Sample Based Generalization Bounds*

Robert C. Williamson‡
bob.williamson@anu.edu.au

John Shawe-Taylor♯♭
jst@dcs.rhbnc.ac.uk

Bernhard Schölkopf†♭
bsc@microsoft.com

Alex J. Smola‡
alex.smola@anu.edu.au

‡ Department of Engineering
Australian National University
0200 Canberra, Australia

♯ Department of Computer Science
Royal Holloway,
University of London
Egham, TW20 0EX, UK

† Microsoft Research Limited
St George House
1 Guildhall Street
Cambridge CB2 3NH, UK

November 15, 1999

*This work was supported by the Australian Research Council and the European Commission under the Working Group Number 27150 (NeuroCOLT2). Parts of earlier versions of this work appeared in the Proceedings of the 4th European Conference on Computational Learning Theory (EUROCOLT'99), and ICANN'99. ♭ Work partially performed whilst at ANU.

Abstract

It is known that the covering numbers of a function class on a double sample (length $2m$, where m is the number of points in the sample) can be used to bound the generalization performance of a classifier by using a margin based analysis. Traditionally this has been done using a "Sauer-like" relationship involving a combinatorial dimension such as the fat-shattering dimension. In this paper we show that one can utilize an analogous argument in terms of the *observed* covering numbers on a single m-sample (being the actual observed data points). The significance of this is that for certain interesting classes of functions, such as support vector machines, one can readily estimate the empirical covering numbers quite well. We show how to do so in terms of the eigenvalues of the Gram matrix created from the data. These covering numbers can be much less than *a priori* bounds indicate in situations where the particular data received is "easy". The work can be considered an extension of previous results which provided generalization performance bounds in terms of the VC-dimension of the class of hypotheses restricted to the sample, with the considerable advantage that the covering numbers can be readily computed, and they often are small.

Index Terms: VC learning theory, sample complexity, support vector machines, covering numbers.

1 Introduction

The PAC framework ("probably approximately correct," sometimes known as the Statistical Learning framework) for analysing the generalization of a learning system bases its analysis on the complexity of the class of hypotheses that can be output by the learning algorithm. Often this leads to poor estimates of generalization as the class must be chosen large enough to solve a wide range of possible tasks. Structural Risk Minimisation [23] counters this problem by placing an *a priori* hierarchy on the class of functions and allowing the learner to seek a function starting in the simpler classes. If a satisfactory function is found in a simple class the corresponding bound on the generalization error is that much tighter. In this sense the estimate is obtained *a posteriori* based on the class determined by the training algorithm.

Only recently have techniques for bounding the tails of the distribution of a data-dependent estimator been proposed [20, 21, 15]. Initially Shawe-Taylor *et al.* [21] showed that the maximal margin hyperplane algorithm used for the support vector machine of Cortes and Vapnik [9] can be analysed in this way using the size of the margin as the predictor of generalization. This should be distinguished from classical Structural Risk Minimisation since the assignment of hypotheses to a complexity class depends on the data and also the target function. The large margin approach has been extended to general neural networks by Bartlett [3]. The line taken in this paper is based on a more general framework developed in [21] which allows inference of good generalization from different measures of performance other than the margin of the classifier.

The general approach described in [21] gives the motivation for the current paper, though a direct application of the techniques introduced there would not give the results obtained here, at least not without an extra $\log(m)$ factor. (In any case direct application of the results would not be a trivial undertaking.) Since the analysis of this paper is endeavouring to lead to more realistic bounds than previous classical and data-dependent techniques we are keen to avoid extra factors if possible.

Our main result is Theorem 4.4 which bounds the generalization error in terms of the covering numbers observed on the training set at a scale determined by the margin of the classifier. Roughly speaking, the role of the VC dimension in the traditional bound on classifier generalization performance is taken by the log of the covering number of the class when restricted to the observed data sample. The scale at which the covering number is measured depends on the observed margin.

The idea of bounding generalization error in terms of the VC dimension measured on the training sample was considered in [21]. The problem with the result there is that there is no simple way of estimating the VC dimension of a set of hypotheses. The approach would also not apply to bounding the generalization of large margin classifiers, since in that case the role of the VC dimension is played

by the fat-shattering dimension at a scale dictated by the size of the margin. The present paper is motivated by the fact that empirical covering numbers can be readily determined for interesting classes of machines, such as Support Vector (SV) machines. As an application of Theorem 4.4 we show in Section 5 how to compute the empirical covering numbers for support vector machines in terms of the eigenvalues of the Gram matrix.

2 Background Results

We will assume that a fixed number m of labelled examples $\mathbf{z} = (\mathbf{x}, t(\mathbf{x}))$ are given to the learner, where $\mathbf{x} = (x_1, \ldots, x_m)$, $x_i \in X$, and $t(\mathbf{x}) = (t(x_1), \ldots, t(x_m))$. (Thus $[\mathbf{x}]$ is a $(\dim X) \times m$ matrix.) We sometimes treat $\mathbf{x}$ as a set, for instance writing $x_i \in \mathbf{x}$ with the obvious meaning. 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 $\mathrm{er}_P(h) = P\{x: h(x) \neq t(x)\}$ to denote the *expected error* (or *generalisation error*) when x is drawn according to P. In what follows we will write $\mathrm{Er}_{\mathbf{x}}(h)$ (rather than $\mathrm{Er}_{\mathbf{z}}(h)$) when the target t is obvious from the context. If $\mathbf{x}, \mathbf{y} \in X^m$, we denote by $\mathbf{xy}$ their concatenation $(x_1, \ldots, x_m, y_1, \ldots, y_m)$. By log we denote logarithms to base 2.

The spaces ℓ_p^m, $1 \leq p \leq \infty$ are m-dimensional vector spaces (m can be infinite). If $x \in \ell_p^m$, $x = (x_1, \ldots, x_m)$, $\|x\|_{\ell_p^m} := (\sum_{i=1}^m |x_i|^p)^{1/p}$ for $p < \infty$ and $\|x\|_{\ell_\infty^m} := \max_{i=1,\ldots,m} |x_i|$ for $p = \infty$. When $m = \infty$, we simply write ℓ_p and elements x of ℓ_p are infinite sequences $(x_1, x_2, \ldots)$ with finite $\|x\|_{\ell_p}$.

We give the definition of the fat-shattering dimension, which was introduced in [13], and has been used for several problems in learning since [1].

Definition 2.1 (Fat-shattering Dimension) *Let $\mathcal{F}$ be a set of real valued functions. We say that a set of points $\mathbf{x}$ is* γ-shattered by $\mathcal{F}$ relative to $r = (r_x)_{x \in X}$ *if there are real numbers r_x indexed by $x \in \mathbf{x}$ such that for all binary vectors b indexed by $\mathbf{x}$, there is a function $f_b \in \mathcal{F}$ satisfying*

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

The fat-shattering dimension $\mathrm{fat}_{\mathcal{F}}$ *of the set $\mathcal{F}$ is a function from the positive real numbers to the nonnegative integers and infinity, mapping a value γ to the size of the largest set which is γ-shattered relative to some $(r_x)_x$, if this is finite, or infinity otherwise.*

Note that in our definition of the fat-shattering dimension we have used a slightly unconventional strict inequality for the value on a positive example. This will prove useful in the technical detail, but also ensures that the definition reduces to the Pollard dimension for $\gamma = 0$.

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 2.2 (Swapping [21]) *Let Σ be the swapping group of permutations on a $2m$ sample of points $\mathbf{xy}$. 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*

$$P_{d,k} \leq \binom{d}{k} 2^{-d}.$$

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

Definition 2.3 (Covering Numbers) *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 $d(a, b) \leq \epsilon$. 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 ∞). We will say the cover is* proper *if $B \subseteq A$.*

We have used a somewhat unconventional less than or equal to in the definition of a cover, as this will prove technically useful in the proofs. We next define the covering numbers that we are concerned with.

Definition 2.4 (Various Metrics) *Let $\mathcal{F}$ be a class of real-valued functions on the space X. For any $m \in \mathbb{N}$ and $\mathbf{x} = (x_1, \ldots, x_m) \in X^m$, we define the pseudo-metric*

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

This is simply $\|f_{|\mathbf{x}} - g_{|\mathbf{x}}\|_{\ell_\infty^m}$ where $f_{|\mathbf{x}}$ and $g_{|\mathbf{x}}$ are the restictions *of f and g to $\mathbf{x}$: $f_{|\mathbf{x}} = (f(x_1), \ldots, f(x_m))$. We write $\mathcal{N}(\epsilon, \mathcal{F}, \mathbf{x}) = \mathcal{N}_{d_{\mathbf{x}}}(\epsilon, \mathcal{F})$. Note that the cover is not required to be proper. Observe that $\mathcal{N}(\epsilon, \mathcal{F}, \mathbf{x}) = \mathcal{N}_{\ell_\infty^m}(\epsilon, \mathcal{F}_{\mathbf{x}})$, the ℓ_∞^m covering number of*

$$\mathcal{F}_{\mathbf{x}} := \{f_{|\mathbf{x}} : f \in \mathcal{F}\},$$

the class $\mathcal{F}$ restricted to the sample $\mathbf{x}$.

We now quote a lemma from [21] which follows directly from a result of Alon *et al.* [1].

Corollary 2.5 (Covering numbers via $\mathrm{fat}_{\mathcal{F}}$) *[21] Let $\mathcal{F}$ be a class of functions $X \to [a, b]$ and P a distribution over X. Choose $0 < \epsilon < 1$ and let $d = \mathrm{fat}_{\mathcal{F}}(\epsilon/4)$. Then*

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

Let $\pi_\gamma(\alpha)$ be the identity function in the range $[\theta - 2.01\gamma, \theta]$, with output θ for larger values and $\theta - 2.01\gamma$ for smaller ones, and let $\pi_\gamma(\mathcal{F}) = \{\pi_\gamma(f)\colon f \in \mathcal{F}\}$. The choice of the threshold θ is arbitrary but will be fixed before any analysis is made.

We will need some compactness properties of the class of functions which will hold in all cases usually considered. We formalise the requirement in the following definition.

Definition 2.6 (Sturdy Functions) *For $m \in \mathbb{N}$, let*

$$\tilde{\mathbf{x}}\colon \mathcal{F} \longrightarrow \mathbb{R}^m, \quad \tilde{\mathbf{x}}\colon f \mapsto f_{|\mathbf{x}}$$

denote the multiple evaluation map *induced by $\mathbf{x} \in X^m$. We say that a class of functions $\mathcal{F}$ is* sturdy *if for all $m \in \mathbb{N}$ and all $\mathbf{x} \in X^m$ the image $\tilde{\mathbf{x}}(\mathcal{F})$ of $\mathcal{F}$ under $\tilde{\mathbf{x}}$ is a compact subset of $\mathbb{R}^m$.*

Recall [14] that a subset of $\mathbb{R}^m$ is compact iff it is closed and bounded.

