(J_k)$ we conclude the proof by substituting for δ'. ∎

A related result, that gives bounds on the misclassification probability of thresholded functions in terms of an error estimate involving the margin of the corresponding real-valued functions, is given in [9]. Using this result and bounds on the fat-shattering dimension of sigmoidal neural networks, that paper also gives bounds on the generalization performance of these networks that depend on the size of the parameters but are independent of the number of parameters.

4 Large Margin Hyperplanes

We will now consider a particular case of the results in the previous section, applicable to the class of of linear threshold functions in Euclidean space. Vapnik and others [46, 48, 14, 16], [18, page 140] have suggested that choosing the maximal margin hyperplane (i.e. the hyperplane which maximises the minimal distance of points — assuming a correct classification can be made) will improve the generalization of the resulting classifier. They give evidence to indicate that the generalization performance is frequently significantly better than that predicted by the VC dimension of the full class of linear threshold functions. In this section of the paper we will show that indeed a large margin does help in this case, and we will give an explicit bound on the generalization error in terms of the margin achieved on the training sample. We do this by first bounding the appropriate fat-shattering function, and then applying theorem 3.9.

The margin also arises in the proof of the perceptron convergence theorem (see for example [23, page 61–62], where an alternate motivation is given for a large margin: noise immunity). The margin occurs even more explictly in the Winnow algorithms and their variants developed by Littlestone and others [30, 31, 32]. The connection between these two uses has not yet been explored.

Consider a hyperplane defined by (w, θ), where w is a weight vector and θ a threshold value. Let X_0 be a subset of the Euclidean space that does not have a limit point on the hyperplane, so that

$$\min_{x \in X_0} |\langle x, w \rangle + \theta| > 0.$$

We say that the hyperplane is in *canonical form* with respect to X_0 if

$$\min_{x \in X_0} |\langle x, w \rangle + \theta| = 1.$$

Let $\|\cdot\|$ denote the Euclidean norm. The maximal margin hyperplane is obtained by minimising $\|w\|$ subject to these constraints. The points in X_0 for which the minimum is attained are called the support vectors of the maximal margin hyperplane.

The following theorem is the basis for our argument for the maximal margin analysis.

Theorem 4.1 (Vapnik [48]) *Suppose X_0 is a subset of the input space contained in a ball of radius R about some point. Consider the set of hyperplanes in canonical form with respect to X_0 that satisfy $\|w\| \leq A$, and let $\mathcal{F}$ be the class of corresponding linear threshold functions,*

$$f(x, w) = \text{sgn}(\langle x, w \rangle + \theta).$$

Then the restriction of $\mathcal{F}$ to the points in X_0 has VC dimension bounded by

$$\min\{R^2 A^2, n\} + 1.$$

Our argument will also be in terms of Theorem 3.9, and to that end we need to bound the fat-shattering dimension of the class of hyperplanes. We do this via an argument concerning the level fat-shattering dimension, defined below.

Definition 4.2 *Let $\mathcal{F}$ be a set of real valued functions. We say that a set of points X is* level γ-shattered by $\mathcal{F}$ at level r *if it can be γ-shattered when choosing the $r_x = r$ for all $x \in X$. The*

level fat shattering dimension $\text{lfat}_{\mathcal{F}}$ *of the set $\mathcal{F}$ is a function from the positive real numbers to the integers which maps a value γ to the size of the largest level γ-shattered set, if this is finite or infinity otherwise.*

The level fat-shattering dimension is a scale sensitive version of a dimension introduced by Vapnik [46]. The scale sensitive version was first introduced by Alon *et al.* [1].

Lemma 4.3 *Let $\mathcal{F}$ be the set of linear functions with unit weight vectors, restricted to points in a ball of radius R,*

$$\mathcal{F} = \{x \mapsto \langle w, x\rangle + \theta : \|w\| = 1\}. \tag{1}$$

Then the level fat shattering function can be bounded from above by

$$\text{lfat}_{\mathcal{F}}(\gamma) \leq \min\{R^2/\gamma^2, n\} + 1.$$

Proof: If a set of points $X = \{x^i\}_i$ is to be level γ-shattered there must be a value r such that each dichotomy b can be realised with a weight vector w^b and threshold θ^b such that

$$\langle w^b, x^i\rangle + \theta^b \begin{cases} \geq r + \gamma & \text{if } b_i = 1 \\ \leq r - \gamma & \text{otherwise.} \end{cases}$$

Let $d = \min_{x\in X} |\langle w^b, x\rangle + \theta_b - r| \geq \gamma$. Consider the hyperplane defined by $(\overline{w}^b, \overline{\theta}^b) = (w^b/d, \theta^b/d - r/d)$. It is in canonical form with respect to the points X, satisfies $\|\overline{w}^b\| = \|w^b/d\| = 1/d$ and realises dichotomy b on X. Hence, the set of points X can be shattered by a subset of canonical hyperplanes $(\overline{w}^b, \overline{\theta}^b)$ satisfying $\|\overline{w}^b\| \leq 1/d \leq 1/\gamma$. The result follows from Theorem 4.1. ∎

Corollary 4.4 *Let $\mathcal{F}$ be the set, defined in (1), of linear functions with unit weight vectors, restricted to points in a ball of n dimensions of radius R about the origin and with thresholds $|\theta| \leq R$. The fat shattering function of $\mathcal{F}$ can be bounded by*

$$\text{fat}_{\mathcal{F}}(\gamma) \leq \min\{9R^2/\gamma^2, n+1\} + 1.$$

Proof: Suppose m points $x^1, \ldots, x^m$ lying in a ball of radius R about the origin are γ-shattered relative to $r = (r_1, \ldots, r_m)$. Since $\|w\| = 1$, $|\langle w, x^i\rangle + \theta| \leq 2R$, and so $|r_i| \leq 2R$. From each x^i, $i = 1, \ldots, m$, we create an extended vector $\overline{x}^i := (x^i_1, \ldots, x^i_n, r_i/\sqrt{2})$. Since $|r_i| \leq 2R$, $\|\overline{x}^i\| \leq \sqrt{3}R$. Let (w^b, θ^b) be the parameter vector of the hyperplane that realizes a dichotomy $b \in \{0,1\}^m$. Set $\overline{w}^b = (w^b_1, \ldots, w^b_n, -\sqrt{2})$.

We now show that the points $\overline{x}^i$, $i = 1, \ldots, m$ are level γ-shattered at level 0 by $\{\overline{w}^b\}_{b\in\{0,1\}^m}$. We have that $\langle \overline{w}^b, \overline{x}^i\rangle + \theta^b = \langle w^b, x^i\rangle + \theta^b - r_i =: t$. But $\langle w^b, x^i\rangle + \theta^b \geq r_i + \gamma$ if $b_i = 1$, and $\langle w^b, x^i\rangle + \theta^b \leq r_i - \gamma$ if $b_i = 0$. Thus

$$\begin{aligned} t &\geq r_i + \gamma - r_i = \gamma \quad \text{if } b_i = 1 \\ t &\leq r_i - \gamma - r_i = \gamma \quad \text{if } b_i = 0. \end{aligned}$$

Now $\|\overline{w}^b\| = \sqrt{3}$. Set $\tilde{w}^b = \overline{w}^b/\sqrt{3}$, and $\tilde{x}^i = \sqrt{3}\overline{x}^i$. Then $\|\tilde{w}^b\| = 1$ and the points $\tilde{x}^i$, $i = 1, \ldots, m$ are level γ-shattered at level 0 by $\{\tilde{w}^b\}_{b\in\{0,1\}^m}$. Since $\dim \tilde{x}^i = n + 1$ and $\|\tilde{x}^i\| \leq \sqrt{3}\sqrt{3}R = 3R$, we have by Lemma 4.3 that $\text{fat}_{\mathcal{F}}(\gamma) \leq \min\{\frac{9R^2}{\gamma^2}, n+1\} + 1$. ∎

Theorem 4.5 *Suppose inputs are drawn independently according to a distribution whose support is contained in a ball in $\mathbb{R}^n$ centered at the origin, of radius R. If we succeed in correctly classifying m such inputs by a canonical hyperplane with $\|w\| = 1/\gamma$ and with $|\theta| \leq R$, then with confidence $1 - \delta$ the generalization error will be bounded from above by*

$$\epsilon(m, \gamma) = \frac{2}{m}\left(k \log\left(\frac{8em}{k}\right)\log(32m) + \log\frac{8m}{\delta}\right),$$

where $k = \lfloor 577R^2/\gamma^2 \rfloor$.

