A Framework for Structural Risk Minimisation

John Shawe-Taylor
Computer Science Dept
Royal Holloway
University of London
Egham, TW20 0EX, UK
john@dcs.rhbnc.ac.uk

Peter L. Bartlett
Systems Engineering Dept
Australian National Univ
Canberra 0200 Australia
Peter.Bartlett@anu.edu.au

Robert C. Williamson
Engineering Dept
Australian National Univ
Canberra 0200 Australia
Bob.Williamson@anu.edu.au

Martin Anthony
Mathematics Dept
London School of Economics
Houghton Street
London WC2A 2AE, UK
anthony@vax.lse.ac.uk

Abstract

The paper introduces a framework for studying structural risk minimisation. The model views structural risk minimisation in a PAC context. It then considers the more general case when the hierarchy of classes is chosen in response to the data. This theoretically explains the impressive performance of the maximal margin hyperplane algorithm of Vapnik. It may also provide a general technique for exploiting serendipitous simplicity in observed data to obtain better prediction accuracy from small training sets.

1 Introduction

The standard 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

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

and be told that the target will lie in one of the sets H_d. Linial, Mansour and Rivest [5] studied learning in such a framework 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 addresses two shortcomings of the above 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 generalisation 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 generalisation error (with a fixed number of examples) than allowing some extra errors in a more restricted class.

The model we consider will allow 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 [8, 10]. Indeed, the results reported in Sections 2 and 3 of this paper are implicit in the cited references, though we feel justified in presenting them in this framework as we believe they deserve greater attention and development. (We also make explicit some of the assumptions inherent in the presentations in [8, 10].) In Sections 4 and 5 we address a shortcoming of the SRM method which Vapnik [8] highlights: *according to the SRM principle the structure has to be defined* a priori *before the training data appear.* An heuristic using maximal separation hyperplanes proposed by Vapnik and coworkers [3] violates this principle in that the hierarchy defined depends on the data. We introduce a framework which allows this dependency on the data and yet can still place rigorous bounds on the generalisation error. As an example, we show that the maximal margin hyperplane approach proposed by Vapnik falls into the framework, thus placing the approach on a firm theoretical foundation and solving a long standing open problem in statistical induction.

The approach can be interpreted as a 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).

2 Basic Example — No Training Errors

As an initial example we consider a hierarchy of classes

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

where we will assume VCdim$(H_d) = d$ for the rest of this section. Such a hierarchy of classes is termed a decomposable concept class by Linial *et al.* [5]. We will assume that a fixed number m of labelled examples are given as a vector $\mathbf{z} = (\mathbf{x}, t(\mathbf{x}))$ to the learner, where $\mathbf{x} = (x_1, \ldots, x_m)$, and $t(\mathbf{x}) = (t(x_1), \ldots, t(x_m))$, 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 $\mathbf{z}$. What we require is a function $\epsilon(m, d, \delta)$ which will give the learner an upper bound on the generalisation error of h with confidence $1-\delta$. The following theorem gives a suitable function. We use $\mathrm{Er}_{\mathbf{z}}(h) = |\{i : h(x_i) \neq t(x_i)\}|$ to denote the *number* of errors that h makes on $\mathbf{z}$, and $\mathrm{er}_P(h) = P\{\mathbf{x}\text{: } h(\mathbf{x}) \neq t(\mathbf{x})\}$ to denote the *expected error* when $x_1, \ldots, x_m$ are drawn independently according to P.

