Generalization Bounds via Eigenvalues of the Gram Matrix

Bernhard Schölkopf, GMD [1]
John Shawe-Taylor, University of London [2]
Alexander J. Smola, GMD [3]
Robert C. Williamson, ANU [4]

NeuroCOLT2 Technical Report Series

NC2-TR-1999-035

March, 1999[5]

Produced as part of the ESPRIT Working Group
in Neural and Computational Learning II,
NeuroCOLT2 27150

For more information see the NeuroCOLT website
`http://www.neurocolt.com`
or email `neurocolt@neurocolt.com`

[1]bs@first.gmd.de GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
[2]jst@dcs.rhbnc.ac.uk Department of Computer Science, Royal Holloway College, University of London, Egham, TW20 0EX, UK
[3]smola@first.gmd.de GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
[4]bob.williamson@anu.edu.au Deptartment of Engineering, Australian National University, 0200 Canberra, Australia
[5]Received 26-MAR-99

Abstract

Model selection in Support Vector machines is usually carried out by minimizing the quotient of the radius of the smallest enclosing sphere of the data and the observed margin on the training set. We provide a new criterion taking the *distribution* within that sphere into account by considering the Gram matrix of the data. In particular, this makes use of the eigenvalue distribution of the matrix. Experimental results on real world data show that this new criterion provides a good prediction of the shape of the curve relating generalization error to kernel width.

1 Introduction

Support Vector (SV) machines traditionally carry out model selection by minimizing the ratio between the radius of the smallest sphere enclosing the data in feature space R and the width of the margin $2/\|w\|$ since this corresponds to a classifier with minimal fat shattering dimension [5]. Whilst in general capturing the correct scaling behaviour in terms of the weight vector w, this approach has several shortcomings. In particular, it completely ignores the information about the *distribution* of the data inside the sphere. In other words — data completely filling the sphere and data restricted to a small cigar–shaped region would lead to identical bounds: the largest variation in any direction determines the bound. This is clearly not what one wants or expects. We provide new bounds taking the distribution of the data in feature space into account by effectively performing Kernel PCA [6] and show that these results are superior to the traditional bounds when it comes to model selection on real world datasets.

2 Background Results

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 $t(\mathbf{X}) = (t(x_1), \ldots, t(x_m))$. We use $\mathrm{Er}_{\mathbf{z}}(h) = |\{i : h(x_i) \neq t(x_i)\}|$ to denote the *number* of errors that some decision function h makes on $\mathbf{z}$, and $\mathrm{er}_P(h) = P\{x : h(x) \neq t(x)\}$ to denote the *expected error* when x is drawn according to P. We give the definition of the fat-shattering dimension, which was first introduced in [4], 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 $r = (r_x)_{x \in \mathbf{X}}$ *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*

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

The fat-shattering dimension $\mathrm{fat}_{\mathcal{F}}$ *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.*

The fat-shattering dimension of linear functions is bounded by the following result (which is a refinement of a result in [11]).[1]

Theorem 2.2 ([2]) *Let X be a ball of radius R in an inner product space and let*

$$\mathcal{F} = \{x \mapsto \langle w, x\rangle : \|w\| \le 1, x \in X\}. \tag{2}$$

Then for all $\gamma > 0$,

$$\mathrm{fat}_{\mathcal{F}}(\gamma) \le \left(\frac{R}{\gamma}\right)^2. \tag{3}$$

Before we can quote the next lemma, we need two more definitions.