Proof : Firstly note that we can restrict our consideration to the subclass of $\mathcal{F}$ with $|\theta| \leq R$. If there is more than one point to be γ-shattered, then it is required to achieve a dichotomy with different signs; that is b is neither all 0s nor all 1s. Since all of the points lie in the ball, to shatter them the hyperplane must intersect the ball. Since $\|w\| = 1$, that means $|\theta| \leq R$. So although one may achieve a greater margin for the all-zero or all-one dichotomy by choosing a larger value of θ, all of the other dichotomies cannot achieve a larger γ. Thus although the bound may be weak in the special case of an all 0 or all 1 classification on the training set, it will still be true.

Hence, we are now in a position to apply Theorem 3.9 with the value of θ given in the theorem taken as 0. Hence,

$$\text{fat}_{\mathcal{F}}(\gamma/8) \leq \lfloor 576R^2/\gamma^2 + 1 \rfloor < \lfloor 577R^2/\gamma^2 \rfloor,$$

since $\gamma < R$. Substituting into the bound of Theorem 3.9 gives the required bound. ∎

In section 6 we will give an analogous result as a special case of the more general framework derived in section 5. Although the sample size bound for that result is weaker (by an additional $\log(m)$ factor), it does allow one to cope with the slightly more general situation of estimating the radius of the ball rather than knowing it in advance.

The fact that the bound in Theorem 4.5 does not depend on the dimension of the input space is particularly important in the light of Vapnik's ingenious construction of his support-vector machines [16, 48]. This is a method of implementing quite complex decision rules (such as those defined by polynomials or neural networks) in terms of linear hyperplanes in very many dimensions. The clever part of the technique is the algorithm which can work in a dual space, and which maximizes the margin on a training set. Thus Vapnik's algorithm along with the bound of Theorem 4.5 should allow good *a posteriori* bounds on the generalization error in a range of applications.

It is important to note that our explanation of the good performance of maximum margin hyperplanes is different to that given by Vapnik in [48, page 135]. Whilst alluding to the result of theorem 4.1, the theorem he presents as the explanation is a bound on the expected generalization error in terms of the number of support vectors. A small number of support vectors gives a good bound. One can construct examples in which all four combinations of small/large margin and few/many support vectors occur. Thus neither explanation is the only one. In the terminology of the next section, the margin and (the reciprocal of) the number of support vectors are both "luckiness" functions, and either could be used to determine bounds on performance.

5 Luckiness: A General Framework for Decomposing Classes

The standard PAC analysis gives bounds on generalization error that are uniform over the hypothesis class. Decomposing the hypothesis class, as described in Section 2, allows us to bias our generalization error bounds in favour of certain target functions and distributions: those for which some hypothesis low in the hierarchy is an accurate approximation. The results of section 4 show that it is possible to decompose the hypothesis class on the basis of the observed data in some cases: there we did it in terms of the margin attained. In this section, we introduce a more general framework which subsumes the standard PAC model, the framework described in Section 2 and can recover (in a slightly weaker form) the results of Section 4 as a special case. This more general decomposition of the hypothesis class based on the sample allows us to bias our generalization error bounds in favour of more general classes of target functions and distributions, which might correspond to more realistic assumptions about practical learning problems.

It seems that in order to allow the decomposition of the hypothesis class to depend on the sample, we need to make better use of the information provided by the sample. Both the standard PAC analysis and structural risk minimisation with a fixed decomposition of the hypothesis class effectively discard the training examples, and only make use of the function $\mathrm{Er}_{\boldsymbol{z}}$ defined on the hypothesis class that is induced by the training examples. The additional information we exploit in the case of sample-based decompositions of the hypothesis class is encapsulated in a *luckiness function*.

The main idea is to fix in advance some assumption about the target function and distribution, and encode this assumption in a real-valued function defined on the space of training samples and hypotheses. The value of the function indicates the extent to which the assumption is satisfied for that sample and hypothesis. We call this mapping a luckiness function, since it reflects how fortunate we are that our assumption is satisfied. That is, we make use of a function

$$L : X^m \times H \to \mathbb{R}^+,$$

which measures the luckiness of a particular hypothesis with respect to the training examples. Sometimes it is convenient to express this relationship in an inverted way, as an unluckiness function,

$$U : X^m \times H \to \mathbb{R}^+.$$

It turns out that only the ordering that the luckiness or unluckiness functions impose on hypotheses is important. We define the *level* of a function $h \in H$ relative to L and $\boldsymbol{x}$ by the function

$$\ell(\boldsymbol{x}, h) = |\{b \in \{0,1\}^m : \exists g \in H, g(\boldsymbol{x}) = b, L(\boldsymbol{x}, g) \geq L(\boldsymbol{x}, h)\}|,$$

or

$$\ell(\boldsymbol{x}, h) = |\{b \in \{0,1\}^m : \exists g \in H, g(\boldsymbol{x}) = b, U(\boldsymbol{x}, g) \leq U(\boldsymbol{x}, h)\}|.$$

Whether $\ell(\boldsymbol{x}, h)$ is defined in terms of L or U is a matter of convenience; the quantity $\ell(\boldsymbol{x}, h)$ itself plays the central role in what follows. If $\boldsymbol{x}, \boldsymbol{y} \in X^m$, we denote by $\boldsymbol{xy}$ their concatenation $(x_1, \ldots, x_m, y_1, \ldots, y_m)$.

5.1 Examples

Example 5.1 *Consider the hierarchy of classes introduced in Section 2 and define*

$$U(\boldsymbol{x}, h) = \min\{d : h \in H_d\}.$$

Then it follows from Sauer's lemma that for any $\boldsymbol{x}$ we can bound $\ell(\boldsymbol{x}, h)$ by

$$\ell(\boldsymbol{x}, h) \leq \left(\frac{em}{d}\right)^d,$$

where $d = U(\boldsymbol{x}, h)$. Notice also that for any $\boldsymbol{y} \in X^m$,

$$\ell(\boldsymbol{x}\boldsymbol{y}, h) \leq \left(\frac{2em}{d}\right)^d.$$

The last observation is something that will prove useful later when we investigate how we can use luckiness on a sample to infer luckiness on a subsequent sample.

We show in Section 6 that the hyperplane margin of Section 5 is a luckiness function which satisfies the technical restrictions we introduce below. We do this in fact in terms of the following unluckiness function, defined formally here for convenience later on.