Theorem 1 *Let H_i be as above, and let p_d be any set of positive numbers satisfying $\sum_{d=1}^{\infty} p_d = 1$. With probability $1 - \delta$ over m independent identically distributed example, for any d for which a learner finds a consistent hypothesis h in H_d, the generalisation error of h is bounded by*

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

Proof : The proof uses the standard bound on generalisation error for each of the classes but divides the confidence between the classes giving proportion p_d to class H_d. Hence, we show that

$$\begin{aligned}\Pr\{\mathbf{z} \quad : \quad & \exists d, \exists h \in H_d, \mathrm{Er}_{\mathbf{z}}(h) = 0, \text{ and} \\ & \mathrm{er}_P(h) > \epsilon(m, d, \delta)\} < \delta,\end{aligned}$$

by showing that for all d

$$\Pr\{\mathbf{z} : \exists h \in H_d, \mathrm{Er}_{\mathbf{z}}(h) = 0, \mathrm{er}_P(h) > \epsilon(m, d, \delta)\} < \delta p_d.$$

The probability on the left of the inequality is, however, bounded with the usual analysis via Sauer's lemma by

$$4\Pi_{H_d}(2m) \exp\left(\frac{-\epsilon(m, d, \delta)m}{4}\right),$$

where $\Pi_{H_d}(m)$ is the growth function: the maximum number of dichotomies implementable on m points by H_d. Substituting the given value of $\epsilon(m, d, \delta)$ gives the required bound of δp_d. ∎

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. (Vapnik [7] implicitly assumes $p_i = 1/d$, $i = 1, \ldots, d$.) The corresponding term

$$-\frac{4}{m}\ln(p_d),$$

represents the overestimate of $\epsilon(m, d, \delta)$ arising from our prior uncertainty about which class to use. If we have any information about which classes are more likely to arise we can use it to improve our bias. For example, if we know that class H_d will occur then we choose $p_d = 1$ and recover the standard PAC model. If on the other hand the probabilities of the function falling in different classes are known to be q_d, the expected overestimate in $\epsilon(m, d, \delta)$ or loss will be

$$L(\mathbf{q}) = \frac{4}{m}\sum_{d=1}^{\infty} -q_d \ln(p_d).$$

Hence, the difference between the loss of $\mathbf{p}$ and that obtained by using the values $\mathbf{q}$ as bias terms is

$$L(\mathbf{p}) - L(\mathbf{q}) = \frac{m}{4}\mathrm{KL}(\mathbf{p}, \mathbf{q}),$$

where $\mathrm{KL}(\cdot, \cdot) \geq 0$ is the Kullback-Leibler divergence between the two distributions. Hence, we have the following corollary.

Corollary 2 *If we know the probabilities q_d of the target falling into the classes H_d, then we minimise the expected overestimate of the generalisation error by choosing $p_d = q_d$. In this case the expected amount of the overestimate is*

$$\frac{4}{m}H(\mathbf{q}),$$

where $H(\mathbf{q})$ is the entropy of the distribution $\mathbf{q}$.

3 Structural Risk Minimisation

Using the framework established in the previous section we now wish to consider the possibility of errors on the training sample.

We will make use of the following result of Vapnik in a slightly improved version of Anthony and Shawe-Taylor [2]. Note also that the result is expressed in terms of the quantity $\mathrm{Er}_{\mathbf{z}}(h)$ which denotes the number of errors of the hypothesis h on the sample $\mathbf{z}$, rather than the usual proportion of errors.

Theorem 3 ([2]) *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 ν be any probability measure on $S = X \times \{0, 1\}$. Then the probability (with respect to ν^m) that for $\mathbf{z} \in S^m$, there is some $h \in H$ such that*

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

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 generalisation 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.

In what follows we will often write $\mathrm{Er}_{\mathbf{x}}(h)$ (rather than $\mathrm{Er}_{\mathbf{z}}(h)$) when the target t is obvious from the context.

Theorem 4 *Let H_i, $i = 1, 2, \ldots$, be defined as in Section 2 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 $\mathbf{x}$, if the learner finds an hypothesis h in H_d with $\mathrm{Er}_{\mathbf{x}}(h) = k$, then the generalisation error of h is bounded by

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

Proof: Again we bound the required probability of failure

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

by showing that for all d and k

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

We will apply Theorem 3 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 $\mathrm{er}_P(h) > \epsilon(m,d,k,\delta)$ and $\mathrm{Er}_{\mathbf{z}}(h) = k$, then

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

as required for an application of the theorem. Substitution of the appropriate parameters into the expression of the probability and then simplifying gives the required result, by ignoring one of the negative terms arising from squaring γ_{dk}. ■

Note that the term ignored at the end of the proof could be used to improve the error bound, particularly for large k, hence reducing the required size of the q_{dk} in these cases. As we will see below we may be content to choose q_{dk} small for these values as we do not expect them to be useful in practice in any case. For small k the term ignored is only marginally less than 1.