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 $d(a, b) \le \epsilon$. The* ϵ-covering number *of A, $\mathcal{N}_d(\epsilon, A)$, is the minimal cardinality of an ϵ-cover[2] for A (if there is no such finite cover then it is defined to be ∞). We will say the cover is proper if $B \subseteq A$.*

Definition 2.4 (Covering Numbers in ℓ_∞^m) *Let $\mathcal{F}$ be a class of real-valued functions on the space X. For any $m \in \mathbb{N}$ and $\mathbf{X} \in X^m$, we define the pseudo-metric*

$$d_{\mathbf{X}}(f, g) := \max_{1 \le i \le m} |f(x_i) - g(x_i)|. \tag{4}$$

This is referred to as the ℓ_∞^m distance over a finite sample $\mathbf{X} = (x_1, \ldots, x_m)$. 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 $\mathcal{N}(\epsilon, \mathcal{F}, \mathbf{X}) = \mathcal{N}_{\ell_\infty^m}(\epsilon, \mathcal{F}_{\mathbf{X}})$, i.e. it is the ℓ_∞^m covering number of

$$\mathcal{F}_{\mathbf{X}} := \{(f(x_1), \ldots, f(x_m)) \colon f \in \mathcal{F}\}, \tag{5}$$

the class $\mathcal{F}$ restricted *to the sample $\mathbf{X}$. It is the* observed *covering number on $\mathbf{X}$.*

We now quote a lemma from [9] which follows directly from a result of Alon *et al.* [1]. It is a growth-function type bound (cf. [11]), i.e. distribution-independent, involving a sup over the domain X.

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

$$\sup_{\mathbf{X} \in X^m} \mathcal{N}(\epsilon, \mathcal{F}, \mathbf{X}) \le 2\left(\frac{4m(b-a)^2}{\epsilon^2}\right)^{d \log\left(\frac{2em(b-a)}{d\epsilon}\right)}. \tag{6}$$

[1] 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 is useful in the technical detail of the proof of the result in [10] which we exploit in the present paper, but also ensures that the definition reduces to the Pollard dimension for $\gamma = 0$.

[2] We have used a somewhat unconventional less than or equal to in the definition of a cover, as this is technically useful in some of the proofs underlying the present results [10].

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 of the evaluation operator.

Definition 2.6 (Evaluation Operator) *Let*

$$\begin{aligned} S_{\mathbf{X}}\colon \mathcal{F} &\longrightarrow \mathbb{R}^m, \\ f &\mapsto (f(x_1), f(x_2), \ldots, f(x_m)) \end{aligned} \tag{7}$$

denote the multiple evaluation map induced by $\mathbf{X} = (x_1, \ldots, x_m) \in X^m$. *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* $S_{\mathbf{X}}(\mathcal{F})$ *of* $\mathcal{F}$ *under* $S_{\mathbf{X}}$ *is a compact subset of* $\mathbb{R}^m$.

3 Covering Numbers on a Double Sample

We begin by presenting a key proposition from [10] that shows with high probability the covering numbers on a sample provide a good estimate of the covering numbers on a double sample.

Definition 3.1 *For* $U \in \mathbb{N}$ *and* $\delta \in \mathbb{R}^+$, *we define the function*

$$\alpha(U, \delta) := 3.08 \left(1 + \tfrac{1}{U} \ln \tfrac{1}{\delta}\right). \tag{8}$$

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

Theorem 3.2 *Consider a sturdy real valued function class* $\mathcal{F}$. *Fix* $\theta \in \mathbb{R}$. *If a learner correctly classifies* m *independently generated examples* $\mathbf{z}$ *with* $h = T_\theta(f) \in T_\theta(\mathcal{F})$ *then for all* γ *such that* $\gamma < \min|f(x_i) - \theta|$, *with confidence* $1 - \delta$ *the expected error of* h *is bounded from above by*

$$\epsilon(m, U, \delta) = \tfrac{2}{m}\Big(U\Big(1 + \alpha\left(U, \tfrac{\delta}{2}\right) \log\left(\tfrac{5em}{U}\right) \log(17m)\Big) + \log\left(\tfrac{16m}{\delta}\right)\Big), \tag{9}$$

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

Here, $\lfloor x \rfloor$ denotes the floor function which returns the largest integer $\leq x$.

4 Main Result

In this section we will restrict consideration to linear functions of arbitrary dimension. By the standard kernel trick [11, 6] this means that all of our reasoning is applicable to support vector machines.