Definition 5.2 *If h is a linear threshold function with separating hyperplane defined by (w, θ), and (w, θ) is in canonical form with respect to an m-sample $\boldsymbol{x}$, then define*

$$U(\boldsymbol{x}, h) = \max_{1 \leq i \leq m} \|x_i\|^2 \|w\|^2.$$

Finally, we give a separate unluckiness function for the maximal margin hyperplane example. In practical experiments it is frequently observed that the number of support vectors is significantly smaller than the full training sample. Vapnik [48, Theorem 5.2] gives a bound on the expected generalization error in terms of the number of support vectors as well as giving examples of classifiers [48, Table 5.2] for which the number of support vectors was very much less than the number of training examples. We will call this unluckiness function the support vectors' unluckiness function.

Definition 5.3 *If h is a linear threshold function with separating hyperplane defined by (w, θ), and (w, θ) is the maximal margin hyperplane in canonical form with respect to an m-sample $\boldsymbol{x}$, then define*

$$U(\boldsymbol{x}, h) = |\{x \in \boldsymbol{x} : |\langle x, w\rangle + \theta| = 1\}|,$$

that is U is the number of support vectors of the hyperplane.

5.2 Probable Smoothness of Luckiness Functions

We now introduce a technical restriction on luckiness functions required for our theorem.

Definition 5.4 *An α-subsequence of a vector $\boldsymbol{x}$ is a vector $\boldsymbol{x}'$ obtained from $\boldsymbol{x}$ by deleting a fraction of at most α coordinates. We will also write $\boldsymbol{x}' \subseteq_\alpha \boldsymbol{x}$. For a partitioned vector $\boldsymbol{xy}$, we write $\boldsymbol{x}'\boldsymbol{y}' \subseteq_\alpha \boldsymbol{xy}$.*
A luckiness function $L(\boldsymbol{x}, h)$ defined on a function class H is probably smooth *with respect to functions $\eta(m, L, \delta)$ and $\phi(m, L, \delta)$, if, for all targets t in H and for every distribution P,*

$$P^{2m}\left\{\boldsymbol{xy} : \exists h \in H, \mathrm{Er}_{\boldsymbol{x}}(h) = 0, \forall \boldsymbol{x}'\boldsymbol{y}' \subseteq_\eta \boldsymbol{xy}, \ell(\boldsymbol{x}'\boldsymbol{y}', h) > \phi(m, L(\boldsymbol{x}, h), \delta)\right\} \leq \delta,$$

where $\eta = \eta(m, L(\boldsymbol{x}, h), \delta)$.

The definition for probably smooth unluckiness is identical except that L's are replaced by U's. The intuition behind this rather arcane definition is that it captures when the luckiness can be estimated from the first half of the sample with high confidence. In particular, we need to ensure that few dichotomies are luckier than h on the double sample. That is, for a probably smooth luckiness function, if an hypothesis h has luckiness L on the first m points, we know that, with high confidence, for most (at least a proportion $\eta(m, L, \delta)$) of the points in a double sample, the growth function for the class of functions that are at least as lucky as h is small (no more than $\phi(m, L, \delta)$).

Theorem 5.5 *Suppose p_d, $d = 1, \ldots, 2m$, are positive numbers satisfying $\sum_{i=1}^{2m} p_i = 1$, L is a luckiness function for a function class H that is probably smooth with respect to functions η and ϕ, $m \in \mathbb{N}$ and $0 < \delta < 1/2$. For any target function $t \in H$ and any distribution P, with probability $1 - \delta$ over m independent examples $\boldsymbol{x}$ chosen according to P, if for any $i \in \mathbb{N}$ a learner finds an hypothesis h in H with $\mathrm{Er}_{\boldsymbol{x}}(h) = 0$ and $\phi(m, L(\boldsymbol{x}, h), \delta) \leq 2^{i+1}$, then the generalization error of h satisfies $\mathrm{er}_P(h) \leq \epsilon(m, i, \delta)$ where*

$$\epsilon(m, i, \delta) = \frac{2}{m}\left(i + 1 + \log\frac{4}{p_i\delta}\right) + 4\eta(m,\ L(\boldsymbol{x}, h),\ p_i\delta/4)\log 4m.$$

Proof : By Lemma 3.8,

$$\begin{aligned} &P^m\left\{\boldsymbol{x} : \exists h \in H, \exists i \in \mathbb{N}, \mathrm{Er}_{\boldsymbol{x}}(h) = 0, \phi(m, L(\boldsymbol{x}, h), \delta) \leq 2^{i+1}, \mathrm{er}_P(h) > \epsilon(m, i, \delta)\right\} \\ &\leq 2P^{2m}\left\{\boldsymbol{xy} : \exists h \in H, \exists i \in \mathbb{N}, \mathrm{Er}_{\boldsymbol{x}}(h) = 0, \phi(m, L(\boldsymbol{x}, h), \delta) \leq 2^{i+1}, \mathrm{Er}_{\boldsymbol{y}}(h) > \tfrac{m}{2}\epsilon(m, i, \delta)\right\}, \end{aligned}$$

provided $m \geq 2/\epsilon(m, i, \delta)$, which follows from the definition of $\epsilon(m, i, \delta)$ and the fact that $\delta \leq 1/2$. Hence it suffices to show that $P^{2m}(J_i) \leq \delta_i = p_i\delta/2$ for each $i \in \mathbb{N}$, where J_i is the event

$$\left\{\boldsymbol{xy} : \exists h \in H, \mathrm{Er}_{\boldsymbol{x}}(h) = 0, \phi(m, L(\boldsymbol{x}, h), \delta) \leq 2^{i+1}, \mathrm{Er}_{\boldsymbol{y}}(h) \geq \tfrac{m}{2}\epsilon(m, i, \delta)\right\}.$$

Let S be the event

$$\left\{\boldsymbol{xy} : \exists h \in H, \mathrm{Er}_{\boldsymbol{x}}(h) = 0, \forall \boldsymbol{x}'\boldsymbol{y}' \subseteq_\eta \boldsymbol{xy}, \ell(\boldsymbol{x}'\boldsymbol{y}', h) > \phi(m, L(\boldsymbol{x}, h), \delta)\right\}$$

with $\eta = \eta(m, L_i, \delta_i/2)$. It follows that

$$\begin{aligned} P^{2m}(J_i) &= P^{2m}(J_i \cap S) + P^{2m}(J_i \cap \bar{S}) \\ &\leq \delta_i/2 + P^{2m}(J_i \cap \bar{S}). \end{aligned}$$

It suffices then to show that $P^{2m}(J_i \cap \bar{S}) \leq \delta_i/2$. But $J_i \cap \bar{S}$ is a subset of

$$\begin{aligned} R &= \{\boldsymbol{xy}\colon \exists h \in H, \mathrm{Er}_{\boldsymbol{x}}(h) = 0, \exists \boldsymbol{x'y'} \subseteq_\eta \boldsymbol{xy}, \\ &\qquad \ell(\boldsymbol{x'y'}, h) \leq 2^{i+1}, \mathrm{Er}_{\boldsymbol{y'}}(h) \geq \tfrac{m}{2}\epsilon(m,i,\delta) - (|\boldsymbol{y}| - |\boldsymbol{y'}|)\big\}, \end{aligned}$$

where $|\boldsymbol{y'}|$ denotes the length of the sequence $\boldsymbol{y'}$.
Now, if we consider the uniform distribution U on the group of permutations on $\{1, \ldots, 2m\}$ that swap elements i and $i+m$, we have

$$P^{2m}(R) \leq \sup_{\boldsymbol{xy}} U\left\{\sigma : (\boldsymbol{xy})^\sigma \in R\right\},$$