The question of how to choose 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 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, 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

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

Hence, for $d \approx m$ we would need to reduce the number of errors by on average about 1.42 per increment of VC class. (Note that the tradeoff between errors on the sample and generalisation error is also discussed in [4].)

4 A General Framework for Decomposing Classes

The standard PAC analysis gives bounds on generalisation error that are uniform over the hypothesis class. Decomposing the hypothesis class, as described in Section 2, allows us to bias our generalisation error bounds in favour of certain target functions and distributions: those for which some hypothesis low in the hierarchy is an accurate approximation. In this section, we introduce a more general framework which subsumes both the standard PAC model and the framework described in Section 2. This can be viewed as one way to allow the decomposition of the hypothesis class to be chosen based on the sample. This allows us to bias our generalisation 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 $\mathrm{Er}_{\mathbf{z}}$ 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. Recall that $h(\mathbf{x}) = (h(x_1), \ldots, h(x_m))$ and let $H_{|\mathbf{x}} = \{h(\mathbf{x}) : h \in H\}$.

We define the *level* of a function $h \in H$ relative to L and $\mathbf{x}$ by the function

$$\ell(\mathbf{x}, h) = |\{g(\mathbf{x}) : \exists g \in H, L(\mathbf{x}, g) \geq L(\mathbf{x}, h)\}|,$$

or

$$\ell(\mathbf{x}, h) = |\{g(\mathbf{x}) : \exists g \in H, U(\mathbf{x}, g) \leq U(\mathbf{x}, h)\}|.$$

Whether $\ell(\mathbf{x}, h)$ is defined in terms of L or U is a matter of convenience; the quantity $\ell(\mathbf{x}, h)$ itself plays the central role in what follows.

4.1 Examples

Example 5 *Consider the hierarchy of classes introduced in Section 2 and define*

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

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

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

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

$$\ell((\mathbf{xy}), h) \leq \left(\frac{2em}{d}\right)^d,$$

where $(\mathbf{xy}) = (x_1, \ldots, x_m, y_1, \ldots, y_m)$*.*

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.

For the case of linear threshold functions in Euclidean space, Boser *et al.* [3] suggest 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 generalisation of the resulting classifier. They give evidence to indicate that the generalisation performance is frequently significantly better than that predicted by the VC dimension of the full class of linear threshold functions. The following theorem gives circumstantial evidence to indicate why this might occur.

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.

Theorem 6 (Vapnik [8]) *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,*

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

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

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

We can analyse the performance of the maximal margin classifier in terms of a certain sample-based decomposition of the class of linear threshold functions. This decomposition favours target functions and distributions for which it is likely that, for a large set of random examples, some linear threshold function will correctly classify the set such that most examples will lie far from the separating hyperplane. In Section 5, we apply the results of this section to show that Vapnik's approach is indeed well founded, and in particular that the generalisation error of the maximal margin classifier is considerably smaller when this assumption is satisfied. We use the following unluckiness function.

Definition 7 *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* $\mathbf{x}$*, then define*

$$U(\mathbf{x}, h) = \min\{256R^2(1 + R^2)A^2, n + 2\} + 1,$$

where $R = \max_{1 \leq i \leq m} \|x_i\|$ *and* $A = \|w\|$*.*

By Theorem 6 the unluckiness function is a loose upper bound on the VC dimension of the set of dichotomies realisable with hyperplanes with at least as large a separation. Hence,

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

where $d = U(\mathbf{x}, h)$.