It is part of the folklore of SV machines that the eigenvalues of the empirical Gram matrix (or kernel matrix, defined below) should somehow influence the generalization performance of a SV machine (e.g. an algorithm involving the latter was proposed in [13]). 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 will take $\mathcal{F}$ to be the class of functions defined on a space X (which we identify with a subset of ℓ_2^M where M may be infinite) via

$$\mathcal{F} = \{x \mapsto \langle w, x\rangle : \|w\|_{\ell_2^M} \leq 1\}. \tag{10}$$

Note that if X is compact, then the sturdiness of $\mathcal{F}$ follows directly from its continuity. In the case of SV machines, which is what we are ultimately interested in, compactness of X is usually required for Mercer's theorem to hold.

The main result of this section is as follows. It gives a bound on the generalization error in terms of the eigenvalues of the Gram matrix of the training points.

Theorem 4.1 *Let $\mathcal{F}$ be the set of linear functions in some feature space defined in (10). For $m \in \mathbb{N}$ let $\mathbf{X} = (x_1, \ldots, x_m)$ and let $G = \mathbf{X}^\top \mathbf{X}$ be the Gram matrix induced by $\mathbf{X}$. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m$ be the eigenvalues of G. Set $\lambda_{m+1} = 0$. Let*

$$k(n) := \min\left\{j : j \in \{1, \ldots, m\}, \lambda_{j+1}^j \leq \left(\tfrac{\lambda_1 \cdots \lambda_j}{n^2}\right)\right\} \tag{11}$$

and for $n = 2, 3, \ldots,$ let

$$\Upsilon(n) := 6 \min_{j \in \{1, \ldots n-1\}} 2^{\frac{1-j}{k(j)}} (\lambda_1 \cdots \lambda_{k(j)})^{\frac{1}{2k(j)}} \cdot \min\left(1, 1.86\sqrt{\frac{\log(\frac{m}{n-j}+1)}{n-j}}\right) \tag{12}$$

Fix $\theta \in \mathbb{R}$. Suppose there exists $f^* \in \mathcal{F}$ such that $T_\theta(f^*)$ correctly classifies $x_1, \ldots, x_m$. Denote by γ the margin on the training set, i.e.

$$\gamma := \min_{i=1,\ldots,n} |f^*(x_i) - \theta|. \tag{13}$$

Let

$$U = \min\{n \in \mathbb{N} : \Upsilon(n) \leq \gamma/8.001\}.$$

Then with confidence $1 - \delta$ the expected error of $T_\theta(f^*)$ is bounded from above by

$$\epsilon(m, U, \delta) = \tfrac{2}{m}\left(U\left(1 + \alpha\left(U, \tfrac{\delta}{2}\right) \log\left(\tfrac{5em}{U}\right) \log(17m)\right) + \log\left(\tfrac{16m}{\delta}\right)\right), \tag{14}$$

where $\alpha(\cdot,\cdot)$ is as in Definition 3.1.

The proof of this result follows from Theorem 4.8 below and Theorem 3.2 by observing that the choice of U implies $\log \mathcal{N}(\gamma/8.0005, \mathcal{F}, \mathbf{X}) \leq U - 1$ and so certainly
$\lfloor \log \mathcal{N}(\gamma/8.0005, \mathcal{F}, \mathbf{X}) \rfloor \leq U$. Note that the result depends only on inner products of the input vectors and so can be applied to Support Vector Machines, where the inner product in a feature space is given by a kernel function $k(x, y)$ in the input space. Thus $\mathbf{X}$ corresponds to the points in feature space. In that case one can explicitly compute the Gram matrix via

$$G = [k(x_i, x_j)]_{i,j=1}^m. \tag{15}$$

We will consider the mapping which takes a weight vector w to its value on the sample. This is the evaluation mapping $S_{\mathbf{x}}$ 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 of $\mathcal{F}_{\mathbf{X}}$ at scale ϵ amounts to calculating the number of ϵ-balls in ℓ_∞^m required to cover $S_{\mathbf{x}}(U_{\ell_2})$, where U_{ℓ_2} is the unit ball in $\mathcal{F}$.