where $\boldsymbol{z}^\sigma = (z_{\sigma(1)}, \ldots, z_{\sigma(2m)})$ for $\boldsymbol{z} \in X^{2m}$. Fix $\boldsymbol{xy} \in X^{2m}$. For a subsequence $\boldsymbol{x'y'} \subseteq_\eta \boldsymbol{xy}$, we let $(\boldsymbol{x'y'})^\sigma$ denote the corresponding subsequence of the permuted version of $\boldsymbol{xy}$ (and similarly for $(\boldsymbol{x'})^\sigma$ and $(\boldsymbol{y'})^\sigma$). Then

$$\begin{aligned} U\left\{\sigma\colon (\boldsymbol{xy})^\sigma \in R\right\} &\leq U\big\{\sigma\colon \exists \boldsymbol{x'y'} \subseteq_\eta \boldsymbol{xy}, \exists h \in H, \ell((\boldsymbol{x'y'})^\sigma, h) \leq 2^{i+1}, \mathrm{Er}_{(\boldsymbol{x'})^\sigma}(h) = 0, \\ &\qquad \mathrm{Er}_{(\boldsymbol{y'})^\sigma}(h) \geq \tfrac{m}{2}\epsilon(m,i,\delta) - (|\boldsymbol{y}| - |\boldsymbol{y'}|)\big\} \\ &\leq \sum_{\boldsymbol{x'y'} \subseteq_\eta \boldsymbol{xy}} U\big\{\sigma\colon \exists h \in H, \ell((\boldsymbol{x'y'})^\sigma, h) \leq 2^{i+1}, \mathrm{Er}_{(\boldsymbol{x'})^\sigma}(h) = 0, \\ &\qquad \mathrm{Er}_{(\boldsymbol{y'})^\sigma}(h) \geq \tfrac{m}{2}\epsilon(m,i,\delta) - (|\boldsymbol{y}| - |\boldsymbol{y'}|)\big\}. \end{aligned}$$

For a fixed subsequence $\boldsymbol{x'y'} \subseteq_\eta \boldsymbol{xy}$, define the event inside the last sum as A. We can partition the group of permutations into a number of equivalence classes, so that, for all i, within each class all permutations map i to a fixed value unless $\boldsymbol{x'y'}$ contains both x_i and y_i. Clearly, all equivalence classes have equal probability, so we have

$$\begin{aligned} U(A) &= \sum_C \Pr(A|C)\Pr(C) \\ &\leq \sup_C \Pr(A|C), \end{aligned}$$

where the sum and supremum are over equivalence classes C. But within an equivalence class, $(\boldsymbol{x'y'})^\sigma$ is a permutation of $\boldsymbol{x'y'}$, so we can write

$$\begin{aligned} \Pr(A|C) &= \Pr\big(\exists h \in H, \ell((\boldsymbol{x'y'})^\sigma, h) \leq 2^{i+1}, \mathrm{Er}_{(\boldsymbol{x'})^\sigma}(h) = 0, \\ &\qquad \mathrm{Er}_{(\boldsymbol{y'})^\sigma}(h) \geq \tfrac{m}{2}\epsilon(m,i,\delta) - (|\boldsymbol{y}| - |\boldsymbol{y'}|) \;\big|\; C\big) \\ &\leq \sup_{\sigma \in C} \left|H_{|(\boldsymbol{x'},\boldsymbol{y'})^\sigma}\right| \sup_h \Pr\left(\mathrm{Er}_{(\boldsymbol{x'})^\sigma}(h) = 0, \mathrm{Er}_{(\boldsymbol{y'})^\sigma}(h) \geq \tfrac{m}{2}\epsilon(m,i,\delta) \big| C\right), \quad (2) \end{aligned}$$

where the second supremum is over the subset of H for which $\ell((\boldsymbol{x'y'})^\sigma, h) \leq 2^{i+1}$. Clearly,

$$\left|H_{|(\boldsymbol{x'},\boldsymbol{y'})^\sigma}\right| \leq 2^{i+1},$$

and the probability in (2) is no more than

$$2^{-m\epsilon(m,i,\delta)/2 + 2\eta m}.$$

Combining these results, we have

$$P^{2m}(J_i \cap \bar{S}) \leq \binom{2m}{2\eta m} 2^{i+1} 2^{-m/2\epsilon(m,i,\delta)+2\eta m},$$

and this is no more than $\delta_i/2 = p_i\delta/2$ if

$$\frac{m}{2}\epsilon(m,i,\delta) \geq 2\eta m \log(2m) + i + 1 + 2\eta m + \log\frac{4}{p_i\delta}.$$

The theorem follows. ∎

6 Examples of Probably Smooth Luckiness Functions

In this section, we consider four examples of luckiness functions and show that they are probably smooth. The first example (Example 5.1) is the simplest; in this case luckiness depends only on the hypothesis h and is independent of the examples $\boldsymbol{x}$. In the second example, luckiness depends only on the examples, and is independent of the hypothesis. The third example allows us to predict the generalization performance of the maximal margin classifier. In this case, luckiness clearly depends on both the examples and the hypothesis. (This is the only example we present here where the luckiness function is *both* a function of the data and the hypothesis.) The fourth example concerns the VC-dimension of a class of functions when restricted to the particular sample available.

First Example

If we consider Example 5.1, the unluckiness function is clearly probably smooth if we choose $\phi(m, U(\boldsymbol{x}, h), \delta) = (2em/U)^U$, and $\eta(m, U, \delta) = 0$ for all m and δ. The bound on generalization error that we obtain from Theorem 5.5 is almost identical to that given in Theorem 2.1.

Second Example

The second example we consider involves examples lying on hyperplanes.

Definition 6.1 *Define the unluckiness function $U(\boldsymbol{x}, h)$ for a linear threshold function h as $U(\boldsymbol{x}, h) = \dim \operatorname{span}\{\boldsymbol{x}\}$, the dimension of the vector space spanned by the vectors $\boldsymbol{x}$.*

Proposition 6.2 *Let H be the class of linear threshold functions defined on $\mathbb{R}^d$. The unluckiness function of Definition 6.1 is probably smooth with respect to $\phi(m, U, \delta) = (2em/U)^U$ and*

$$\eta(m, U, \delta) = \frac{4}{m}\left(U \ln\left(\frac{2em}{U}\right) + \ln\left(\frac{4d}{\delta}\right)\right).$$

Proof: The recognition of a k dimensional subspace is a learning problem for the indicator functions H_k of the subspaces. These have VC dimension k. Hence, applying the hierarchical approach of Theorem 2.1 taking $p_k = 1/d$, we obtain the given error bound for the number of examples in the second half of the sequence lying outside the subspace. Hence, with probability $1 - \delta$ there will be a $(1 - \eta)$-subsequence of points all lying in the given subspace. For this sequence the growth function is bounded by $\phi(m, U, \delta)$. ∎

The above example will be useful if we have a distribution which is highly concentrated on the subspace with only a small probability of points lying outside it. We conjecture that it is possible to relax the assumption that the probability distribution is concentrated exactly on the subspace, to take advantage of a situation where it is concentrated around the subspace and the classifications are compatible with a perpendicular projection onto the space. This will also make use of both the data and the classification to decide the luckiness.

Third Example

We are now in a position to state the result concerning maximal margin hyperplanes.

Proposition 6.3 *The unluckiness function of Definition 5.2 is probably smooth with* $\phi(m, U, \delta) = (2em/(9U))^{9U}$, *and*

$$\eta(m, U, \delta) = \frac{1}{2m}\left(k \log\left(\frac{8em}{k}\right)\log(32m) + \log(4m + 2) + 2\log\left(\frac{2}{\delta}\right)\right),$$

where $k = \lfloor 1297U \rfloor$.