4.2 Probable Smoothness

Definition 8 *An* α*-subsequence of a vector* $\mathbf{x}$ *is a vector* $\mathbf{x}'$ *obtained from* $\mathbf{x}$ *by deleting a fraction of at most* α *coordinates. We will also write* $\mathbf{x}' \subseteq_\alpha \mathbf{x}$*.*

A luckiness function $L(\mathbf{x}, h)$ *defined on a function class* H *is* probably smooth *with respect to functions* $\eta(m, L, \delta)$ *and* $\phi(m, L)$*, if, for all targets* t *in* H *and for every distribution* P*,*

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

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

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 $\phi(m, L)$).

Theorem 9 *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* $\mathbf{x}$ *chosen according*

to P, if for any $i \in \mathbb{N}$ a learner finds an hypothesis h in H with $\mathrm{Er}_{\mathbf{x}}(h) = 0$ and $\phi(m, L(\mathbf{x}, h)) \leq 2^{i+1}$, then the generalisation error of h satisfies $\mathrm{er}_P(h) \leq \epsilon(m, i, \delta)$ where

$$\begin{aligned}\epsilon(m,i,\delta) &= \frac{2}{m}\left(i+1+\log_2\frac{4}{p_i\delta}\right) \\ &\quad +4\eta(m, L(\mathbf{x},h), p_i\delta/4)(\log_2 2m+1).\end{aligned}$$

Proof : Using Chernoff bounds,

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

provided $m \geq 8/\epsilon(m,i,\delta)$, 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

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

Let S be the event

$$\begin{aligned}\{(\mathbf{xy}) : \exists h \in H, \mathrm{Er}_{\mathbf{x}}(h) = 0, \forall (\mathbf{x}'\mathbf{y}') \subseteq_\eta (\mathbf{xy}) \\ \ell((\mathbf{x}'\mathbf{y}'), h) > \phi(m, L(\mathbf{x},h))\}\end{aligned}$$

with $\eta = \eta(m, L_i, \delta_i/2)$. Since L is probably smooth, $P^{2m}(S) \leq \delta_i/2$. 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

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

where $|\mathbf{y}'|$ denotes the length of the sequence $\mathbf{y}'$.

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

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

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

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

For a fixed subsequence $(\mathbf{x}'\mathbf{y}') \subseteq_\eta (\mathbf{xy})$, 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 $(\mathbf{x}'\mathbf{y}')$ 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, $(\mathbf{x}'\mathbf{y}')^\sigma$ is a permutation of $(\mathbf{x}'\mathbf{y}')$, so we can write

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

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

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

and the probability in (1) 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

$$\frac{m}{2}\epsilon(m,i,\delta) \geq 2\eta m \log_2(2m) + i + 1 + 2\eta m + \log_2\frac{4}{p_i\delta}.$$

The theorem follows. ∎

5 Examples of Probable Smoothness

In this section, we consider three examples of luckiness functions and show that they are probably smooth. The first example (Example 5) is the simplest; in this case luckiness depends only on the hypothesis h and is independent of the examples $\mathbf{x}$. In the second example, luckiness depends only on the examples, and is independent of the hypothesis. The third example allows us to predict the generalisation performance of the maximal margin classifier. In this case, luckiness clearly depends on both the examples and the hypothesis.

If we consider Example 5, the unluckiness function is clearly probably smooth if we choose $\phi(m, U(\mathbf{x},h)) = \left(2em/U\right)^U$, and $\eta(m, U, \delta) = 0$ for all m and δ. The bound on generalisation error that we obtain from Theorem 9 is almost identical to that given in Theorem 1.

The second example we consider involves examples lying on hyperplanes.

Definition 10 *Define the unluckiness function $U(\mathbf{x}, h)$ for a linear threshold function h as $U(\mathbf{x}, h) = \dim \text{span}\{\mathbf{x}\}$, the dimension of the vector space spanned by the vectors* **x**.

Proposition 11 *Let H be the class of linear threshold functions defined on $\mathbb{R}^d$. The unluckiness function of Definition 10 is probably smooth with respect to $\phi(m, U) = \left(2em/U\right)^U$ 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 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 $\phi(m, U)$. ∎

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.

We turn for the rest of the paper to the example of the maximal margin hyperplanes and show that the unluckiness function defined in Definition 7 is probably smooth for appropriate (useful) choices of η and ϕ. We begin by some preliminary definitions and results.