Lemma 2.7 ($\mathcal{N}(\gamma, \mathcal{F}, \mathbf{x})$ is Right continuous) *Let $\mathcal{F}$ be a sturdy class of functions. Then for each $N \in \mathbb{N}$ and any fixed sequence $\mathbf{x} \in X^m$, the infimum $\gamma_N = \inf\{\gamma : \mathcal{N}(\gamma, \mathcal{F}, \mathbf{x}) = N\}$, is attained.*

Proof We first note that for any fixed subset $g := (g_1, \ldots, g_N) \in \mathbb{R}^{m \times N}$ the quantity

$$\gamma_g := \max_{\mathbf{z} \in \tilde{\mathbf{x}}(\mathcal{F})} \min_{1 \le i \le N} \|\mathbf{z} - g_i\|_{\ell_\infty^m} \tag{1}$$

is well defined. This is the case since $\min_{1 \le i \le N} \|\mathbf{z} - g_i\|_{\ell_\infty^m}$ is a continuous function in $\mathbf{z}$ and the maximum over $\mathbf{z}$ is taken in a compact set $\tilde{\mathbf{x}}(\mathcal{F})$.

Secondly we note that γ_g is also a continuous function in g. Hence also the minimum over $g \in 2RU_{\ell_\infty^{m \times N}}$ is well defined. Here $U_{\ell_\infty^{m \times N}}$ denotes the (compact) unit ball with respect to the $\ell_\infty^{m \times N}$ norm and $R = \max_{\mathbf{z} \in \tilde{\mathbf{x}}(\mathcal{F})} \|\mathbf{z}\|_{\ell_\infty^m}$.

Next we have to show that

$$\min_{g \in 2RU_{\ell_\infty^{m \times N}}} \gamma_g = \gamma^* := \min_{g \in \mathbb{R}^{m \times N}} \gamma_g.$$

This is true since clearly $\gamma^* \le R$ and any $g_i \notin 2RU_{\ell_\infty^m}$ can be dropped from the minimization over i in (1) without increasing γ_g. Thus only $g_i \in 2RU_{\ell_\infty^m}$ have to be considered.

Finally, assume that $\gamma_N = \inf\{\gamma : \mathcal{N}(\gamma, \mathcal{F}, \mathbf{x}) = N\} < \gamma^*$. This implies that there exists some $\gamma_N < \gamma' < \gamma^*$ and a corresponding cover $(g'_1, \ldots, g'_N) \subset \mathbb{R}^{m \times N}$ such that $\gamma' = \max_{\mathbf{z} \in \tilde{\mathbf{x}}(\mathcal{F})} \min_{1 \le i \le N} \|\mathbf{z} - g'_i\|_{\ell_\infty^m}$. This, however, contradicts the definition of γ^*. Hence γ^* exists and is optimal. ∎

We will use the following lemma, which in the form below is given by Vapnik [23, page 168].

Lemma 2.8 (Bounds on the error using the double sample) *Let X be a set and S a system of sets on X, and P a probability measure on X. For $\mathbf{x} \in X^m$, $\mathbf{y} \in X^m$, and $A \in S$, define the frequency of occurence of the event A on the sample $\mathbf{x}$ as*

$$\nu_{\mathbf{x}}(A) := \frac{1}{m}|\{x_i : x_i \in A,\ i = 1, \ldots, m\}|.$$

If $m > 2/\epsilon$, then

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

3 Growth Function Bounds

In this section we consider sample based estimates of the growth function of a set of hypotheses and generalization bounds in terms of them. This will set the stage for the results involving covering numbers which will form the subject of the next section.

Let H be a class of functions mapping an input space X to $\{-1, 1\}$. Let $\mathbf{x} = (x_1, \ldots, x_m) \in X^m$. Recall that the class $H_{|\mathbf{x}}$ is simply H restricted to $\mathbf{x}$: $H_{|\mathbf{x}} = \{\mathbf{x} \mapsto h_{|\mathbf{x}} : h \in H\}$. The *growth function of H restricted to* $\mathbf{x}$ is

$$B_H(\mathbf{x}) = \left|H_{|\mathbf{x}}\right|.$$

The *growth function* of H is

$$B_H^m = \sup\{B_H(\mathbf{x}) : \mathbf{x} \in X^m\}.$$

The *VC Dimension* of H is

$$\mathrm{VCdim}(H) = \sup\{m \in \mathbb{N} : B_H^m = 2^m\}.$$

A key result concerning the VC Dimension is that if $\mathrm{VCdim}(H) = d$, $B_H^m \leq (em/d)^d$. This is the basis for the use of the VC dimension in existing bounds on generalization performance of classifiers. Furthermore, it follows immediately from the definitions that $2^{\mathrm{VCdim}(H_{|\mathbf{x}})} \leq B_H(\mathbf{x})$.

We first quote Corollary 6.31 of [21], which bounds the generalization of a classifier in terms of the VC dimension of the set of hypotheses restricted to the sample.

Theorem 3.1 (Generalization Bound via $\mathrm{VCdim}(H_{|\mathbf{x}})$) *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 $\mathrm{Er}_{\mathbf{x}}(h) = 0$, and in addition bounds the quantity $\mathrm{VCdim}(H_{|\mathbf{x}})$ by U, then the generalization error of h is no more than*

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

We now consider an application of Theorem 3.1. Ruiz and López-de-Teruel [16, 17] describe a method of verifying the generalization ability of a classifier which they test experimentally on benchmark datasets. Their method relies on randomizing the labels given to the training examples and ascertaining the minimum number of training errors that can be obtained for the randomized labels. If the error rate is high, then good generalization can be expected. Our technique allows one to bound the performance of this method. Ruiz presents an alternative analysis that is not readily comparable with ours. A related idea can be found in [25].

We will provide a PAC style bound on the generalization error of a classifier in terms of the number of mistakes made when the classifications are randomised. We first prove a lemma giving a probabilistic bound on the VC dimension in terms of the number of mistakes. If the VC dimension on the sample is d, then there is a set of that size that can be shattered and so there is a hypothesis h' that makes no mistakes on that set. The expected number of mistakes that h' makes on the remaining $m-d$ points is $(m-d)/2$. Hence we would expect $d = m - 2k$ to be an estimate of the VC dimension. In order to obtain a reliable bound on the probabilities involved we must take a larger value of $d = m - 1.8k - 1$. To prove the following technical lemma we will need a version of Bernstein's inequality (cf. e.g. [10] for the formulation below).

Proposition 3.2 (Bernstein) *Let X_i, $i = 1, \ldots, n$, be identically distributed random variables taking values $-1, +1$ with probability $1/2$. Let $S_n = \sum_{i=1}^{n} X_i$. Then for $x > 0$,*

$$P(S_n \geq x) \leq \exp\left\{\frac{-x^2}{2(n+x)}\right\}.$$

Lemma 3.3 (VC Dimension Bounds from Random Labels) *Fix $1 \leq \alpha < 2$. Suppose we are given a training sample $\mathbf{x} \in X^m$ for a set of hypotheses H. Suppose the labels $\mathbf{y}$ are chosen at random (uniformly) from $\{-1, 1\}^m$. Then the following probability (with respect to this random labelling) satisfies*

$$P\Big(\mathrm{VCdim}(H_{|\mathbf{x}}) > m - \lfloor \alpha k \rfloor - 1\Big) \leq \exp\left\{-\left(1 - \frac{\alpha}{2}\right)^2\right\}^k$$

where k is defined by

$$k = \min_{h \in H}(\mathrm{Er}_{\mathbf{x}}(h)),$$

with respect to the target values $\mathbf{y}$.

Proof Suppose that $\mathrm{VCdim}(H_{|\mathbf{x}}) > d - 1 = m - \lfloor \alpha k \rfloor - 1$. Assume without loss of generality that $x_1, \ldots, x_d$ can be shattered. Hence, whatever the random choice of the y_i for these inputs, there exists $h' \in H$ which realises those values. For the remaining points with probability $1/2$, $h'(x_i) = y_i$, for $i > d$. But

$$k \;=\; \min_{h \in H}(\mathrm{Er}_{\mathbf{x}}(h)) \leq \mathrm{Er}_{\mathbf{x}}(h')$$

and so h' must make at least k errors for $i > d$. Hence, we have that

$$P\Big(\mathrm{VCdim}(H_{|\mathbf{x}}) \geq m - \lfloor \alpha k \rfloor\Big) \;\leq\; P\Big(|\{i > d : h'(x_i) = y_i\}| \geq k\Big).$$

We can express this second probability in the following way

$$P\Big(|\{i > d : h'(x_i) = y_i\}| \geq k\Big) \;=\; P\Big(\sum_{i=1}^{m-d} X_i \geq 2k + d - m\Big),$$

where X_i is the random variable taking value 1 if $h'(x_i) = y_i$ and -1 otherwise. Since $2k + d - m > 0$, we can now apply Proposition 3.2 to obtain

$$\begin{aligned}
P\Big(|\{i > d : h'(x_i) = y_i\}| \geq k\Big) &\leq \exp\left\{\frac{-(2k+d-m)^2}{2(m-d+2k+d-m)}\right\} \\
&= \exp\left\{\frac{-(2k+d-m)^2}{4k}\right\} \\
&= \exp\left\{-\left(1 - \frac{m-d}{2k}\right)^2 k\right\} \\
&= \exp\left\{-\left(1 - \frac{\lfloor \alpha k \rfloor}{2k}\right)^2 k\right\} \qquad (2) \\
&\leq \exp\left\{-\left(1 - \frac{\alpha k}{2k}\right)^2 k\right\} \qquad (3) \\
&= \exp\left\{-\left(1 - \frac{\alpha}{2}\right)^2 k\right\}.
\end{aligned}$$

The step from (2) to (3) is justified since $\alpha k < 2k$. ∎

We can now state a theorem bounding generalization in terms of the number of errors made on the randomized problem.