Proof: By the definition of the unluckiness function U, we have that the maximal margin hyperplane has margin γ satisfying,

$$U = R^2/\gamma^2,$$

where

$$R = \max_{1 \le i \le m} \|x_i\|.$$

The proof works by allowing two sets of points to be excluded from the second half of the sample, hence making up the value of η. By ignoring these points with probability $1 - \delta$ the remaining points will be in the ball of radius R about the origin and will be correctly classified by the maximal margin hyperplane with a margin of $\gamma/3$. Provided this is the case then the function $\phi(m, U, \delta)$ gives a bound on the growth function on the double sample of hyperplanes with larger margins. Hence, it remains to show that with probability $1 - \delta$ there exists a fraction of $\eta(m, U, \delta)$ points of the double sample whose removal leaves a subsequence of points satisfying the above conditions. First consider the class

$$\mathcal{H} = \{f_\rho | \rho \in \mathbb{R}^+\},$$

where

$$f_\rho(x) = \begin{cases} 1; & \text{if } \|x\| \le \rho, \\ 0; & \text{otherwise.} \end{cases}$$

The class has VC dimension 1 and so by the permutation argument with probability $1-\delta/2$ at most a fraction η_1 of the second half of the sample are outside the ball B centered at the origin containing with radius R, where

$$\eta_1 = \frac{1}{m}\left(\log(2m+1) + \log\frac{2}{\delta}\right),$$

since the growth function $B_{\mathcal{H}}(m) = m+1$. We now consider the permutation argument applied to the points of the double sample contained in B to estimate how many are closer to the hyperplane than $\gamma/3$ or are incorrectly classified. This involves an application of Lemma 3.6 with $\gamma/3$ substituted for γ and using the folding argument introduced just before that Lemma. We have by Corollary 4.4 that

$$\text{fat}_{\mathcal{F}}(\gamma) \leq \min\{9R^2/\gamma^2, n+1\} + 1,$$

where $\mathcal{F}$ is the set of linear threshold functions with unit weight vector restricted to points in a ball of radius R about the origin. Hence, with probability $1-\delta/2$ at most a fraction η_2 of the second half of the sample that are in B are either not correctly classified or within a margin of $\gamma/3$ of the hyperplane, where

$$\eta_2 = \frac{1}{m}\left(k\log\frac{8em}{k}\log(32m) + \log\frac{4}{\delta}\right),$$

for $k = \lfloor 1297R^2/\gamma^2 \rfloor = \lfloor 1297U \rfloor$. The result follows by adding the numbers of excluded points $\eta_1 m$ and $\eta_2 m$ and expressing the result as a fraction of the double sample as required. ∎

Combining the results of Theorem 5.5 and Proposition 6.3 gives the following corollary.

Corollary 6.4 *Suppose p_d, for $d = 1, \ldots, 2m$, are positive numbers satisfying $\sum_{d=1}^{2m} p_d = 1$. Suppose $0 < \delta < 1/2$, $t \in H$, and P is a probability distribution on X. Then with probability $1-\delta$ over m independent examples $\boldsymbol{x}$ chosen according to P, if a learner finds an hypothesis h that satisfies $\text{Er}_{\boldsymbol{x}}(h) = 0$, then the generalization error of h is no more than*

$$\begin{aligned}\epsilon(m,U,\delta) \;=\; & \frac{2}{m}\left(\left\lceil 9U\log\frac{2em}{9U}\right\rceil + 2\log\sqrt{32}m\log\frac{4}{p_i\delta} + \right.\\ & \left. + \left(\lfloor 1297U\rfloor\log\left(\frac{8em}{\lfloor 1297U\rfloor}\right)\log(32m) + \log(16m+8)\right)\log 4m\right)\end{aligned}$$

where $U = U(\boldsymbol{x}, h)$ for the unluckiness function of Definition 5.2.

If we compare this corollary with Theorem 4.5, there is an extra $\log(m)$ factor that arises from the fact that we have to consider all possible permutations of the omitted η subsequence in the general proof, whereas that is not necessary in the direct argument based on fat-shattering. The additional generality obtained here is that the support of the probability distribution does not need to be known, and even if it is we may derive advantage from observing points with small norms, hence giving a better value of $U = R^2/\gamma^2$ than would be obtained in Theorem 4.5 where the *a priori* bound on R must be used.

Vapnik has used precisely the expression for this unluckiness function (given in Definition 5.2) as an estimate of the effective VC dimension of the Support Vector Machine [48, p.139]. The functional obtained is used to locate the best suited complexity class among different polynomial kernel functions in the Support Vector Machine [48, Table 5.6]. The result above shows that this strategy is well-founded by giving a bound on the generalization error in terms of this quantity.

It is interesting to note that the support vectors' unluckiness function of Definition 5.3 relates to the same classifiers but one which for a given same sample defines a different ordering on the functions in the class in the sense that a large margin can occur with a large number of support vectors, while a small margin can be forced by a small number of support vectors.

We will omit the proof of the probable smoothness of the support vectors' unluckiness function since a more direct bound on the generalization error can be obtained using the results of Floyd and Warmuth [19]. Since the set of support vectors is a compression scheme, Theorem 5.1 of [19] can be rephrased as follows.

Let MMH be the function that returns the maximal margin hyperplane consistent with a labelled sample. Note that applying the function MMH to the labelled support vectors returns the maximal margin hyperplane of which they are the support vectors.

Theorem 6.5 (Littlestone and Warmuth [33]) *Let D be any probability distribution on a domain X, c be any concept on X. Then the probability that $m \geq d$ examples drawn independently at random according to D contain a subset of at most d examples that map via* MMH *to a hypothesis that is both consistent with all m examples and has error larger than ϵ is at most*

$$\sum_{i=0}^{d} \binom{m}{i} (1-\epsilon)^{m-i}.$$

The theorem implies that the generalization error of a maximal margin hyperplane with d support vectors among a sample of size m can with confidence $1-\delta$ be bounded by

$$\frac{1}{m-d}\left(d \log \frac{em}{d} + \log \frac{m}{\delta}\right),$$

where we have allowed different numbers of support vectors by applying standard SRM to the bounds for different d. Note that the value d, that is the unluckiness of Definition 5.3, plays the role of the VC dimension in the bound.

Using a similar technique to that of the above theorem it is possible to show that the support vectors' unluckiness function is indeed probably smooth with respect to

$$\begin{aligned} \eta(m,d,\delta) &= \frac{1}{2m}\left(d \log \frac{2em}{d} + \log \frac{1}{\delta}\right) \\ \text{and} \quad \phi(m,d,\delta) &= \left(\frac{2em}{d}\right)^{d}. \end{aligned}$$

However, the resulting bound on the generalization error involves an extra log factor.

Fourth Example

The final example is more generic in nature as we do not indicate how the luckiness function might be computed or estimated. This might vary according to the particular representation. If H is a class of functions and $\boldsymbol{x} \in X^m$, we write $H_{|\boldsymbol{x}} = \{h_{|\boldsymbol{x}} : h \in H\}$.

Definition 6.6 *Consider a hypothesis class H and define the unluckiness function $U(\boldsymbol{x}, h)$ for a function $h \in H$ as*

$$U(\boldsymbol{x}, h) = \text{VCdim}(H_{|\boldsymbol{x}}).$$

The motivation for this example can be found in a number of different sources.