Definition 12 *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 values 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.*

Definition 13 *Let $\mathcal{F}$ be a set of real valued functions. We say that a set of points X is level γ-shattered by $\mathcal{F}$ if it can be γ-shattered when choosing the r_x constant across all inputs 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.*

Note that the second dimension is a scale sensitive version of a dimension introduced by Vapnik [7]. The scale sensitive version was first introduced by Alon *et al.* [1].

Lemma 14 *Let $\mathcal{F}$ be the set of linear functions with unit weight vectors, restricted to points in a ball of radius R,*

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

Then the level fat shattering function can be bounded by

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

Proof: If a set of points X 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 such that

$$\langle w_b, x\rangle \begin{cases} \geq r + \gamma & \text{if } b_x = 1 \\ \leq r - \gamma & \text{otherwise} \end{cases}$$

Let $d = \min_{x \in X} |\langle w_b, x\rangle - r| \geq \gamma$. Consider the hyperplane defined by $(w_b', r') = (w_b/d, r/d)$. It is in canonical form with respect to the points X, satisfies $\|w_b/d\| = 1/d$ and realises the given dichotomy. Hence, the set of points X can be shattered by a subset of canonical hyperplanes satisfying $\|w_b'\| \leq 1/d \leq 1/\gamma$. The result follows from Theorem 6. ∎

Corollary 15 *Let $\mathcal{F}$ be the set of linear functions with unit weight vectors, restricted to points in a ball of radius R about the origin. The fat shattering function of $\mathcal{F}$ can be bounded by*

$$\text{Fat}_{\mathcal{F}}(\gamma) \leq \min\{4R^2/\gamma^2, n + 1\} + 1.$$

Proof: Suppose m points lying in a ball of radius R about the origin are γ-shattered with corresponding output values r_x. Clearly, $r_x < R$. We extend the points into one extra dimension adding the value r_x for the point x. That is, $x = (x_1, \ldots, x_d)$ becomes $(x_1, \ldots, x_d, r_x)$. Since $r_x < R$, the norm of the new vectors is bounded by $\sqrt{2}R$. By extending each of the weight vectors realising the fat shattering of the points with coordinate -1, we see that the points are level γ-shattered at level 0. The norm of the weight vectors is $\sqrt{2}$, so to obtain unit weight vectors we must divide them by this amount. If we multiply the m vectors by the same amount the effect is to leave the output unchanged. Hence, we have generated a set of m points in $n + 1$ dimensional space lying in a ball of radius $2R$, which are level γ-shattered by linear functions with unit weight vectors. By the lemma, we have $m \leq \min\{4R^2/\gamma^2, n + 1\} + 1$. ∎

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

Definition 16 *Let (X, d) be a (pseudo-) metric space, let A be a subset of X and $\epsilon > 0$. A set $B \subseteq A$ 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 and allow finite arguments to be used while covering the full space A. In our case we will use the l^∞ distance over a finite sample $\mathbf{x} = (x_1, \ldots, x_m)$ for the pseudo-metric in the space of functions,

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

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

We now quote a lemma from Alon *et al.* [1] which we will need in the proof of the next proposition.

Lemma 17 *(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*

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

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

Corollary 18 *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*

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

where the expectation E is over samples $\mathbf{x} \in X^m$ 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

$$E[\mathcal{N}(\epsilon, \mathcal{F}, \mathbf{x})] = E[\mathcal{N}(\epsilon/(b-a), \mathcal{F}', \mathbf{x})].$$