Theorem 3.4 (Uniform Convergence Bounds from Random Labels) *Suppose $0 < \delta < 1/2$, $t \in H$, and P is a probability distribution on X. Assume $m \geq 140 \log(2/\delta)$. 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 generalization error of h is no more than*

$$\epsilon(m, U, \delta) \;=\; \frac{2}{m}\left\{3.08\left(U + \ln\frac{2}{\delta}\right)\log\frac{2em}{3.08U} + \log\frac{16m}{\delta}\right\},$$

where $U = m - 1.8k - 1$, and $k = \min_{h \in H}(\mathrm{Er}_{\mathbf{x}}(h))$ with respect to a set of random target values $\mathbf{y}$, provided $k \geq m/2$.

Proof We apply Lemma 3.3 with $\alpha = 1.8$. The bound will follow from an application of Theorem 3.1 using $\delta/2$ provided we can show that the bound on the probability of Lemma 3.3 that U is an upper bound on $\mathrm{VCdim}(H_{|\mathbf{x}})$ is smaller than $\delta/2$. Substituting $\alpha = 1.8$ and the minimal value of $k = m/2$ gives

$$\exp\left\{-\left(1-\frac{\alpha}{2}\right)^2\right\}^k \leq \exp\left\{-\left(1-\frac{\alpha}{2}\right)^2\right\}^{m/2} \leq \exp\{-0.01\}^{m/2} \leq \frac{\delta}{2}$$

provided $m \geq 140\log(2/\delta)$. Hence the theorem follows. ∎

Note that the restriction we have placed on k is not significant, since the bound would be trivial if k were as small as $m/2$.

We have used a tail bound on the probability that the VC dimension estimate is misleading, and so there will be only a small price to pay if we were to consider doing several rerandomizations and taking the minimum of the VC dimensions obtained. For example if we made t such experiments, we would only require that the probability that any one should be misleading be less than $\delta/(2t)$, that is

$$\exp\{-0.01\}^{m/2} \quad \leq \quad \delta/(2t),$$

which will follow provided $m \geq 140\log(2t/\delta)$.

We now present a simple corollary to Theorem 3.1 which illustrates the form of the main result we develop for covering numbers in the next section.

Corollary 3.5 (Generalization Bound via $B_H(\mathbf{x})$) *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 generalization error of h is no more than*

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

where $U = \lfloor \log(B_H(\mathbf{x})) \rfloor$.

Proof Simply observe that

$$\mathrm{VCdim}(H_{|\mathbf{x}}) \leq \lfloor \log(B_H(\mathbf{x})) \rfloor = U,$$

since $2^{\mathrm{VCdim}(H_{|\mathbf{x}})} \leq B_H(\mathbf{x})$ and $\mathrm{VCdim}(H_{|\mathbf{x}}) \in \mathbb{N}$. The result then follows from an application of the theorem. ∎

In the next section we will consider estimates of the covering numbers on a sample at a scale chosen according to the margin of the classifier. This will allow us to infer a lower bound on the fat-shattering dimension on the sample, giving with

high probability a bound on the fat-shattering dimension on the double sample, hence giving the bound on generalization error. This outline argument is refined a little by bounding the covering numbers on the double sample in terms of the product of the covering numbers on the two halves. We can therefore slightly improve the overall bound by making use of the given bound for the covering numbers on the first half.

4 Covering Numbers on a Double Sample

We begin by presenting a key proposition that shows with high probability the covering numbers on a sample provide a good estimate of the covering numbers on a double sample. Although the result contains no reference to the fat-shattering dimension, it does play a key role in the proof. It is the combinatorial properties of the fat-shattering dimension which make it possible to infer the properties of the second half of the sample from the first. The probabilistic inference of the fat-shattering dimension on the double sample in terms of its value on the first half involves a multiplicative factor slightly larger than three. Its precise form is given in the following definition.

Definition 4.1 *For $U \in \mathbb{N}$ and $\delta \in \mathbb{R}^+$, we define the function*

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

Proposition 4.2 *Suppose $\mathcal{F}$ is a set of functions mapping from X to $\mathbb{R}$. For any probability distribution P on X, For fixed $U \in \mathbb{N}$, for all $\epsilon > 0$, $\delta \in (0, 1)$ and $m \in \mathbb{N}$,*

$$P^{2m}\Big\{\mathbf{xy} \in X^{2m} \;:\; \Big(\lfloor \log \mathcal{N}(\epsilon/4, \mathcal{F}, \mathbf{x}) \rfloor = U \text{ and } 2\mathcal{N}(\epsilon, \mathcal{F}, \mathbf{x}) 2^{\alpha(U,\delta) U \log(17m) \log(5em/U)} < \mathcal{N}(\epsilon, \pi_\epsilon(\mathcal{F}), \mathbf{xy})\Big)\Big\} \leq \delta.$$

Proof

Part 1. For notational convenience, if $\mathcal{F}$ is a set of functions, we write $\mathcal{F}_{\mathbf{x}}$ for $\mathcal{F}_{|\mathbf{x}}$. If $B_{\mathbf{x}}$ is an ϵ-cover of $\mathcal{F}_{\mathbf{x}}$ for the function class $\mathcal{F}$ and $B_{\mathbf{y}}$ is an ϵ-cover for $\mathcal{F}_{\mathbf{y}}$, we can form an (improper) cover B of $\mathcal{F}_{\mathbf{xy}}$ by simply choosing a function which agrees with each pair of functions from $B_{\mathbf{x}} \times B_{\mathbf{y}}$ on their respective domains. If the sequence $\mathbf{y}$ contains common points with $\mathbf{x}$ delete all such common points from $\mathbf{y}$. The size of the cover required for $\mathbf{y}$ will decrease as a result. Hence, $|B| \leq |B_{\mathbf{x}}||B_{\mathbf{y}}|$. It follows that

$$\mathcal{N}(\epsilon, \mathcal{F}, \mathbf{xy}) \leq \mathcal{N}(\epsilon, \mathcal{F}, \mathbf{x}) \mathcal{N}(\epsilon, \mathcal{F}, \mathbf{y}). \tag{4}$$

Next observe that for $\epsilon > 0$, the fat-shattering dimension $\mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon)$ satisfies

$$\mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon) \leq \lfloor \log \mathcal{N}(\epsilon, \mathcal{F}, \mathbf{x}) \rfloor, \tag{5}$$

since any pair of functions realising a distinct dichotomy with margin ϵ must differ by more than 2ϵ at some point in $\mathbf{x}$ and hence cannot be covered by the same function in any cover.

Part 2. For any $\epsilon > 0$ let

$$A^{\epsilon}_{\mathbf{xy}} := \{\mathbf{xy}\colon \alpha(\mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon), \delta)\mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon) < \mathrm{fat}_{\mathcal{F}_{\mathbf{xy}}}(\epsilon)\}.$$

Following an argument similar to one in [21] we will show that for any $\epsilon > 0$, $P^{2m}(A^{\epsilon}_{\mathbf{xy}}) \leq \delta$. Let $d = \mathrm{fat}_{\mathcal{F}_{\mathbf{xy}}}(\epsilon)$ and suppose $\mathbf{z} = (z_1, \ldots, z_d) \subset \mathbf{xy}$ are ϵ-shattered by $\mathcal{F}$. We use the usual permutation argument. Let

$$E_k := \{\mathbf{xy}\colon k = \mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon),\ \alpha(k, \delta)k < d\}$$

and observe that $A^{\epsilon}_{\mathbf{xy}} = \bigcup_k E_k$. Since if $|\mathbf{z} \cap \mathbf{x}| = k$, $\mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon) \geq k$, we have

$$E_k \subset G_k := \{\mathbf{xy}\colon |\mathbf{z} \cap \mathbf{x}| = k, \alpha(k, \delta)k < d\}$$

and by the union bound,

$$P^{2m}(A^{\epsilon}_{\mathbf{xy}}) \leq \sum_k P^{2m}(G_k) = \sum_{k:\alpha(k,\delta)k<d} P^{2m}\{\mathbf{xy}\colon |\mathbf{z} \cap \mathbf{x}| = k\}.$$

But $\alpha(k, \delta)k < d \Rightarrow 3k < d$ for all $\delta \in (0, 1)$. Thus by setting U to satisfy $\alpha(U, \delta)U = d = \alpha U$ we can write

$$P^{2m}(A^{\epsilon}_{\mathbf{xy}}) \leq \sum_{k=0}^{U} P_{d,k} \leq \sum_{k=0}^{U} \binom{d}{i} 2^{-d} \leq 2^{-d} \left(\frac{ed}{U}\right)^{U} = 2^{-\alpha U}(e\alpha)^{U},$$

where we have used Lemma 2.2. One can readily check that for all $\delta \in (0, 1)$ and all $U \in \mathbb{N}$, $2^{-\alpha(U,\delta)U}(e\alpha(U, \delta))^{U} \leq \delta$. Thus $P^{2m}(A^{\epsilon}_{\mathbf{xy}}) \leq \delta$, for all $\epsilon > 0$.

Part 3. Let $B^{\epsilon}_{\mathbf{xy}}$ be the event in the statement of the proposition. We will show that $B^{\epsilon}_{\mathbf{xy}} \subseteq A^{\epsilon/4}_{\mathbf{xy}}$ and thus $P^{2m}(B^{\epsilon}_{\mathbf{xy}}) \leq P^{2m}(A^{\epsilon/4}_{\mathbf{xy}}) \leq \delta$. We do this by showing that "$B^{\epsilon}_{\mathbf{xy}}$ is true" $\Rightarrow$ "$A^{\epsilon/4}_{\mathbf{xy}}$ is true". Now

$$\begin{aligned}
& 2\mathcal{N}(\epsilon, \mathcal{F}, \mathbf{x})2^{\alpha(U,\delta)U \log(17m) \log(5em/U)} < \mathcal{N}(\epsilon, \pi_{\epsilon}(\mathcal{F}), \mathbf{xy}) \\
\Rightarrow\ & 2\mathcal{N}(\epsilon, \mathcal{F}, \mathbf{x})2^{\alpha(U,\delta)U \log(17m) \log(5em/U)} < \mathcal{N}(\epsilon, \pi_{\epsilon}(\mathcal{F}), \mathbf{x})\mathcal{N}(\epsilon, \pi_{\epsilon}(\mathcal{F}), \mathbf{y}) \quad (6) \\
\Rightarrow\ & 2\mathcal{N}(\epsilon, \mathcal{F}, \mathbf{x})2^{\alpha(U,\delta)U \log(17m) \log(5em/U)} \\
& \qquad < \mathcal{N}(\epsilon, \pi_{\epsilon}(\mathcal{F}), \mathbf{x})2\left(\tfrac{4m(2.01\epsilon)^2}{\epsilon^2}\right)^{\mathrm{fat}_{\mathcal{F}_{\mathbf{xy}}}(\frac{\epsilon}{4}) \log\left(\frac{2em2.01\epsilon}{\epsilon \mathrm{fat}_{\mathcal{F}_{\mathbf{xy}}}(\frac{\epsilon}{4})}\right)} \quad (7)
\end{aligned}$$