Recently Sontag [44] showed the following result for smoothly parametrized classes of functions: Under mild conditions, if all sets in general position of size equal to the VC dimension of the class are shattered, then the VC dimension is bounded by half the number of parameters. This implies that even if the VC dimension is super-linear in the number of the parameters, it will not be so on all sets of points. In fact the paper shows that there are nonempty open sets of samples which cannot be shattered. Hence, though we might consider a hypothesis space such as a multi-layer sigmoidal neural network whose VC dimension can be quadratic [27] in the number of parameters, it is possible that the VC dimension when restricted to a particular sample is only linear in the number of parameters. However there are as yet no learning results of the standard kind that take advantage of this result (to get appropriately small sample size bounds) when the conditions of his theorem hold. The above luckiness function does take advantage of Sontag's result implicitly in the sense that it can detect, whether the situation which Sontag predicts will sometimes occur, has in fact occurred. Further, it can then exploit this to give better bounds on generalization error.

A further motivation can be seen from the distribution dependent learning described in [5], where it is shown that classes which have infinite VC dimension may still be learnable provided that the distribution is sufficiently concentrated on regions of the input space where the set of hypotheses has low VC dimension. The problem with that analysis is that there is no apparent way of checking *a priori* whether the distribution is concentrated in this way. The probable smoothness of the unluckiness function of Definition 6.6 shows that we can effectively estimate the distribution from the sample and learn successfully if it witnesses a region of low VC dimension.

In addition to the above two motivations, the approach mirrors closely that taken in a recent paper by Lugosi and Pintér [36]. They divide the original sample in two and use the first part to generate a covering set of functions for the hypothesis class in a metric derived from the function values on these points. They then choose the function from this cover which minimises the empirical error on the second half of the sample. They bound the error of the function in terms of the size of the cover derived on the first set of points. However, the size of this cover can be bounded by the VC dimension of the set of hypotheses when restricted to these points. Hence, the generalization is effectively bounded in terms of a VC-dimension estimate derived from the sample. The bound they obtain is difficult to compare directly with the one given below, since it is expressed in terms of the *expected* size of the cover. In addition, their estimator must build a (potentially very large) empirical cover of the function class. Lugosi and Nobel [35] have more recently extended this work in a number of ways, in particular to general regression problems. However their bounds are all still in terms of expected size of covers.

We begin with a technical lemma which analyses the probabilities under the swapping group of permutations used in the symmetrisation argument. The group Σ consists of all 2^m permutations which exchange corresponding points in the first and second halves of the sample, i.e. $x_j \leftrightarrow y_j$ for $j \in \{1, \ldots, m\}$.

Lemma 6.7 *Let Σ be the swapping group of permutations on a $2m$ sample of points $\boldsymbol{xy}$. Consider any fixed set $z_1, \ldots, z_d$ of the points. For $3k < d$ the probability $P_{d,k}$ under the uniform distribution over permutations that exactly k of the points $z_1, \ldots, z_d$ are in the first half of the sample is bounded by*

$$P_{d,k} \leq \binom{d}{k} 0.5^d.$$

Proof : The result is immediate if no pair of z_i's is in corresponding positions in opposite halves of the sample, since the expression counts the fraction of permutations which leave exactly k points in the first half. The rest of the proof is concerned with showing that when pairs of z_i's do occur in opposite positions the probability is reduced. Let $P_{d,k}^{2l}$ be the probability when l pairs are matched in this way. In this case, whatever the permutation, l points are in the first half, and to make up the number to k a further $k - l$ trials must succeed out of $d - 2l$, each trial having probability 0.5. Hence

$$P_{d,k}^{2l} = \binom{d-2l}{k-l} 0.5^{d-2l}.$$

Note that

$$\begin{aligned} P_{d,k}^{2(l+1)} &= \binom{d-2l-2}{k-l-1} 0.5^{d-2l-2} \\ &= g(k,l) P_{d,k}^{2l}, \end{aligned}$$

where

$$g(k,l) = \frac{4(k-l)(d-k-l)}{(d-2l)(d-2l-1)}.$$

The result will follow if we can show that $g(k,l) \leq 1$ for all relevant values of k and l. The function $g(k,l)$ attains its maximum value for $k = d/2$ and since it is a quadratic function of k with negative coefficient of the square term, its maximum in the range of interest is strictly less than

$$g(d/3, l) = \frac{4(d-3l)(2d-3l)}{9(d-2l)(d-2l-1)}.$$

Hence, in the range of interest $g(k,l) < 1$, if

$$\begin{aligned} 4(d-3l)(2d-3l) &\leq 9(d-2l)(d-2l-1) \\ \Leftarrow d^2 - 9d + 18l &\geq 0 \\ \Leftarrow d &\geq 9. \end{aligned}$$

Hence, for $d \geq 9$ we have for all l that

$$P_{d,k}^{2l} \leq P_{d,k}^{0} = \binom{d}{k} 0.5^d,$$

and the result follows. For $d < 9$ a problem could only arise for the case when $l = 0$ in view of the $18l$ in the above equation, i.e. the case when a single linked pair is introduced. Hence, one point is automatically in the first half. Since $3k < d < 9$, we need only consider $k = 2$ and $d = 7, 8$. By the equation above $d = 7$ will be the worst case. It is, however, easily verified that $P^2_{7,2} \leq P^0_{7,2}$ as required. ∎

Proposition 6.8 *The unluckiness function of Definition 6.6 is probably smooth with respect to* $\phi(m, U, \delta)$ *and* $\eta(m, U, \delta) = 0$*, where*

$$\phi(m, U, \delta) = \left(\frac{2em}{\alpha U}\right)^{\alpha U},$$

and

$$\alpha = 3.08\left(1 + \frac{1}{U}\ln\frac{1}{\delta}\right).$$

Proof: Let $\alpha = \alpha(U, \delta)$ be as in the proposition statement. The result will follow if we can show that with high probability the ratio between the VC dimensions obtained by restricting H to the single and double samples is not greater than α. Formally expressed it is sufficient to show that

$$P^{2m}\left\{\boldsymbol{xy} : \alpha(\text{VCdim}(H_{|\boldsymbol{x}}), \delta)\text{VCdim}(H_{|\boldsymbol{x}}) < \text{VCdim}(H_{|\boldsymbol{xy}})\right\} \leq \delta,$$

since ϕ gives a bound on the growth function for a set of functions with VC dimension αU, where U is the VC dimension measured on the first half of the sample. We use the symmetrisation argument to bound the given probability. Let the VC dimension on the double sample be d and consider points $z_1, \ldots, z_d \in \boldsymbol{xy}$ which are shattered by H. We stratify the bound by considering the case when k of these d points are on the left hand side under the given permutation. By Lemma 6.7 the probability $P_{d,k}$ that this occurs is bounded by

$$P_{d,k} \leq \binom{d}{k} 0.5^d,$$

provided $k < 3d$. Having k points in the first half will not violate the condition if

$$\alpha(U, \delta)U \geq d,$$

for all $U \geq k$. This is because with k of the points $z_1, \ldots, z_d$ on the left hand side we must have

$$U = \text{VCdim}(H_{|\boldsymbol{x}}) \geq k.$$

Since $\alpha(U, \delta)U$ is monotonically increasing we can bound the probability of the condition being violated by summing the probabilities $P_{d,k}$ for k such that $\alpha(k, \delta)k < d$. Let U satisfy the equation $\alpha(U, \delta)U = d = \alpha U$. Hence, since $3U < d$, it suffices to show that

$$L = \sum_{k=0}^{\lfloor U \rfloor} P_{d,k} \leq \sum_{k=0}^{\lfloor U \rfloor} \binom{d}{i} 0.5^d \leq \delta.$$