The result follows. ∎

Proposition 19 *Suppose $\mathcal{F}$ is a set of functions that map from X to $[a,b]$ with finite fat-shattering dimension. Then for any distribution P on X,*

$$P^{2m}\left\{(\mathbf{xy}) : \exists\gamma \in \mathbb{R}^+, \exists f \in \mathcal{F}, \right.$$
$$\left.\tfrac{1}{m}\left|\left\{i : |f(y_i)| \leq \min_i |f(x_i)| - 2\gamma\right\}\right| > \epsilon(m,\gamma,\delta)\right\} < \delta$$

where $\epsilon(m,\gamma,\delta)$ is

$$\tfrac{1}{m}\left(k_{\gamma/2}\log_2\left(\tfrac{8em(b-a)}{k_{\gamma/2}\gamma}\right)\log_2\left(\tfrac{32m(b-a)^2}{\gamma^2}\right) + \log_2\left(\tfrac{4(b-a)}{\gamma\delta}\right)\right),$$

and $k_\gamma = \text{Fat}_{\mathcal{F}}(\gamma)$.

Proof: The required probability is no more than

$$P^{2m}\left\{(\mathbf{xy}) : \exists\gamma \in \{(b-a)2^{-i} : i \in \mathbb{N}\}\right.$$
$$\exists f \in \mathcal{F}, \tfrac{1}{m}\left|\left\{i : |f(y_i)| \leq \min_i |f(x_i)| - 2\gamma\right\}\right| >$$
$$\left.\epsilon(m, 2\gamma, \delta)\right\}$$
$$\leq \sum_{i\in\mathbb{N}} P^{2m}\left\{(\mathbf{xy}) : \exists f \in \mathcal{F},\right.$$
$$\tfrac{1}{m}\left|\left\{i : |f(y_i)| \leq \min_i |f(x_i)| - 2\gamma_i\right\}\right| >$$
$$\left.\epsilon(m, 2\gamma_i, \delta)\right\},$$

where $\gamma_i = (b-a)2^{-i}$. It suffices to show that the i'th term in this sum is no more than $2^{-i}\delta$.

Fix i and let $k = k_{\gamma_i}$ and $\gamma = \gamma_i$. Using the standard permutation argument, we may fix a sequence $(\mathbf{xy})$ and bound the probability under the uniform distribution on swapping permutations that the permuted sequence satisfies the condition stated. Consider a minimal γ-cover B of $\mathcal{F}$ in the standard pseudo-metric with respect to the points $\mathbf{xy}$. Let

$$r = \min\{|f(x)| : x \in \mathbf{x}\}.$$

If we now pick any $f \in \mathcal{F}$, there exists $\hat{f} \in B$, with $|f(x) - \hat{f}(x)| < \gamma$ for all $x \in \mathbf{xy}$. Hence, for $x \in \mathbf{x}$, $|\hat{f}(x)| > r - \gamma$, and there are at least $\epsilon(m, 2\gamma, \delta)m$ points $x \in \mathbf{y}$ which satisfy $|\hat{f}(x)| < r - 2\gamma + \gamma = r - \gamma$. By the permutation argument at most $2^{-\epsilon(m,2\gamma,\delta)m}$ of the sequences obtained by swapping corresponding points satisfy the conditions. Hence, by Corollary 18 we may bound the probability of the event occurring by

$$E\left(\mathcal{N}(\gamma, \mathcal{F}, \mathbf{xy})\right) 2^{-\epsilon(m,2\gamma,\delta)m}$$
$$\leq 2\left(\frac{8m(b-a)^2}{\gamma^2}\right)^{k\log(4em(b-a)/(k\gamma))} 2^{-\epsilon(m,2\gamma,\delta)m}.$$

Now, this is no more than $2^{-i}\delta$ if

$$\epsilon(m, 2\gamma_i, \delta) \geq$$
$$\tfrac{1}{m}\left(k_{\gamma_i}\log_2\tfrac{8m(b-a)^2}{\gamma_i^2}\log_2\tfrac{4em(b-a)}{k_{\gamma_i}\gamma_i} + \log_2\tfrac{2^{i+1}}{\delta}\right),$$

and so we choose $\epsilon(m,\gamma,\delta)$ as in the proposition. ∎

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

Proposition 20 *The unluckiness function of Definition 7 is probably smooth with $\phi(m,U) = \left(2em/U\right)^U$, and*

$$\eta(m,U,\delta) = \frac{1}{m}\left(8\ln(2em) + 8\ln\left(\frac{12}{\delta}\right) + U\log_2\left(\frac{8em}{\sqrt{U}}\right)\log_2(32Um) + \log_2\left(\frac{12\sqrt{U}}{\delta}\right)\right).$$

Proof: Notice first that, for a set X_0 of points and a hyperplane (w,θ) in canonical form with respect to X_0, the distance from the plane to the closest point in X_0 is

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

We can therefore interchange the quantity $A = \|w\|$ in Definition 7 with the inverse of this minimum distance.