We will use results from [3] to bound the entropy numbers of the operator $S_{\mathbf{x}}$. Suppose (X, d) is a normed space. The nth *entropy number* of a set $S \subset X$ is defined by

$$\epsilon_n(S) = \epsilon_n(S, d) := \inf\{\epsilon > 0 \colon \mathcal{N}(\epsilon, S, d) \leq n\}. \tag{16}$$

We denote by U_d the (closed) unit ball: $U_d := \{x \in X \colon d(x, 0) \leq 1\}$. Suppose X and Y are Banach spaces and T is a linear operator mapping from X to Y. Then the *operator norm* of T is defined by

$$\|T\| := \sup\{\|Tx\|_Y \colon \|x\|_X \leq 1\}. \tag{17}$$

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

$$\epsilon_n(T) := \epsilon_n(T(U_X)). \tag{18}$$

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

$$e_n(T) := \epsilon_{2^{n-1}}(T) \text{ with } n \in \mathbb{N} \tag{19}$$

(This particular definition ensures $e_1 = \epsilon_1$.) The *factorization theorem for entropy numbers* is extremely useful:

Lemma 4.2 (Carl and Stephani [3]) *Let A, B, C be Banach 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. $\forall k, l \in \mathbb{N},\ e_{k+l-1}(RS) \leq e_k(R) e_l(S).$

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.

Theorem 4.3 (Carl and Stephani [3]) *Suppose $1 \leq p \leq \infty$. 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*

$$D \colon \ell_p \to \ell_p \text{ with } D\mathbf{x} = (\sigma_1 x_1, \sigma_2 x_2, \ldots, \sigma_j x_j, \ldots) \tag{20}$$

for $\mathbf{x} = (x_1, x_2, \ldots, x_j, \ldots) \in \ell_p$ be the diagonal operator generated by the sequence $(\sigma_j)_j$. Then for all $n \in \mathbb{N}$,

$$\sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}} \leq \epsilon_n(D) \leq 6 \sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}}. \tag{21}$$

An alternative form of this result in terms of the dyadic entropy numbers is convenient for us. The upper bound becomes: for all $n \in \mathbb{N}$,

$$e_n(D) \leq 6 \sup_{j \in \mathbb{N}} 2^{-\frac{n-1}{j}} (\sigma_1 \cdots \sigma_j)^{1/j}. \tag{22}$$

The bound of Theorem 4.3 is worth analysing more closely. The following lemma shows for which value of j the sup is obtained.

Lemma 4.4 *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 $n \in \mathbb{N}$. Let*

$$n_j := \frac{\prod_{i=1}^{j-1} \sigma_i}{\sigma_j^{j-1}} = \frac{\prod_{i=1}^{j} \sigma_i}{\sigma_j^{j}}. \tag{23}$$

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

$$\sigma_{k+1} \leq \sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}} = n^{-\frac{1}{k}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{\frac{1}{k}} \leq \sigma_k. \tag{24}$$

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

$$\frac{n_{k+1}}{n_k} = \frac{\sigma_k^{k-1} \prod_{i=1}^{k} \sigma_i}{\sigma_{k+1}^{k} \prod_{i=1}^{k-1} \sigma_i} = \frac{\sigma_k^k}{\sigma_{k+1}^k} \geq 1. \tag{25}$$

Next we show the parts of (24) not involving $\sup_j$. Assume $j > k$:

$$\sigma_j \leq \sigma_{k+1} = \left(n_{k+1}^{-1} \prod_{i=1}^{k} \sigma_i \right)^{\frac{1}{k}} \leq \left(n^{-1} \prod_{i=1}^{k} \sigma_i \right)^{\frac{1}{k}} \tag{26}$$