$$\Rightarrow \quad 2\mathcal{N}(\epsilon,\mathcal{F},\mathbf{x})2^{\alpha(\mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon/4),\delta)\mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon/4)\log(17m)\log(5em/\mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon/4))}$$
$$< \mathcal{N}(\epsilon,\mathcal{F},\mathbf{x})2(17m)^{\mathrm{fat}_{\mathcal{F}_{\mathbf{xy}}}(\epsilon/4)\log(5em/\mathrm{fat}_{\mathcal{F}_{\mathbf{xy}}}(\epsilon/4))} \quad (8)$$
$$\Rightarrow \quad \alpha(\mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon/4),\delta)\mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon/4)\log(17m)\log(5em/\mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon/4)) + 1$$
$$< \mathrm{fat}_{\mathcal{F}_{\mathbf{xy}}}(\epsilon/4)\log(5em/\mathrm{fat}_{\mathcal{F}_{\mathbf{xy}}}(\epsilon/4))\log(17m) + 1 \quad (9)$$
$$\Rightarrow \quad \alpha(\mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon/4),\delta)\mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon/4) < \mathrm{fat}_{\mathcal{F}_{\mathbf{xy}}}(\epsilon/4) \quad (10)$$

where (6) follows from (4); (7) follows from the fact that $\mathrm{fat}_{\mathcal{F}_{\mathbf{y}}}(\epsilon/4) \leq \mathrm{fat}_{\mathcal{F}_{\mathbf{xy}}}(\epsilon/4)$, that the range of functions in $\pi_\epsilon(\mathcal{F})$ is an interval $[a,b]$ with $b-a = 2\epsilon$ and Corollary 2.5; (8) follows from (5) and the fact that $\mathcal{N}(\epsilon,\pi_\epsilon(\mathcal{F}),\mathbf{x}) \leq \mathcal{N}(\epsilon,\mathcal{F},\mathbf{x})$; (9) follows by dividing both sides of (8) by $\mathcal{N}(\epsilon,\mathcal{F},\mathbf{x})$ and taking logs; (10) follows from the fact that $\mathrm{fat}_{\mathcal{F}_{\mathbf{x}}}(\epsilon/4) \leq \mathrm{fat}_{\mathcal{F}_{\mathbf{xy}}}(\epsilon/4)$ and dividing out common terms on both sides. Now (10) defines the event $A^{\epsilon/4}_{\mathbf{xy}}$ as required. ∎

Lemma 4.3 *Suppose $\mathcal{F}$ is a sturdy set of functions that map from X to $\mathbb{R}$. Then for any distribution P on X, and any $U \in \mathbb{N}$ and any $\theta \in \mathbb{R}$*

$$P^{2m}\Big\{\mathbf{xy}\colon \Big(\exists f\in\mathcal{F},\ r=\max_j\{f(x_j)\},\ 2\gamma<\theta-r,\ \lfloor\log\mathcal{N}(\gamma/4,\mathcal{F},\mathbf{x})\rfloor = U,$$
$$\tfrac{1}{m}\left|\{i: f(y_i)\geq\theta\}\right| > \epsilon(m,k,\delta)\Big)\Big\} < \delta,$$

where $\epsilon(m,k,\delta) = \frac{1}{m}(U(\log\frac{5em}{U}\log(17m)\alpha(U,\delta/2)+1)+\log\frac{4}{\delta})$.

Proof Using the standard permutation argument (as in [24]), 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. Let

$$\gamma_U := \min\{\gamma'\colon \lfloor\log\mathcal{N}(\gamma'/4,\mathcal{F},\mathbf{x})\rfloor = U\}.$$

By Lemma 2.7 and the sturdiness of $\mathcal{F}$, the minimum is attained by some choice of 2^U functions from $\mathcal{F}$. The probability above is no greater than

$$P^{2m}\left\{\mathbf{xy} : \exists\gamma\in\mathbb{R}^+, \lfloor\log\mathcal{N}(\gamma'/4,\mathcal{F},\mathbf{x})\rfloor = U, \exists f\in\mathcal{F}, A_f(2\gamma_U)\right\},$$

where $A_f(\gamma)$ is the event that $f(y_i) > \max_j\{f(x_j)\}+\gamma$ for at least $m\epsilon(m,k,\delta)$ points y_i in $\mathbf{y}$. Note that $r+2\gamma<\theta$. Consider a minimal γ_U-cover $B_{\mathbf{xy}}$ of $\pi_{\gamma_U}(\mathcal{F})$ in the pseudo-metric $d_{\mathbf{xy}}$. In the remainder we will suppress the γ_U subscript to the function π to simplify the notation. We have that for any $f\in\mathcal{F}$, there exists $\tilde{f}\in B_{\mathbf{xy}}$, with $|\pi(f)(x)-\pi(\tilde{f})(x)|\leq\gamma_U$ for all $x\in\mathbf{xy}$. Thus since for all $x\in\mathbf{x}$, by the definition of r, $f(x)\leq r<\theta-2\gamma$, $\pi(f)(x)<\max\{\theta-2\gamma,\theta-2.005\gamma_U\}$, and so $\pi(\tilde{f})(x)<\theta-\gamma_U$. However there are at least $m\,\epsilon(m,k,\delta)$ points $y\in\mathbf{y}$ such that $f(y)\geq\theta>r+2\gamma$, so $\pi(\tilde{f})(y)>r+2\gamma-\gamma_U>\max_j\{\pi(\tilde{f})(x_j)\}$. 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

$$\begin{aligned} & P^{2m}\left\{\mathbf{xy} : \exists \gamma \in \mathbb{R}^+, \lfloor \log \mathcal{N}(\gamma'/4, \mathcal{F}, \mathbf{x}) \rfloor = U, \exists f \in \mathcal{F}, A_f(2\gamma_U)\right\} \\ \leq \; & E(|B_{\mathbf{xy}}|) 2^{-\epsilon(m,k,\delta)m}, \end{aligned}$$

where the expectation is over $\mathbf{xy}$ drawn according to P^{2m}. Hence, by Proposition 4.2 with probability at least $1 - \delta/2$

$$\begin{aligned} E(|B_{\mathbf{xy}}|) &\leq 2\mathcal{N}(\gamma_U, \mathcal{F}, \mathbf{x}) 2^{\alpha(U,\delta/2)U \log(17m) \log(5em/U)} \\ &\leq 2^{1+U[\log(17m)\log(5em/U)\alpha(U,\delta/2)+1]}. \end{aligned}$$

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

$$\epsilon(m,k,\delta) \geq \tfrac{1}{m}\left(U(1 + \log(5em/U)\log(17m)\alpha(U,\delta/2)) + \log\tfrac{4}{\delta}\right),$$

as required. ∎

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.

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 4.4 (Generalization Bounds via Empirical Covering Numbers) *Suppose $\mathcal{F}$ is a sturdy real valued function class. Fix $\theta \in \mathbb{R}$. With probability $1-\delta$ over m independently drawn examples $\mathbf{z}$, if $h = T_\theta(f) \in T_\theta(\mathcal{F})$ correctly classifies $\mathbf{z}$ then for all γ such that $\gamma < \min|f(x_i) - \theta|$, the expected error of h is bounded from above by*

$$\epsilon(m,U,\delta) = \frac{2}{m}\left(U\left(1 + \alpha(U,\delta/2)\log\left(\frac{5em}{U}\right)\log(17m)\right) + \log\left(\frac{16m}{\delta}\right)\right),$$

where $U = \lfloor \log \mathcal{N}(\gamma/8, \mathcal{F}, \mathbf{x}) \rfloor$.

Proof Making use of lemma 2.8 we will move to the double sample and stratify by U. By the union bound, it thus suffices to show that $\sum_{U=1}^{2m} P^{2m}(J_U) < \delta/2$, where

$$\begin{aligned} J_U = \{\mathbf{xy} \;:\; & \exists h = T_\theta(f) \in T_\theta(\mathcal{F}), \mathrm{Er}_{\mathbf{x}}(h) = 0, U = \lfloor \log \mathcal{N}(\gamma/8, \mathcal{F}, \mathbf{x}) \rfloor, \\ & \gamma < \min|f(x_i) - \theta|, \mathrm{Er}_{\mathbf{y}}(h) \geq m\epsilon(m,U,\delta)/2\}. \end{aligned}$$

(The largest value of U we need consider is $2m$, since for larger values the bound will in any case be trivial). It is sufficient if $P^{2m}(J_U) \leq \frac{\delta}{4m} = \delta'$. Consider $\hat{\mathcal{F}} = \widehat{T_\theta(\mathcal{F})}$. 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 $y \in \mathbf{y}$ to be misclassified, it must have $\hat{f}(\hat{y}) \geq \theta > \max\{\hat{f}(\hat{x}) : \hat{x} \in \hat{\mathbf{x}}\} + \gamma$, so that

$$J_k \subseteq \Big\{\hat{\mathbf{x}}\hat{\mathbf{y}} \in (X \times \{0,1\})^{2m} \;\; : \;\; \exists \hat{f} \in \hat{\mathcal{F}}, r = \max\{\hat{f}(\hat{x}) : \hat{x} \in \hat{\mathbf{x}}\}, \gamma < \theta - r,$$
$$U = \lfloor \log \mathcal{N}(\gamma/8, \mathcal{F}, \mathbf{x}) \rfloor, \left|\{\hat{y} \in \hat{\mathbf{y}} : \hat{f}(\hat{y}) \geq \theta\}\right| \geq m\epsilon(m, U, \delta)/2\Big\}.$$

Replacing γ by $\gamma/2$ in Lemma 4.3 and appealing to Lemma 2.8 we obtain $P^{2m}(J_U) \leq \delta'$, for

$$\epsilon(m, U, \delta) = \frac{2}{m}\left(U(1 + \alpha(U, \delta/2)\log(5em/U)\log(17m)) + \log(4/\delta')\right).$$

The condition of Lemma 2.8 is satisfied by this linking of ϵ and m. Substituting for δ' gives the result. ∎

Despite superficial appearances Theorem 4.4 is quite different from results obtained in [21]. For example, the bound involving the margin of a classifier given there relies on an *a priori* bound on the fat-shattering dimension for the whole class, not the fat-shattering dimension (or in our case the logarithm of the covering numbers) of the class restricted to the training set. The other result of [21] which is reminiscent of Theorem 4.4 involves bounding the generalization error in terms of the VC dimension of the set of hypotheses restricted to the training set. This result cannot take into account the margin of a large margin classifier, but refers to classical generalization bounds in terms of the VC dimension. The motivation for obtaining Theorem 4.4 came from recent work [26] on bounding $\sup_{\mathbf{x} \in X^m} \mathcal{N}(\epsilon, \mathcal{F}, \mathbf{x})$, where $\mathcal{F}$ is a class of SV machines, in terms of the eigenvalues of an integral operator induced by the kernel used in the SV machine. This in turn suggested one could compute $\mathcal{N}(\epsilon, \mathcal{F}, \mathbf{x})$ for a *given* $\mathbf{x}$ in terms of the eigenvalues of the Gram matrix. That is indeed the case as we will show in the next section. We are unaware of any calculation of the fat-shattering dimension for such classes that gives competitive bounds on the covering numbers. Thus even though the log covering numbers and fat shattering dimension can only differ by $\log(m)$ factors, the *bounds one has* on the quantities can differ significantly; that is certainly the current situation with regard to support vector machines.