We will show that

$$P^{2m}\left\{(\mathbf{xy}) : \forall h \in H, \text{Er}_{\mathbf{x}}(h) = 0 \Rightarrow \exists(\mathbf{x}'\mathbf{y}') \subseteq_\eta (\mathbf{xy}),\right.$$
$$\max_i\|y_i'\| \leq \max_i\|x_i\| \text{ and}$$
$$\left.\min_i|\langle w_h, y_i'\rangle + \theta_h| \geq \min_i|\langle w_h, x_i\rangle + \theta_h|/2\right\} \geq 1-\delta,$$

where $\eta = \eta(m, U(\mathbf{x},h), \delta)$, and (w_h, θ_h) are such that

$$h : x \mapsto \text{sgn}\left(\langle w_h, x\rangle + \theta_h\right)$$

and $\|w_h\| = 1$. It follows from this that

$$P^{2m}\{(\mathbf{xy}) : \forall h \in H, \mathrm{Er}_{\mathbf{x}}(h) = 0 \to \exists(\mathbf{x}'\mathbf{y}') \subseteq_\eta (\mathbf{xy}), U((\mathbf{x}'\mathbf{y}'), h) \leq 2U(\mathbf{x}, h)\} \geq 1 - \delta,$$

and Theorem 6 implies the result for

$$\phi(m, U) = (2em/U)^U.$$

Now, define the events

$$E = \Big\{(\mathbf{xy}) : \exists h \in H, \mathrm{Er}_{\mathbf{x}}(h) = 0 \; \forall(\mathbf{x}'\mathbf{y}') \subseteq_\eta (\mathbf{xy}), \max_i \|y'_i\| > \max_i \|x_i\| \text{ or } \min_i |\langle w_h, y'_i\rangle + \theta_h| < \min_i |\langle w_h, x_i\rangle + \theta_h|/2\Big\},$$

and

$$S = \Big\{(\mathbf{xy}) : \exists(\mathbf{x}'\mathbf{y}') \subseteq_{\eta_1} (\mathbf{xy}), \max_i \|y'_i\| \leq \max_i \|x_i\|\Big\},$$

for some η_1. Then

$$\Pr(E) \leq \Pr(E|S) + \Pr(\bar{S}). \tag{2}$$

The second term is the probability that an excessive number of points in the second half sample lie outside the minimal ball containing the points of the first half sample. This is the probability of failure in learning a class with VC dimension 1, and hence it can be bounded by $\delta/3$ if

$$\eta_1 = \tfrac{1}{m}\left(\ln(2em) + \ln(12/\delta)\right).$$

A similar argument allows us to neglect linear threshold functions defined by (w, θ) with $\|w\| = 1$ and $|\theta| > R$, where $R = \max_i \|x_i\|$ (at the expense of another $\delta/3$ and another proportion η_1 of the second sample).

To estimate the first probability in (2), we can restrict our attention to the subsequence $\mathbf{y}'$ of $m' = m(1 - 2\eta_1)$ points which lie in a ball of radius R. Hence, the conditional probability in (2) is no more than

$$P^{2m'}\Big\{(\mathbf{xy}) : \exists h \in H, \mathrm{Er}_{\mathbf{x}}(h) = 0, \forall(\mathbf{x}'\mathbf{y}') \subseteq_\alpha (\mathbf{xy}), \min_i |\langle w_h, y'_i\rangle + \theta_h| < \min_i |\langle w_h, x_i\rangle + \theta_h|/2\Big\}, \tag{3}$$

where $\alpha = (\eta - 2\eta_1)/(1 - 2\eta_1)$.