Similarly for $j \leq k$ one has

$$\sigma_j \geq \sigma_k = \left(n_k^{-1} \prod_{i=1}^{k} \sigma_i \right)^{\frac{1}{k}} \geq \left(n^{-1} \prod_{i=1}^{k} \sigma_i \right)^{\frac{1}{k}} \tag{27}$$

which, with (26) proves the right hand side of (24). Finally we have to show that the $\sup_j$ is obtained for $j = k$. Again, consider $j > k$ and observe that due to (26)

$$\begin{aligned} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}} &\leq n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{\frac{1}{j}} n^{-\frac{j-k}{jk}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{\frac{j-k}{jk}} \\ &= n^{-\frac{1}{k}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{\frac{1}{k}} \end{aligned}$$

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 (26) since we bounded $\sigma_{j'}$ by the right-hand side of (26) for all $j' > k$. In a similar fashion for $j < k$

$$\begin{aligned} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}} &\leq n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{\frac{1}{j}} n^{\frac{k-j}{jk}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{-\frac{k-j}{jk}} \\ &= n^{-\frac{1}{k}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{\frac{1}{k}} \end{aligned}$$

which followed by the expansion by $\sigma_{j+1} \ldots \sigma_k$ of the expression and exploiting (27). Thus the $\sup_j$ is obtained for $j = k$ which concludes the proof. ∎

In order to compute the bound in (21), we need only find a k such that $n_k \leq n \leq n_{k+1}$. To obtain the best bound we want the smallest k such that $n \leq n_{k+1} = \frac{\sigma_1 \cdots \sigma_k}{\sigma_{k+1}^k}$ or equivalently $\sigma_{k+1} \leq \left(\frac{\sigma_1 \cdots \sigma_k}{n}\right)^{1/k}$. One can show that for any $n \in \mathbb{N}$ such a k always exists[3] for any $(\sigma_j)_j$ such that $\lim_{j \to \infty} \sigma_j = 0$. This justifies the definition of $k(n)$ in (11).

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

$$\begin{array}{rcl} \mathrm{id}_{2,\infty}^m \colon \ell_2^m & \to & \ell_\infty^m \\ x & \mapsto & x \end{array} \tag{28}$$

Theorem 4.5 (Entropy Numbers of the Identity) *For all $m \in \mathbb{N}$ and $k \in \mathbb{N}$,*

$$e_{k+1}(\mathrm{id}_{2,\infty}^m) \leq \min\left(1, c\sqrt{\frac{\log(\frac{m}{k}+1)}{k}}\right), \tag{29}$$

where $c = 1.86$.

This result, without the explicit constant and the $\min(1, \ldots)$ is due to Schütt [8]. The $\min(1, \ldots)$ is an immediate consequence of Lemma 4.2 (1) since $\|\mathrm{id}_{2,\infty}^m\| = 1$. The explicit value of c is determined in [12].

We first decompose the multiple evaluation operator $S_{\mathbf{x}}$ (cf. Definition 2.6 and (10)) into two operators,

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

where $\tilde{S}_{\mathbf{x}}$ 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 defined in (28).

We then decompose $\tilde{S}_{\mathbf{x}}$ into a sequence of three operators given by a singular value decomposition. This will allow us to bound the entropy numbers of $\tilde{S}_{\mathbf{x}}$ using the bound for diagonal operators of Theorem 4.3. The situation is summarized in the following diagram.

$$\begin{tikzcd} \ell_2^M \arrow[rr, "S_{\mathbf{x}}"] \arrow[d, "W_m^\top"] \arrow[drr, "\tilde{S}_{\mathbf{x}}"] & & \ell_\infty^m \\ \ell_2^m \arrow[r, "\Sigma"] & \ell_2^m \arrow[r, "V"] & \ell_2^m \arrow[u, "\mathrm{id}"'] \end{tikzcd} \tag{30}$$

Lemma 4.6 *Let $\mathbf{X} = WSV^\top$ be the singular value decomposition of the matrix $\mathbf{X}$ whose columns are the points of the training sample. Then we can write*

$$\tilde{S}_{\mathbf{x}} = V \circ \Sigma \circ W_m^\top, \tag{31}$$

where the mapping W_m consists of the first m columns of W and $\Sigma = D(\sigma_i)$ is the leading $m \times m$ principal submatrix of S. Note that $W_m^\top$ 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 $W_m^\top$ and V satisfy $\|W_m^\top\| = \|V\| = 1$, while $\sigma_i = \sqrt{\lambda_i}$, where λ_i is the i-th eigenvalue of the Gram matrix $G = \mathbf{X}^\top\mathbf{X}$. Thus (30) commutes.