5 Generalization Bounds from Eigenvalues of a Gram Matrix

In this section we will restrict consideration to linear functions of arbitrary input dimension. By the standard kernel trick this means that all of our reasoning is

applicable to support vector machines. Briefly, the kernel trick provides a method for computing dot products in feature spaces F nonlinearly related to the input space via some map $\Phi : X \to F$, using a kernel k, i.e. [4]

$$k(x, y) = \langle \Phi(x), \Phi(y) \rangle. \tag{11}$$

For instance if $X = \mathbb{R}^N$, the polynomial kernel

$$k(x, y) = (\langle x, y \rangle)^d$$

can be shown to correspond to a map Φ into the space spanned by all products of exactly d dimensions of $\mathbb{R}^N$. More generally, Mercer's theorem shows that kernels k of positive integral operators give rise to maps Φ such that (11) holds. We quote a version given in [11]:

Theorem 5.1 (Mercer) *If k is a continuous symmetric kernel of a positive integral operator T, i.e.*

$$(Tf)(y) = \int_X k(x, y) f(x)\, dx \tag{12}$$

with

$$\int_{X \times X} k(x, y) f(x) f(y)\, dx\, dy \geq 0 \tag{13}$$

for all $f \in L_2(X)$ (X being a compact subset of $\mathbb{R}^N$), it can be expanded in a uniformly convergent series (on $X \times X$) in terms of the eigenfunctions ψ_j of T and their positive eigenvalues λ_j:

$$k(x, y) = \sum_{j=1}^{N_F} \lambda_j \psi_j(x) \psi_j(y), \tag{14}$$

where $N_F \leq \infty$ is the number of nonzero eigenvalues.

Note that the eigenfunctions ψ_j corresponding to nonzero eigenvalues can be shown to be continuous [2, p.270]. From (14), it is straightforward to construct a map Φ into a potentially infinite-dimensional ℓ_2 space which satisfies (11). For instance, we may use

$$\Phi(x) = (\sqrt{\lambda_1}\psi_1(x), \sqrt{\lambda_2}\psi_2(x), \ldots). \tag{15}$$

SV Machines [4, 22, 18] exploit the kernel trick to estimate functions linear in F, taking the form

$$f(x) = \langle w, \Phi(x) \rangle + b \tag{16}$$

(regression estimation) or thresholded versions thereof (classification). Here, $w \in F$ is some weight vector. Using a regularizer of the form $\|w\|_F^2$ and suitable cost

functions, the training of SV Machines can be shown to reduce to a quadratic program in terms of the kernel Gram matrix $G_{ij} = [k(x_i, x_j)]_{i,j=1}^m$ where $x_1, \ldots, x_m$ are the training points. Moreover, the solution has a representation in terms of the training points,

$$w = \sum_{i=1}^{m} \alpha_i \Phi(x_i), \tag{17}$$

which together with (16) provides us with a kernel expansion for the estimated function.

It is part of the folklore of SV machines that the eigenvalues of the Gram matrix G should somehow influence the generalization performance of a SV machine. In this section we present a bound utilizing empirical covering numbers that shows this folklore is justified. The key trick is to find good bounds on the empirical covering number in terms of eigenvalues of the Gram matrix. We do this by using the machinery of entropy numbers of operators which is explained below. We also exploit the fact that by definition, the eigenvalues of the Gram matrix are the squares of the *singular values* of the data matrix $\mathbf{x}$ (see [12] e.g.).

We will take $\mathcal{F}$ to be the class of functions

$$\mathcal{F} = \left\{ x \mapsto \sum_{i=1}^{m} \alpha_i k(x, x_i) : \sum_{i=1}^{m} \sum_{j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \leq 1 \right\}. \tag{18}$$

Observe that from (17),

$$\|w\|_F^2 = \langle w, w \rangle = \left\langle \sum_{i=1}^{m} \alpha_i \Phi(x_i), \sum_{j=1}^{m} \alpha_j \Phi(x_j) \right\rangle = \sum_{i,j=1}^{m} \alpha_i \alpha_j k(x_i, x_j) \tag{19}$$

and thus the inequality in (18) is equivalent to $\|w\|_F \leq 1$.

The main result of this section is now stated. It gives a bound on the generalization error when using $\mathcal{F}$ in terms of the eigenvalues of the Gram matrix of the training points.

Theorem 5.2 (Generalization Bounds via Eigenvalues of Gram Matrix) *Let k be a Mercer kernel and $\mathcal{F}$ be defined by (18). For $m \in \mathbb{N}$ let $\mathbf{x} = (x_1, \ldots, x_m)$ where x_i $(i = 1, \ldots, m)$ are points in some input space X which is a compact subset of $\mathbb{R}^N$. Let $G = [k(x_i, x_j)]_{i,j=1}^m$ be the Gram matrix induced by k and $\mathbf{x}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the eigenvalues of G. Set $\lambda_{m+1} = 0$. Fix $\theta \in \mathbb{R}$. With probability $1 - \delta$ over m iid examples $x_1, \ldots, x_m$, if there exists $f^* \in \mathcal{F}$ such that $T_\theta(f^*)$ correctly classifies $x_1, \ldots, x_m$ and attains margin*

$$\gamma = \min_{i=1,\ldots,m} |f^*(x_i) - \theta| < 1382 \lambda_2 \log(m+1)$$

the expected error of $T_\theta(f^*)$ is bounded from above by

$$\epsilon(m, U, \delta) = \frac{2}{m} \left(U \left(1 + \alpha(U, \delta/2) \log \left(\frac{5em}{U} \right) \log(17m) \right) + \log \left(\frac{16m}{\delta} \right) \right)$$

where $\alpha(\cdot,\cdot)$ is as in Definition 4.1,

$$U = \frac{k^\star}{2\ln(2)} + \frac{k^\star}{2}\log\left(\frac{\mathfrak{a}(\lambda_1\cdots\lambda_{k^\star})^{1/k^\star}\log(m+1)}{(k^\star-1)\gamma^2}\right) + 2,$$

$$k^\star = \min\left\{k\in\mathbb{N}\colon \lambda_{k+1} \le \frac{k\gamma^2}{\mathfrak{a}\log(m+1)}\right\} \tag{20}$$

and $\mathfrak{a} = 11051$. In fact $k^\star \ge 2$.

Observe that

$$\epsilon(m,U,\delta) = O\left(\frac{U}{m}\log^2 m\log\frac{1}{\delta}\right).$$

The value of $\mathfrak{a}$ is not the best possible. Our guess for the best would be around $2\ln 2(1.86)^2 3^2 2^2 \approx 129$ if we could replace the 8 in Theorem 4.4 by 2, $C = 3$, and $c = 1.86$. Of course whilst it is to be desired it is not essential to obtain tight bounds on the constants in order to make use of them: see e.g. [8].

Proof The theorem follows from Theorem 4.4, Theorem 5.3 and Lemma 5.4 by observing that the choice of U implies $\log\mathcal{N}(\gamma/8,\mathcal{F},\mathbf{x}) \le U$ and that $\mathfrak{a} > 64\mathfrak{c}$. The fact that Theorem 4.4 requires γ is strictly less than the observed margin is handled by the fact that this inequality involving $\mathfrak{a}$ and $\mathfrak{c}$ is strict. The floor in the definition of U in Theorem 4.4 is dropped since $\epsilon(m,U,\delta)$ is increasing in U. The constant 1382 is $\lceil 8\mathfrak{c}\rceil$. ∎

Theorem 5.3 (Covering Numbers via Eigenvalues of Gram Matrix) *Let $\mathcal{F}$ be as in (18). Suppose $m\in\mathbb{N}$, and $\mathbf{x} = (x_1,\ldots,x_m)$ is an arbitrary sequence of m points in X. Define G, $\lambda_1,\ldots,\lambda_{m+1}$ as in Theorem 5.2. Let $\sigma_i = \sqrt{\lambda_i}$ for $i = 1,\ldots,m+1$. Let $\mathfrak{c} = 172.66$. Suppose $\epsilon < \sigma_2^2\mathfrak{c}\log(m+1)$. Then*

$$\log\mathcal{N}(\epsilon,\mathcal{F},\mathbf{x}) \le \frac{k^*}{2\ln(2)} + \frac{k^*}{2}\log\left(\frac{\mathfrak{c}(\sigma_1\cdots\sigma_{k^*})^{2/k^*}\log(m+1)}{(k^*-1)\epsilon^2}\right) + 2 \tag{21}$$

where

$$k^* := \min\left\{k\in\mathbb{N}\colon \sigma_{k+1}^2 \le \frac{k\epsilon^2}{\mathfrak{c}\log(m+1)}\right\} \tag{22}$$

and in fact $k^* \ge 2$.

The proof is given later in this section. The constant $\mathfrak{c}$ can quite likely be improved. See the remarks about c and C after Theorems 5.8 and 5.6.

Lemma 5.4 (Sturdiness of Mercer kernel machines) *Suppose $\mathcal{F}$ is given by (18), and k a Mercer kernel. Then $\mathcal{F}$ is sturdy.*

Proof We can write any $f \in \mathcal{F}$ as $f(x) = \langle \Phi(x), w \rangle$ and thus $\tilde{\mathbf{x}}(\mathcal{F})$ is a *linear* operator from $\mathcal{F}$ to $\mathbb{R}^m$. If the unit ball $U_F \ni w$ was compact we would be done. The trick is to replace U_F by a compact object with the same image as $\tilde{\mathbf{x}}(\mathcal{F})$. We will work with the mapping $\tilde{\mathbf{x}}' : w \to (\langle \Phi(x_1), w \rangle, \ldots, \langle \Phi(x_m), w \rangle)$ Observe $\tilde{\mathbf{x}}(\mathcal{F}) = \tilde{\mathbf{x}}(U_F)$ where F can be identified with ℓ_2.