Consider the class

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

By augmenting the input vector with an extra (constant) component, we can transform F_R into a linear class. Corollary 15 shows that

$$\mathrm{Fat}_{\mathcal{F}_R}(\gamma) \leq \min\left\{\frac{4R^2(1 + R^2)}{\gamma^2}, n + 2\right\} + 1.$$

Notice that functions in $\mathcal{F}_R$ map from points in a radius R ball about the origin to $[-2R, 2R]$. Let $d = \min_i |\langle w_h, x_i\rangle + \theta_h|$.

If we substitute

$$\begin{aligned} \gamma &:= d/4, \\ k_{\gamma/2} &:= \mathrm{Fat}_{\mathcal{F}_R}(\gamma/2) \\ &= \min\left\{256R^2(1 + R^2)/d^2, n + 2\right\} + 1 \\ &= U, \\ m &:= m', \\ \delta &:= \delta/3, \\ b - a &:= 4R \end{aligned}$$

in Proposition 19, we have that the probability (3) is less than $\delta/3$ if $(\eta - 2\eta_1)/(1 - 2\eta_1)$ is at least

$$\tfrac{1}{m'}\left(U\log_2\left(\tfrac{128em'R}{Ud}\right)\log_2\left(\tfrac{8192m'R^2}{d^2}\right) + \log_2\left(\tfrac{192R}{d\delta}\right)\right),$$

and so it will suffice if we set η equal to

$$\begin{aligned} &2\eta_1 + \tfrac{1}{m}\left(U\log_2\left(\tfrac{8em}{\sqrt{U}}\right)\log_2(32Um) + \log_2\left(\tfrac{12\sqrt{U}}{\delta}\right)\right) \\ &= \tfrac{1}{m}\Big(8\ln(2em) + 8\ln(12/\delta) + U\log_2\left(8em/\sqrt{U}\right)\log_2(32Um) + \log_2\left(12\sqrt{U}/\delta\right)\Big). \end{aligned}$$

■

Combining the results of Theorem 9 and Proposition 20 gives the following corollary.

Corollary 21 *Suppose p_d, for $d = 1, \ldots, 2m$, are positive numbers satisfying $\sum_{i=1}^{2m} p_i = 1$. 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 $\mathbf{x}$ chosen according to P, if a learner finds an hypothesis h that satisfies $\mathrm{Er}_{\mathbf{x}}(h) = 0$, then the generalisation error of h is no more than*

$$\begin{aligned} &\tfrac{2}{m}\left(U\log_2\left(\tfrac{2em}{U}\right) + \log_2\left(\tfrac{4}{p_i\delta}\right)\right) + \\ &\tfrac{4}{m}\Big(8\ln(2em) + 8\ln\left(\tfrac{48}{p_i\delta}\right) + U\log_2\left(\tfrac{8em}{\sqrt{U}}\right)\log_2(32Um) + \log_2\left(\tfrac{48\sqrt{U}}{p_i\delta}\right)\Big)\left(\log_2(2m) + 1\right) \\ &= O\left(\tfrac{1}{m}\left(U\log^2(Um) + \log\left(\tfrac{U}{p_i\delta}\right)\right)\log(m)\right), \end{aligned}$$

where $U = U(\mathbf{x}, h)$ for the unluckiness function of Definition 7, and $i = \lfloor U\log_2(2em/U)\rfloor$.

It seems likely that the bound in Corollary 21 can be improved. Nevertheless, the improvement over the standard PAC bounds can already be considerable, because the difference between the dimension of the space and the VC dimension of the set of hypotheses with similar margins (that is, the value of U) can be very large. Vapnik [8] cites an example where the space is 10^{13} dimensional while the effective VC dimension is approximately 10^3.

Acknowledgements

This work was supported by the Australian Research Council.

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, D. 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 John Shawe-Taylor, "A Result of Vapnik with Applications," *Discrete Applied Mathematics*, **47**, 207–217, (1993).

[3] B. Boser, I. Guyon, and V.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)

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

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

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

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

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

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

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