But we can bound L as follows:

$$\begin{aligned} L &\leq \frac{1}{2^d}\left(\frac{ed}{U}\right)^U \\ &\leq \frac{(e\alpha)^U}{2^{\alpha U}}, \end{aligned}$$

Hence, $L \leq \delta$, provided

$$\alpha(U,\delta) = \alpha \geq \log(\alpha e) + \frac{1}{U}\log\frac{1}{\delta}.$$

Using Lemma 3.2 from [42] with $c = 1$, the above holds provided

$$\alpha(\ln 2 - 1/e) \geq \frac{1}{U}\ln\frac{1}{\delta} + 1,$$

and this holds when

$$\alpha(U,\delta) = 3.08\left(1 + \frac{1}{U}\ln\frac{1}{\delta}\right),$$

as required. ∎

Corollary 6.9 *Suppose $0 < \delta < 1/2$, $t \in H$, and P is a probability distribution on X. Then with probability $1-\delta$ over m independent examples $\boldsymbol{x}$ chosen according to P, if a learner finds an hypothesis h that satisfies $\mathrm{Er}_{\boldsymbol{x}}(h) = 0$, and in addition bounds the quantity $\mathrm{VCdim}(H_{|\boldsymbol{x}})$ by U, then the generalization error of h is no more than*

$$\epsilon(m,U,\delta) = \frac{2}{m}\left\{3.08\left(U + \ln\frac{1}{\delta}\right)\log\frac{2em}{3.08U} + \log\frac{8m}{\delta}\right\}$$

Proof: We apply the proposition together with Theorem 5.5, choosing $p_i = 2/m$, for $i = 1,\ldots,2m$. ∎

Observe that this corollary could be interpreted as a result about "effective VC-dimension." A similar notion was introduced in [22], but the precise definition was not given there. The above corollary is the first result along these lines of which we are aware, that gives a theoretical performance bound in terms of quantities that can be determined empirically (albeit at a potentially large computational cost).

7 Conclusions

The aim of this paper has been to show that structural risk minimisation can be performed by specifying in advance a more abstract stratification of the overall hypothesis class. In this new inductive framework the subclass of the resulting hypothesis depends on its relation to the observed data and not just a predefined partition of the functions. The luckiness function of the data and hypothesis captures the stratification implicit in the approach, while probable smoothness is the property required to ensure that the 'luckiness' observed on the sample can

be used to reliably infer 'lucky' generalization. We have shown that Vapnik's maximal margin hyperplane algorithm is an example of implementing this strategy where the luckiness function is the ratio of the maximum size of the input vectors to the maximal margin observed.

Since lower bounds exist on *a priori* estimates of generalization derived from VC dimension bounds, the better generalization bounds must be a result of a non-random relation between the probability distribution and the target hypothesis. This is most evident in the maximal margin hyperplane case where the distribution must be concentrated away from the separating hyperplane.

There are many different avenues that might be pursued through the application of the ideas in practical learning algorithms, since it allows practitioners to take advantage of their intuitions about structure that might be present in particular problems. By encapsulating these ideas in an appropriate luckiness function, they can potentially derive algorithms and generalization bounds significantly better than the normal worst case PAC estimates, if their intuitions are correct.

From the analytic point of view many questions are raised by the paper. Corresponding lower bounds would help place the theory on a tighter footing and might help resolve the role of the additional $\log(m)$ factor introduced by the luckiness framework. Alternatively, it may be possible to either refine the proof or the definition of probable smoothness to eliminate this apparent looseness in the bound.

Another exciting prospect from a theoretical angle is the possibility of linking this work with other a posteriori bounds on generalization. The most notable example of such bounds is that provided by the Bayesian approach, where the volume of weight space consistent with the hypothesis is treated in much the same manner as a luckiness function (see for example [39, 40]). Indeed, the size of the maximal margin can be viewed as a way of bounding from below the volume of weight space consistent with the hyperplane classification. Hence, other weight space volume estimators could be considered though it seems unlikely that the true volume itself would be probably smooth since accurate estimation of the true volume requires too many sample points. If Bayesian estimates could be placed in this framework the role of the prior distribution, which has been a source of so much criticism of the approach, could be given a more transparent status, while the bounds themselves would become distribution independent.

Acknowledgements

We would like to thank Vladimir Vapnik for useful discussions at a Workshop on Artificial Neural Networks: Learning, Generalization and Statistics at the Centre de Recherches Mathématiques, Université de Montréal, where some of these results were presented.

This work was carried out in part whilst John Shawe-Taylor and Martin Anthony were visiting the Australian National University, and whilst Robert Williamson was visiting Royal Holloway and Bedford New College, University of London.

This work was supported by the Australian Research Council, and the ESPRIT Working Group in Neural and Computational Learning (NeuroCOLT Nr. 8556). Martin Anthony's visit to Australia [25] was partly financed by the Royal Society.

Much of this work was done whilst the authors were at ICNN95, and we would like to thank the organizers for providing the opportunity.

References

[1] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi and David Haussler, "Scale-sensitive Dimensions, Uniform Convergence, and Learnability," in *Proceedings of the Conference on Foundations of Computer Science (FOCS)*, (1993). Also to appear in *Journal of the ACM*.

[2] Martin Anthony and Peter Bartlett, "Function learning from interpolation", Technical Report, (1994). (An extended abstract appeared in *Computational Learning Theory, Proceedings 2nd European Conference, EuroCOLT'95*, pages 211–221, ed. Paul Vitanyi, (Lecture Notes in Artificial Intelligence, 904) Springer-Verlag, Berlin, 1995).

[3] Martin Anthony, Norman Biggs and John Shawe-Taylor, The Learnability of Formal Concepts," pages 246–257 in *Proceedings of the Third Annual Workshop on Computational Learning Theory*, Rochester Morgan Kaufmann, (1990).

[4] Martin Anthony and John Shawe-Taylor, "A Result of Vapnik with Applications," *Discrete Applied Mathematics*, **47**, 207–217, (1993).

[5] Martin Anthony and John Shawe-Taylor, "A sufficient condition for polynomial distribution-dependent learnability," *Discrete Applied Mathematics*, **77**, 1–12, (1997).

[6] Andrew R. Barron, "Approximation and Estimation Bounds for Artificial Neural Networks," *Machine Learning*, **14**, 115–133, (1994).

[7] Andrew R. Barron, "Complexity Regularization with Applications to Artificial Neural Networks," pages 561–576 in G. Roussas (Ed.) *Nonparametric Functional Estimation and Related Topics* Kluwer Academic Publishers, 1991.

[8] Andrew R. Barron and Thomas M. Cover, "Minimum Complexity Density Estimation," *IEEE Transactions on Information Theory*, **37**, 1034–1054; 1738, (1991).

[9] Peter 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," Technical Report, Department of Systems Engineering, Australian National University, May 1996.

[10] Peter L. Bartlett and Philip M. Long, "Prediction, Learning, Uniform Convergence, and Scale-Sensitive Dimensions," Preprint, Department of Systems Engineering, Australian National University, November 1995.

[11] Peter L. Bartlett, Philip M. Long, and Robert C. Williamson, "Fat-shattering and the learnability of Real-valued Functions," *Journal of Computer and System Sciences*, **52**(3), 434-452, (1996).

[12] Gyora M. Benedek and Alon Itai, "Dominating Distrubutions and Learnability," pages 253–264 in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh ACM, (1992).