Since the linear map $\tilde{\mathbf{x}}'$ has at most rank m we may decompose F orthogonally into the nullspace $F_\parallel$ of $\tilde{\mathbf{x}}'$ and its finite dimensional complement $F_\perp$. It can then be seen that $\tilde{\mathbf{x}}(\mathcal{F})$ is in fact the image of a compact linear operator applied to a finite dimensional unit ball and is therefore compact.

∎

5.1 Entropy Numbers

We will consider the mapping which takes a weight vector w to its value on the sample. This is the evaluation mapping $\tilde{\mathbf{x}}$ of Definition 2.6. We can view this mapping as being from ℓ_2 into ℓ_∞^m, by considering the ℓ_2 metric in weight space and the $d_{\mathbf{x}}$ metric in the image space. Bounding the covering numbers at scale ϵ amounts to calculating the number of ϵ-balls in ℓ_∞^m required to cover $\tilde{\mathbf{x}}(U_F)$, where U_F is the unit ball in $\mathcal{F}$. Thus $\mathcal{N}(\epsilon, \tilde{\mathbf{x}}(U_F)) = \mathcal{N}(\epsilon, \mathcal{F})$.

We will use results from [7] to bound the entropy numbers of this operator.

Suppose (X, d) is a normed space. The nth *entropy number* of a set $S \subset X$ is defined by

$$\epsilon_n(S) = \epsilon_n(S, d) := \inf\{\epsilon > 0 : \mathcal{N}(\epsilon, S, d) \leq n\}.$$

We denote by U_d the (closed) unit ball: $U_d := \{x \in X : \|x\| \leq 1\}$. If d is implicit from the context, we will sometimes write U_X. Suppose X and Y are normed spaces and T is a linear operator mapping from X to Y. Then the *operator norm* of T is defined by

$$\|T\| := \sup\{\|T\mathbf{x}\|_Y : \mathbf{x} \in U_d\}.$$

and T is *bounded* if $\|T\| < \infty$. We denote by $\mathcal{L}(X, Y)$ the set of all bounded linear operators from X to Y.

If $T \in \mathcal{L}(X, Y)$ the *entropy numbers of the operator* T are defined by

$$\epsilon_n(T) := \epsilon_n(T(U_X)), \qquad n \in \mathbb{N}.$$

The *dyadic entropy numbers* $e_n(T)$ are defined by

$$e_n(T) := \epsilon_{2^n - 1}(T) \qquad n \in \mathbb{N}$$

(This particular definition ensures $e_1 = \epsilon_1$.)

The *factorization theorem for entropy numbers* is extremely useful:

Lemma 5.5 *Let A, B, C be normed spaces and let $S, T \in \mathcal{L}(A, B)$ and $R \in \mathcal{L}(B, C)$. Then*

1. $\|T\| = e_1(T) \geq e_2(T) \geq \cdots \geq 0$.
2. $\forall k, l \in \mathbb{N}$, $e_{k+l-1}(RS) \leq e_k(R) e_l(S)$ *and* $\epsilon_{kl}(RS) \leq \epsilon_k(R)\epsilon_l(S)$.

The following theorem characterises to within a factor of 6 the entropy numbers of a diagonal operator. When working in ℓ_2 this also characterises the entropy numbers of any operator in terms of its eigenvalues since rotations can be performed at both ends with no cost (i.e. represent an arbitrary $T \in \mathcal{L}(\ell_2, \ell_2)$ by $T = S^{-1}DS$ where D is diagonal and S and S^{-1} are rotations and have norm 1 and then appeal to Lemma 5.5.)

Theorem 5.6 (Carl and Stephani, 1990 [6]) *Suppose $1 \leq p \leq \infty$ and let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_j \geq \ldots \geq 0$ be a non-increasing sequence of non-negative numbers and let*

$$D(\mathbf{x}) = (\sigma_1 x_1, \sigma_2 x_2, \ldots, \sigma_j x_j, \ldots)$$

for $\mathbf{x} = (x_1, x_2, \ldots, x_j, \ldots) \in \ell_p$ be the diagonal operator from ℓ_p into itself, generated by the sequence $(\sigma_j)_j$, where $1 \leq p \leq \infty$. Then for all $n \in \mathbb{N}$,

$$\sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}} \leq \epsilon_n(D) \leq C \sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}}, \tag{23}$$

where $C = 6$.

(6 is not the best possible value for C. Clearly it can not be less than 1. We believe a value of 3 is possible.)

The bound of Theorem 5.6 is worth analysing more closely.

Lemma 5.7 *Let $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_j \geq \ldots \geq 0$ be a non-increasing sequence of non-negative numbers. Let*

$$n_j := \frac{\prod_{i=1}^{j-1} \sigma_i}{\sigma_j^{j-1}} = \frac{\prod_{i=1}^{j} \sigma_i}{\sigma_j^{j}}. \tag{24}$$

Then $1 = n_1 \leq n_i \leq n_j$ for $i < j$ and for $n_k \leq n \leq n_{k+1}$,

$$\sigma_{k+1} \leq \sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}} = n^{-\frac{1}{k}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{\frac{1}{k}} \leq \sigma_k. \tag{25}$$

Proof To show that $n_k \leq n_{k+1}$, we compute the quotient

$$\frac{n_{k+1}}{n_k} = \frac{\sigma_k^{k-1} \prod_{i=1}^{k} \sigma_i}{\sigma_{k+1}^{k} \prod_{i=1}^{k-1} \sigma_i} = \frac{\sigma_k^k}{\sigma_{k+1}^k} \geq 1. \tag{26}$$

Next we show the parts of (25) not involving $\sup_j$. Assume $j > k$:

$$\sigma_j \le \sigma_{k+1} = \left(n_{k+1}^{-1} \prod_{i=1}^{k} \sigma_i \right)^{\frac{1}{k}} \le \left(n^{-1} \prod_{i=1}^{k} \sigma_i \right)^{\frac{1}{k}} . \tag{27}$$

Similarly for $j \le k$ one has

$$\sigma_j \ge \sigma_k = \left(n_k^{-1} \prod_{i=1}^{k} \sigma_i \right)^{\frac{1}{k}} \ge \left(n^{-1} \prod_{i=1}^{k} \sigma_i \right)^{\frac{1}{k}} \tag{28}$$

which, with (27) proves the right hand side of (25). Finally we have to show that the $\sup_j$ is obtained for $j = k$. Again, consider $j > k$ and observe that due to (27)

$$n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}} \le n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{\frac{1}{j}} n^{-\frac{j-k}{jk}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{\frac{j-k}{jk}} = n^{-\frac{1}{k}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{\frac{1}{k}}$$

where the first inequality follows by splitting the product $\sigma_1 \cdots \sigma_j$ into $\sigma_1 \cdots \sigma_k$ and $\sigma_{k+1} \cdots \sigma_j$ and then bounding the latter using (27) since we bounded $\sigma_{j'}$ by the right-hand side of (27) for all $j' > k$. In a similar fashion for $j < k$

$$n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}} \le n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{\frac{1}{j}} n^{\frac{k-j}{jk}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{-\frac{k-j}{jk}} = n^{-\frac{1}{k}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{\frac{1}{k}}$$

which followed by the expansion by $\sigma_{j+1} \ldots \sigma_k$ of the expression and exploiting the lhs of (25). Thus the $\sup_j$ is obtained for $j = k$ which concludes the proof. ∎

The other operator whose entropy numbers we must bound is $\mathrm{id}_{2,\infty}^m$ which is defined by

$$\begin{aligned} \mathrm{id}_{2,\infty}^m = \mathrm{id}\colon \ell_2^m &\to \ell_\infty^m \\ \mathrm{id}\colon x &\mapsto x \end{aligned}$$

Theorem 5.8 *For all $m \in \mathbb{N}$ and $k \in \mathbb{N}$,*

$$e_{k+1}(\mathrm{id}_{2,\infty}^m) \le \min\left(1, c\sqrt{\frac{\log(\frac{m}{k}+1)}{k}}\right),$$

where $c = 1.86$.

This result, without the explicit constant and the $\min(1,\ldots)$ is due to Schütt [19]. The $\min(1,\ldots)$ is an immediate consequence of Lemma 5.5 part 1 since $\|\mathrm{id}_{2,\infty}^m\| = 1$. The explicit value of c is determined in [27]. We believe it to be the best possible. One can restate the result in terms of ϵ_l. Setting $l = 2^k$, we have for $l \ge 2$

$$\epsilon_l(\mathrm{id}_{2,\infty}^m) \le \min\left(1, c\sqrt{\frac{\log(\frac{m}{\log(l)}+1)}{\log(l)}}\right). \tag{29}$$

Since $\epsilon_1(\mathrm{id}_{2,\infty}^m) = \|\mathrm{id}_{2,\infty}^m\| = 1$, in fact (29) holds for all $l \in \mathbb{N}$.

5.2 Proof of Theorem 5.3

As foreshadowed above we bound $\mathcal{N}(\epsilon, \mathcal{F})$ by bounding $\mathcal{N}(\epsilon, \tilde{\mathbf{x}}(U_F))$. We do that by bounding $\epsilon_n(\tilde{\mathbf{x}})$. Since $\tilde{\mathbf{x}}: F \to \ell_\infty^m$ and by (19) $\|w\|_F \leq 1$ we have $\epsilon_n(\tilde{\mathbf{x}}) = \epsilon(\mathcal{F})$. We identify F with $\ell_2^{N_F}$ where N_F may be infinite. We first decompose the multiple evaluation operator $\tilde{\mathbf{x}}$ into two operators,

$$\tilde{\mathbf{x}} = \mathrm{id}_{2,\infty}^m \circ \tilde{\mathbf{x}}_2,$$

where $\tilde{\mathbf{x}}_2$ is the multiple evaluation map with the metric of the output space $\mathbb{R}^m$ now taken to be ℓ_2^m, while $\mathrm{id}_{2,\infty}^m$ is the identity mapping on $\mathbb{R}^m$ with metric ℓ_2^m on the input and ℓ_∞^m on the output.