[3] This is achieved by proving $\lim_{k \to \infty} n_k = \infty$, hence for every n there exists a k with $n \in [n_k, n_{k+1}]$.

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

$$\tilde{S}_{\mathbf{x}}(w) = w^\top \mathbf{X} = w^\top W S V^\top = w^\top W_m \Sigma V^\top = (V \circ \Sigma \circ W_m^\top)(w).$$

Hence, the decomposition is shown. Since, W and V are unitary matrices, we have $\|V\| = 1$ and $\|W_m^\top\| \leq 1$. Choosing the first column w_1 of W and observing that $\|W_m^\top w_1\| = 1$ shows that equality also holds for the norm of $W_m^\top$. Finally, observe that $G = V\Sigma^2 V^\top = V\Lambda V^\top$, 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 the growth function $\sup_{\mathbf{X}} \mathcal{N}(\epsilon, \mathcal{F}, \mathbf{X})$, an alternative strategy is warranted. See [12] for examples of the calculation of growth functions suitable for traditional generalization error bounds.

Corollary 4.7 *For any $n \in \mathbb{N}$ and $j = 1, \ldots, n$, the dyadic entropy numbers of the multiple evaluation map satisfy*

$$e_n(S_{\mathbf{x}}) \leq e_j(\Sigma) e_{n-j+1}(\mathrm{id}_{2,\infty}^m). \tag{32}$$

Proof: We apply Lemma 4.2 several times:

$$e_n(S_{\mathbf{x}}) \leq e_1(W_m^\top) e_j(\Sigma) e_1(V) e_{n-j+1}(\mathrm{id}_{2,\infty}^m). \tag{33}$$

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

Lemma 4.4 allows us to prove the following bound on the covering numbers of the function class $\mathcal{F}$.

Theorem 4.8 *Let $\mathcal{F}$ be as in (10). Suppose $m \in \mathbb{N}$, and $\mathbf{X} = (x_1, \ldots, x_m)$ is an arbitrary sequence of m points in X. Define G, $\lambda_1, \ldots, \lambda_{m+1}$, $k(n)$ and $\Upsilon(n)$ as in Theorem 4.1. Then for all $n \in \mathbb{N}$, and for all $\epsilon > \Upsilon(n)$,*

$$\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbf{X}) \leq n. \tag{34}$$

Proof: We will compute the entropy numbers of the operator $S_{\mathbf{x}}$ using Corollary 4.7:

$$e_n(S_{\mathbf{x}}) \leq e_j(\Sigma) e_{n-j+1}(\mathrm{id}_{2,\infty}^m) \tag{35}$$

where $j \in \{1, \ldots, n\}$ is arbitrary and where $U\Sigma V^\top$ is a singular value decomposition of $\mathbf{X}$. Theorem 4.3 and Lemma 4.4 give

$$e_j(\Sigma) \leq 6 \cdot 2^{-(j-1)/k(j)} (\sigma_1 \cdots \sigma_{k(j)})^{1/k(j)} \tag{36}$$

This, together with (35) and Theorem 4.5 gives

$$e_n(S_{\mathbf{x}}) \leq 6 \cdot 2^{-(j-1)/k(j)} (\sigma_1 \cdots \sigma_{k(j)})^{1/k(j)} \cdot \min\left(1, 1.86\sqrt{\frac{\log(\frac{m}{n-j}+1)}{n-j}}\right) \tag{37}$$