[13] Michael Biehl and Manfred Opper, "Perceptron Learning: The Largest Version Space," in *Neural Networks: The Statistical Mechanics Perspective.* Proceedings of the CTP–PBSRI Workshop on Theoretical Physics, World Scientific. Also available at: `http://brain.postech.ac.kr/nnsmp/compressed/biehl.ps.Z`

[14] Bernhard E. Boser, Isabelle M. Guyon, and Vladimir N. Vapnik, "A Training Algorithm for Optimal Margin Classifiers," pages 144–152 in *Proceedings of the Fifth Annual Workshop on Computational Learning Theory*, Pittsburgh ACM, (1992).

[15] Kevin L. Buescher and P.R. Kumar, "Learning by Canonical Smooth Estimation, Part I: Simultaneous Estimation," *IEEE Transactions on Automatic Control*, **41**(4), 545 (1996).

[16] Corinna Cortes and Vladimir Vapnik, "Support-Vector Networks," *Machine Learning*, **20**, 273–297 (1995).

[17] Thomas M. Cover and Joy Thomas, *Elements of Information Theory*, Wiley, New York, 1994.

[18] Richard O. Duda and Peter E. Hart, *Pattern Classification and Scene Analysis*, John Wiley and Sons, New York, 1973.

[19] Sally Floyd and Manfred Warmuth, "Sample Compression, learnability, and the Vapnik-Chervonenkis Dimension," *Machine Learning*, **21**, 269–304 (1995).

[20] Frederico Girosi, Michael Jones and Tomaso Poggio, "Regularization Theory and Neural Networks Architecture," *Neural Computation*, **7**, pages 219–269, (1995).

[21] Leonid Gurvits and Pascal Koiran, "Approximation and Learning of Convex Superpositions," pages 222–236 in Paul Vitanyi (Ed.), *Proceedings of EUROCOLT95* (Lecture Notes in Artificial Intelligence 904), Springer, Berlin, 1995.

[22] Isabelle Guyon, Vladimir N. Vapnik Bernhard E. Boser, Leon Bottou and Sara A. Solla, "Structural Risk Minimization for Character Recognition," pages 471–479 in John E. Moody *et al.* (Eds.) *Advances in Neural Information Processing Systems 4*, Morgan Kaufmann Publishers, San Mateo, CA, 1992.

[23] Mohamad H. Hassoun, *Fundamentals of Artificial Neural Networks*, MIT Press, Cambridge, MA, 1995.

[24] David Haussler, "Decision Theoretic Generalizations of the PAC Model for Neural Net and Other Learning Applications," *Information and Computation*, **100**, 78–150 (1992).

[25] Donald Horne, *The Lucky Country: Australia in the Sixties*, Penguin Books, Ringwood, Victoria, 1964.

[26] Michael J. Kearns and Robert E. Schapire, "Efficient Distribution-free Learning of Probabilistic Concepts," pages 382–391 in *Proceedings of the 31st Symposium on the Foundations of Computer Science*, IEEE Computer Society Press, Los Alamitos, CA, 1990.

[27] Pascal Koiran and Eduardo D. Sontag, "Neural Networks with Quadratic VC Dimension," to appear in NIPS95 and also *Journal of Computer and System Sciences*; also available as a NeuroCOLT Technical Report NC-TR-95-044 (`ftp://ftp.dcs.rhbnc.ac.uk/pub/neurocolt/tech_reports`).

[28] P.R. Kumar and Kevin L. Buescher, "Learning by Canonical Smooth Estimation, Part 2: learning and Choice of Model Complexity," *IEEE Transactions on Automatic Control*, **41**(4), 557 (1996).

[29] Nathan Linial, Yishay Mansour and Ronald L. Rivest, "Results on Learnability and the Vapnik-Chervonenkis Dimension," *Information and Computation*, **90**, 33–49, (1991).

[30] Nick Littlestone, "Learning Quickly When Irrelevant Attributes Abound: A New Linear Threshold Algorithm," *Machine Learning* **2**, 285–318 (1988).

[31] Nick Littlestone, "Mistake-driven bayes Sports: Bounds for Symmetric Apobayesian Learning Algorithms," Technical Report, NEC Research Center, New Jersey, (1996).

[32] Nick Littlestone and Chris Mesterham, "An Apobayesian Relative of Winnow," Preprint, NEC Research Center, New Jersey, (1996).

[33] Nick Littlestone and Manfred Warmuth, "Relating Data Compression and Learnability", unpublished manuscript, University of California Santa Cruz, 1986.

[34] Lennart Ljung, *System Identification: Theory for the User*, Prentice-Hall PTR, Upper Saddle River, New Jersey, 1987.

[35] Gábor Lugosi and Andrew B. Nobel, "Adaptive Model Selection Using Empirical Complexities," Preprint, Department of Mathematics and Computer Science, technical University of Budapest, Hungary, (1996).

[36] Gábor Lugosi and Márta Pintér, "A Data-dependent Skeleton Estimate for Learning," pages 51–56 in *Proceedings of the Ninth Annual Workshop on Computational Learning Theory*, Association for Computing Machinery, New York, 1996.

[37] Gábor Lugosi and Kenneth Zeger, "Nonparametric Estimation via Empirical Risk Minimization," *IEEE Transactions on Information Theory*," **41**(3), 677–687, (1995).

[38] Gábor Lugosi and Kenneth Zeger, "Concept Learning Using Complexity Regularization," *IEEE Transactions on Information Theory*, **42**, 48–54, (1996).

[39] David J.C. MacKay, "Bayesian Model Comparison and Backprop Nets," pages 839–846 in John E. Moody *et al.* (Eds.) *Advances in Neural Information Processing Systems 4*, Morgan Kaufmann Publishers, San Mateo, CA, 1992.

[40] David J.C. MacKay, "Probable Networks and Plausible Predictions — A Review of Practical Bayesian methods for Supervised Neural Networks," Preprint, Cavendish Laboratory, Cambridge (1996).

[41] David Pollard, *Convergence of Stochastic Processes*, Springer, New York, 1984.

[42] John Shawe-Taylor, Martin Anthony and Norman Biggs, "Bounding sample size with the Vapnik-Chervonenkis dimension", *Discrete Applied Mathematics*, **42**, 65–73, (1993).

[43] John Shawe-Taylor, Peter Bartlett, Robert Williamson and Martin Anthony, "A Framework for Structural Risk Minimization", pages 68–76 in *Proceedings of the 9th Annual Conference on Computational Learning Theory*, Association for Computing Machinery, New York, 1996.

[44] Eduardo D. Sontag, "Shattering all Sets of k points in 'General Position' Requires $(k-1)/2$ Parameters," Rutgers Center for Systems and Control (SYCON) Report 96-01; Also NeuroCOLT Technical Report NC-TR-96-042 (`ftp://ftp.dcs.rhbnc.ac.uk/pub/neurocolt/tech_reports`).

[45] Aad W. van der Vaart and Jon A. Wellner, *Weak Convergence and Empirical Processes*, Springer, New York, 1996.

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

[47] Vladimir N. Vapnik, "Principles of Risk Minimization for Learning Theory," pages 831–838 in John E. Moody *et al.* (Eds.) *Advances in Neural Information Processing Systems 4*, Morgan Kaufmann Publishers, San Mateo, CA, 1992.

[48] Vladimir N. Vapnik, *The Nature of Statistical Learning Theory*, Springer-Verlag, New York, 1995.

[49] Vladimir N. Vapnik and Aleksei Ja. Chervonenkis, "On the Uniform Convergence of Relative Frequencies of Events to their Probabilities," *Theory of Probability and Applications*, **16**, 264–280 (1971).

[50] Vladimir N. Vapnik and Aleksei Ja. Chervonenkis, "Ordered Risk Minimization (I and II)", *Automation and Remote Control*, **34**, 1226–1235 and 1403–1412 (1974).