We will now decompose $\tilde{\mathbf{x}}_2$ into a sequence of three operators given by a singular value decomposition. This will allow us to bound the entropy numbers of $\tilde{\mathbf{x}}_2$ using the bound for diagonal operators of Theorem 5.6. The situation is summarized in the following diagram.

$$\begin{array}{ccccc} \ell_2^{N_F} & & \xrightarrow{\quad\tilde{\mathbf{x}}\quad} & & \ell_\infty^m \\ \Big\downarrow{\scriptstyle W_m^T} & & \searrow^{\tilde{\mathbf{x}}_2} & & \Big\uparrow{\scriptstyle \mathrm{id}} \\ \ell_2^m & \xrightarrow{\Sigma} & \ell_2^m & \xrightarrow{V} & \ell_2^m \end{array} \tag{30}$$

For the sake of adherence to usual notational conventions, it is convenient to use $\mathbf{X}$ to denote $\mathbf{x}$ considered as a matrix.

Lemma 5.9 *Let $\mathbf{X} = WSV^T$ be the singular value decomposition of the matrix $\mathbf{X}$ whose columns are the points of the training sample. Then we can write*

$$\tilde{\mathbf{x}}_2 = V \circ \Sigma \circ W_m^T,$$

where the mapping W_m consists of the first m columns of W and $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_m)$ is the leading $m \times m$ principal submatrix of S. Note that W_m^T is a mapping between ℓ_2^M and ℓ_2^m and the other two maps are between ℓ_2^m spaces. Furthermore the norms of W_m^T and V satisfy $\|W_m^T\| = \|V\| = 1$, while $\sigma_i = \sqrt{\lambda_i}$, where λ_i is the i-th eigenvalue of the Gram matrix $G = \mathbf{X}^T\mathbf{X}$. Thus (30) commutes.

Proof First observe that the evaluation mapping $\tilde{\mathbf{x}}_2$ can be written as

$$\tilde{\mathbf{x}}_2(w) = w^T\mathbf{X} = w^T WSV^T = w^T W_m \Sigma V^T = (V \circ \Sigma \circ W_m^T)(w).$$

Hence, the decomposition is shown. Since, W and V are unitary matrices, we have $\|V\| = 1$ and $\|W_m^T\| \leq 1$. Choosing the first column w_1 of W and observing that $\|W_m^T w_1\| = 1$ shows that equality also holds for the norm of W_m^T. Finally, observe that $G = V\Sigma^2 V^T = V\Lambda V^T$, and that the singular values σ_i are positive to prove the final assertion. ∎

Note that the strategy of this proof only makes sense since we have a *fixed* $\mathbf{X}$; if we required a result that held for all $\mathbf{X}$ subject to some condition, in order, for example, to compute $\sup_{\mathbf{X}\in X^m} \mathcal{N}(\epsilon, \mathcal{F}, \mathbf{X})$, an alternative strategy would be warranted. See [26] and [27] for examples of the calculation of such suprema suitable for traditional generalization error bounds.

Corollary 5.10 *For any* $l, k \in \mathbb{N}$

$$\epsilon_{lk}(\tilde{\mathbf{x}}) \leq \epsilon_l(\Sigma)\epsilon_k(\mathrm{id}_{2,\infty}^m).$$

Proof We apply Lemma 5.5 several times:

$$\epsilon_{lk}(\tilde{\mathbf{x}}) \leq \epsilon_1(W)\epsilon_l(\Sigma)\epsilon_1(V)\epsilon_k(\mathrm{id}_{2,\infty}^m).$$

Recalling that $\epsilon_1(A) = \|A\|$, for all operators A, and utilising that $\|W_m^T\| = \|V\| = 1$ we are done. ∎

Proof (Theorem 5.3) Lemma 5.7 allows us to prove Theorem 5.3

Part 1. First we show if we have formulas $\bar{\epsilon}(t, \Sigma)$ and $\bar{\epsilon}(t, \mathrm{id}_{2,\infty}^m)$ defined for $t \in \mathbb{R}^+$ such that $\bar{\epsilon}(t, \cdot)$ is nonincreasing in t,

$$\epsilon_t(\Sigma) \leq \bar{\epsilon}(t, \Sigma) \qquad t \in \mathbb{N},$$

and

$$\epsilon_t(\mathrm{id}_{2,\infty}^m) \leq \bar{\epsilon}(t, \mathrm{id}_{2,\infty}^m) \qquad t \in \mathbb{N}, \tag{31}$$

then

$$\epsilon_{4n}(\tilde{\mathbf{x}}) \leq \bar{\epsilon}(t, \Sigma)\bar{\epsilon}(n/t, \mathrm{id}_{2,\infty}^m) \quad \text{for all } n \in \mathbb{N} \text{ and } t \in [1, n]. \tag{32}$$

We have $\bar{\epsilon}(\lceil t \rceil, \Sigma) \leq \bar{\epsilon}(t, \Sigma)$ and $\bar{\epsilon}(\lceil n/t \rceil, \mathrm{id}_{2,\infty}^m) \leq \bar{\epsilon}(n/t, \mathrm{id}_{2,\infty}^m)$ and thus from Corollary 5.10,

$$\begin{aligned} \epsilon_{\lceil t \rceil \lceil n/t \rceil}(\tilde{\mathbf{x}}) &\leq \epsilon_{\lceil t \rceil}(\Sigma)\epsilon_{\lceil n/t \rceil}(\mathrm{id}_{2,\infty}^m) \quad t \in [1, n] \\ &\leq \bar{\epsilon}(\lceil t \rceil, \Sigma)\bar{\epsilon}(\lceil n/t \rceil, \mathrm{id}_{2,\infty}^m) \\ &\leq \bar{\epsilon}(t, \Sigma)\bar{\epsilon}(n/t, \mathrm{id}_{2,\infty}^m). \end{aligned} \tag{33}$$

But $\lceil t \rceil \lceil n/t \rceil \leq (t+1)(n/t+1) = n + t + n/t + 1 \leq 4n$ (since $t \in [1, n]$). Furthermore since $k > j \Rightarrow \epsilon_k \leq \epsilon_j$, we obtain

$$\epsilon_{4n}(\tilde{\mathbf{x}}) \leq \epsilon_{\lceil t \rceil \lceil n/t \rceil}(\tilde{\mathbf{x}}). \tag{34}$$

Combining (34) with (33) we conclude that (32) holds.

Part 2. Observe that with

$$\bar{\epsilon}(t, \mathrm{id}_{2,\infty}^m) := c\sqrt{\frac{\log(m+1)}{\log t}} \tag{35}$$

(31) holds. (For $t \in \mathbb{N}$ and $t \geq 2$, this follows immediately from (29); for $t = 1$, $\bar{\epsilon}(t, \mathrm{id}_{2,\infty}^m) = \infty$.)

Part 3. We substitute previously obtained bounds into (32), optimize for $t \in [1, n]$ and invert to bound $\log n$ in terms of ϵ.

Using (34), Corollary 5.10, (35), Theorem 5.6 and Lemma 5.7 we have for $n, m \in \mathbb{N}$, $t \in [1, n]$,

$$\epsilon_{4n} \leq cCt^{-1/k}(\sigma_1 \cdots \sigma_k)^{1/k}\sqrt{\frac{\log(m+1)}{\log(n/t)}} \tag{36}$$

where k is such that

$$n_k \leq t \leq n_{k+1} \tag{37}$$

and n_k is given by (24). Thus

$$\begin{aligned} \epsilon_{4n}^2 &\leq c^2C^2t^{-2/k}(\sigma_1 \cdots \sigma_k)^{2/k}\frac{\log(m+1)}{\log(n/t)} \\ \Rightarrow \quad \log n &\leq \frac{c^2C^2t^{-2/k}(\sigma_1 \cdots \sigma_k)^{2/k}\log(m+1)}{\epsilon_{4n}^2} + \log(t), \quad n_k \leq t \leq n_{k+1} \end{aligned} \tag{38}$$

Differentiate the RHS of (38) with respect to t and set to zero:

$$\begin{aligned} & \frac{-2}{k}\frac{c^2C^2(\sigma_1 \cdots \sigma_k)^{2/k}\log(m+1)}{t^{2/k+1}\epsilon_{4n}^2} + \frac{1}{t\ln(2)} = 0 \\ \Rightarrow \quad & \ln(2)t = \frac{\frac{k}{2}t^{2/k+1}\epsilon_{4n}^2}{c^2C^2(\sigma_1 \cdots \sigma_k)^{2/k}\log(m+1)} \\ \Rightarrow \quad & t^{2/k} = \frac{2\ln(2)c^2C^2(\sigma_1 \cdots \sigma_k)^{2/k}\log(m+1)}{k\epsilon_{4n}^2} =: \hat{t}_k^{2/k}. \end{aligned}$$

Substitute $\hat{t}_k^{2/k}$ into (38): If $t = \hat{t}_k \in [n_k, n_{k+1}]$,

$$\begin{aligned} \log n &\leq \frac{c^2C^2k\epsilon_{4n}^2(\sigma_1 \cdots \sigma_k)^{2/k}\log(m+1)}{2\ln(2)c^2C^2(\sigma_1 \cdots \sigma_k)^{2/k}\log(m+1)\epsilon_{4n}^2} + \log(\hat{t}_k) \\ &= \frac{k}{2\ln(2)} + \frac{k}{2}\log(\hat{t}_k^{2/k}) - 1 \\ &= \frac{k}{2\ln(2)} + \frac{k}{2}\log\left(\frac{2\ln(2)c^2C^2(\sigma_1 \cdots \sigma_k)^{2/k}\log(m+1)}{k\epsilon_{4n}^2}\right) \end{aligned} \tag{39}$$

provided $\hat{t}_k \in [n_k, n_{k+1}]$.

Part 4. We note some implications following from $\hat{t}_k \in [n_k, n_{k+1}]$.

a) Observe that

$$\begin{aligned}
& \hat{t}_k \leq n_{k+1} \\
\Leftrightarrow \quad & \hat{t}_k^{2/k} \leq n_{k+1}^{2/k} = \left(\frac{\sigma_1 \cdots \sigma_k}{\sigma_{k+1}^k}\right)^{2/k} = \frac{(\sigma_1 \cdots \sigma_k)^{2/k}}{\sigma_{k+1}^2} \\
\Leftrightarrow \quad & \frac{2\ln(2)c^2C^2(\sigma_1 \cdots \sigma_k)^{2/k}\log(m+1)}{k\epsilon_{4n}^2} \leq \frac{(\sigma_1 \cdots \sigma_k)^{2/k}}{\sigma_{k+1}^2} \qquad (40) \\
\Leftrightarrow \quad & \sigma_{k+1}^2 \leq \frac{k\epsilon_{4n}^2}{2\ln(2)c^2C^2\log(m+1)}. \qquad (41)
\end{aligned}$$

b) Also observe that

$$\begin{aligned}
& n_k \leq \hat{t}_k \\
\Leftrightarrow \quad & n_k^{2/k} \leq \frac{2\ln(2)c^2C^2(\sigma \cdots \sigma_k)^{2/k}\log(m+1)}{k\epsilon_{4n}^2} \\
\Leftrightarrow \quad & \sigma_k^2 \geq \frac{k\epsilon_{4n}^2}{2\ln(2)c^2C^2\log(m+1)}.
\end{aligned}$$

Part 5. Choose the minimum $k \in \mathbb{N}$ such that (41) is satisfied. Denote this value k^*. Thus k^* is defined by (22) (with $\epsilon = \epsilon_{4n}$). Observe that by assumption $k^* \geq 2$. From the sequence of equivalences leading to (41), we conclude

$$\hat{t}_{k^*} \leq n_{k^*+1}. \qquad (42)$$

Part 6. We now consider the two possibilities that $\hat{t}_{k^*} \geq n_{k^*}$ and $\hat{t}_{k^*} < n_{k^*}$.

a) If $\hat{t}_{k^*} \geq n_{k^*}$ then (42) implies $\hat{t}_{k^*} \in [n_{k^*}, n_{k^*+1}]$ and so by (39), we obtain

$$\log n \leq \frac{k^*}{2\ln 2} + \frac{k^*}{2}\log\left(\frac{2\ln(2)c^2C^2(\sigma_1 \cdots \sigma_{k^*})^{2/k^*}\log(m+1)}{k^*\epsilon_{4n}^2}\right). \qquad (43)$$

b) If $\hat{t}_{k^*} < n_{k^*}$ then we will in fact choose $t = n_{k^*} = \frac{\sigma_1 \cdots \sigma_{k^*}}{\sigma_{k^*}^{k^*}}$ and so we can use (38) to obtain

$$\begin{aligned}
\log n &\leq \frac{c^2C^2(\sigma_1 \cdots \sigma_{k^*})^{2/k^*}\log(m+1)}{n_{k^*}^{2/k^*}\epsilon_{4n}^2} + \log(n_{k^*}) - 1 \\
&= \frac{c^2C^2(\sigma_1 \cdots \sigma_{k^*})^{2/k^*}\log(m+1)\sigma_{k^*}^2}{(\sigma_1 \cdots \sigma_{k^*})^{2/k^*}\epsilon_{4n}^2} + \log\left(\frac{\sigma_1 \cdots \sigma_{k^*}}{\sigma_{k^*}^{k^*}}\right) \\
&= \frac{c^2C^2\log(m+1)\sigma_{k^*}^2}{\epsilon_{4n}^2} + \frac{k^*}{2}\log\left(\frac{(\sigma_1 \cdots \sigma_{k^*})^{2/k^*}}{\sigma_{k^*}^2}\right). \qquad (44)
\end{aligned}$$

Now by definition of k^*,

$$\sigma_{k^*}^2 > \frac{(k^*-1)\epsilon_{4n}^2}{2\ln(2)c^2C^2\log(m+1)}. \tag{45}$$

Furthermore our current assumption

$$\begin{aligned}
& \hat{t}_{k^*} < n_{k^*} \\
\Rightarrow \quad & \hat{t}_{k^*}^{2/k^*} < n_{k^*}^{2/k^*} \\
\Rightarrow \quad & \frac{2\ln(2)c^2C^2(\sigma_1\cdots\sigma_{k^*})^{2/k^*}\log(m+1)}{k^*\epsilon_{4n}^2} < \frac{(\sigma_1\cdots\sigma_{k^*})^{2/k^*}}{\sigma_{k^*}^2} \\
\Rightarrow \quad & \sigma_{k^*}^2 < \frac{k^*\epsilon_{4n}^2}{2\ln(2)c^2C^2\log(m+1)}.
\end{aligned} \tag{46}$$

Now we use the bounds (45) and (46) in (44) (2nd and 1st occurrence of $\sigma_{k^*}^2$ respectively) to obtain

$$\begin{aligned}
\log n \quad \leq \quad & \frac{c^2C^2\log(m+1)k^*\epsilon_{4n}^2}{\epsilon_{4n}^2 2\ln(2)c^2C^2\log(m+1)} \qquad (47) \\
& + \frac{k^*}{2}\log\left(\frac{(\sigma_1\cdots\sigma_{k^*})^{2/k^*}2\ln(2)c^2C^2\log(m+1)}{(k^*-1)\epsilon_{4n}^2}\right) \\
= \quad & \frac{k^*}{2\ln(2)} + \frac{k^*}{2}\log\left(\frac{(\sigma_1\cdots\sigma_{k^*})^{2/k^*}2\ln(2)c^2C^2\log(m+1)}{(k^*-1)\epsilon_{4n}^2}\right) \qquad (48)
\end{aligned}$$

Summarising cases a) and b) we have: If $\hat{t}_{k^*} \geq n_{k^*}$ then (43); If $\hat{t}_{k^*} < n_{k^*}$ then (48). Since (48) $\geq$ the RHS of (43) always, we conclude that (48) holds in both cases.

Observe that (48) is trivial if $k^* = 1$ and that

$$\begin{aligned}
k^* \geq 2 \quad & \Leftrightarrow \quad \sigma_2^2 > \frac{\epsilon^2}{2\ln(2)c^2C^2\log(m+1)} \\
& \Leftrightarrow \quad \epsilon < \sigma_2^2 2\ln(2)c^2C^2\log(m+1).
\end{aligned} \tag{49}$$

Part 7. Finally, let $N = 4n$ and so $n = N/4$ and $\log(N/4) \leq A \Rightarrow \log(N) \leq A + 2$. Hence with $\epsilon = \epsilon_N$ and $\mathfrak{c}$ set to $2\ln(2)c^2C^2$ with numerical values substituted for c and C, (21) holds.

∎

6 Conclusions

This paper has presented a method by which bounds on the covering numbers of a function class on a training set can be used to bound the generalization error of the resulting classifier. In the previous section we have shown how this method can then be used to derive alternative bounds on the generalization error derived from observed properties of the margin and inner product matrix of a Support Vector Machine.

Improved bounds can be used to guide more refined Structural Risk Minimization over choices of different kernels for example, or model selection. Hence, the approach developed here may well have applications in practical learning systems. Our hope is that these methods may also be able to give bounds that are more realistic than previous PAC estimates.

Finally we remark that recently a very nice result concerning the stability of the VC dimension on a random sample has been obtained using concentration of measure techniques [5]. We conjecture that one may be able to obtain refinements of the results of the present paper concerning empirical covering numbers using those techniques.

References

[1] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale sensitive dimensions, uniform convergence and learnability. *Journal of the ACM*, 44(4):615–631, 1997.

[2] Robert Ash. *Information Theory*. Interscience Publishers, New York, 1965.

[3] P. L. Bartlett. The sample complexity of pattern classsification with neural networks: the size of the weights is more important than the size of the network. *IEEE Trans. Information theory*, 44(2):525–536, 1998.

[4] B. E. Boser, I. M. Guyon, and V. N. Vapnik. A training algorithm for optimal margin classifiers. In D. Haussler, editor, *5th Annual ACM Workshop on COLT*, pages 144–152, Pittsburgh, PA, 1992. ACM Press.

[5] S. Boucheron, G. Lugosi, and P. Massart. A sharp concentration inequality with applications. Preprint, CNRS-Université Paris-Sud, April 1999.

[6] B. Carl and I. Stephani. *Entropy, compactness, and the approximation of operators*. Cambridge University Press, Cambridge, UK, 1990.

[7] Bernd Carl and Irmtraud Stephani. *Entropy, Compactness and the Approximation of Operators*. Cambridge University Press, Cambridge, 1990.

[8] V. Cherkassky, X. Shao, F.M. Mulier, and V.N. Vapnik. Model complexity control for regression using VC generalization bounds. *IEEE Transactions on Neural Networks*, 10(5):1075–1089, 1999.

[9] C. Cortes and V. Vapnik. Support vector networks. *M. Learning*, 20:273 – 297, 1995.

[10] L. Devroye, L. Györfi, and G. Lugosi. *A Probabilistic Theory of Pattern Recognition*. Number 31 in Applications of mathematics. Springer, New York, 1996.

[11] N. Dunford and J. T. Schwartz. *Linear Operators Part II: Spectral Theory, Self Adjoint Operators in Hilbert Space*. Number VII in Pure and Applied Mathematics. John Wiley & Sons, New York, 1963.

[12] Roger A. Horn and Charles R. Johnson. *Matrix Analysis*. Cambridge University Press, Cambridge, 1992.

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

[14] A. N. Kolmogorov and S. V. Fomin. *Introductory Real Analysis*. Dover, 1970.

[15] G. Lugosi and M. Pinter. A data-dependent skeleton estimate for learning. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pages 51–58, New York, 1996. Association for Computing Machinery.

[16] A. Ruiz. Practical prediction of generalization by the method of random falsifiability. Technical Report DIS–9–97, Department of Computer Science, University of Murcia, Spain, 1997.

[17] A. Ruiz and P. E. López de Teruel. Random falsifiability and support vector machines. In *Proceedings of Learning 98*, Getafe, 1998.

[18] B. Schölkopf, C. J. C. Burges, and A. J. Smola. *Advances in Kernel Methods — Support Vector Learning*. MIT Press, Cambridge, MA, 1999.

[19] C. Schütt. Entropy numbers of diagonal operators between symmatric Banach spaces. *Journal of Approximation Theory*, 40:121–128, 1984.

[20] J. Shawe-Taylor, P. Bartlett, R. Williamson, and M. Anthony. A framework for structural risk minimization. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pages 68–76, New York, 1996. Association for Computing Machinery.

[21] J. Shawe-Taylor, P. L. Bartlett, R. C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.

[22] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer, N.Y., 1995.

[23] V. N. Vapnik. *Estimation of Dependences Based on Empirical Data*. Springer-Verlag, Berlin, 1982.

[24] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probab. and its Applications*, 16(2):264–280, 1971.

[25] Vladimir Vapnik, Esther Levin, and Yann Le Cun. Measuring the VC-dimension of a learning machine. *Neural Computation*, 6(5):851–876, 1994.

[26] R. Williamson, A. Smola, and B. Schölkopf. Entropy numbers, operators and support vector kernels. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 127–144, Cambridge, MA, 1999. MIT Press.

[27] R.C. Williamson, A.J. Smola, and B. Schölkopf. A maximum margin miscellany. Preprint, 1998-9.