From Lemma 4.6, $\sigma_i = \sqrt{\lambda_i}$ for $i = 1, \ldots, m$. Since j was arbitrary, we can minimize over j and so for all $n \in \mathbb{N}$, $e_n(S_{\mathbf{x}}) \leq \Upsilon(n)$. Finally observe that $S_{\mathbf{x}}(U_{\ell_2}) = \mathcal{F}(\mathbf{X}) = (\mathcal{F}(x_1), \ldots, \mathcal{F}(x_m))$ where $\mathcal{F}(x_i) := \{f(x_i): f \in \mathcal{F}\}$. Thus for all $\epsilon > e_n(S_{\mathbf{x}})$,

$$\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbf{X}) \leq n - 1 \tag{38}$$

and so for all $\epsilon > \Upsilon(n)$, $\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbf{X}) \leq n - 1$. ∎

The theorem is for data $\mathbf{X}$ that may not be centered in feature space. This could result in a poor bound. Thus we may wish to translate all points by t; i.e. use $x_i' = x_i + t$. Observe that the $k = k(n)$ obtained for the optimal translation vector t is the dimension of the affine space which contains the points to within a margin specified by ϵ. Hence, it is clear that a good choice of t will be the centre of gravity of the training points

$$-t = \bar{x} = \frac{1}{m} \sum_{i=1}^{m} x_i. \tag{39}$$

This will not only choose an origin guaranteed to be in the centre of any affine subspace containing the points, but will also minimise

$$\sum_{i=1}^{m} \|x_i + t\|^2 = \sum_{i=1}^{m} \lambda_i, \tag{40}$$

a fact which is well known from the study of PCA. One can also place the centre at a point which minimises the radius R of the ball containing the points and using Theorem 2.2 obtain a bound on the fat-shattering dimension. However one can not utilize the result without modication in this case. Note that translation by t is possible since the "$+b$" offset commonly used in SV machines is already taking this fact into account:

$$\langle w, x + t \rangle + b = \langle w, x \rangle + \langle w, t \rangle + b = \langle w, x \rangle + b'. \tag{41}$$

As explained in [13] this means one needs to add a correction term to the covering numbers. In our case this will mean replacing U by U' where $U' = U + \frac{16B}{\gamma}$ where B is an upper bound on $|b|$. Except in situations where λ_i decay *extremely* rapidly, this will have negligble effect; i.e $U' \approx U$.

5 Experiments

To test the utility of the novel bounds for model selection, we ran a set of experiments on the well-known US postal service handwritten digit recognition benchmark. The dataset consists of 7291 digits of size 16×16, with a test set of 2007 patterns. To keep the computational complexity associated with computing the eigenvalue decomposition limited, we considered two-class subproblems.

In all experiments, we divided the training set into 23 random subsets of size 317 ($23 \cdot 317$ being the prime factorization of the training set size). Error bars in figures 1 – 3 denote standard deviations over the 23 trials. At the beginning of the experiment, the whole USPS set (training plus test set) was permuted, to ensure that the distribution of training and test data is the same.[4] We considered three tasks. In the first case, we separated digits 0 through 4 from 5 through 9; in the second one, we separated even from odd digits. The third task, finally, differs from the above two in that the two classes do not have similar size: in that case, we separated digit 4 from all the others. Digit 4 was chosen since it is known to be the hardest in this data set, and for larger error rates we expect to find a more reliable minimum in the error curve.

SVMs usually come with two free parameters: the regularization constant C, or ν (depending on which parametrization one prefers, cf. [11, 7]), and the kernel parameter. To make our bounds applicable, we chose the regularization such that the SVM attains zero training error over all experiments, thus focusing our model selection efforts on the kernel parameter, which in our case was the width σ of a Gaussian kernel

$$k(x,y) = \exp\left(-\frac{\|x-y\|^2}{\sigma}\right). \tag{42}$$

Note that using k corresponds to an inner product in a feature space nonlinearly related to input space by a map Φ induced by the kernel,

$$k(x,y) = \langle \Phi(x), \Phi(y) \rangle, \tag{43}$$

and the SVM constructs a separating hyperplane for the data mapped by Φ. All the reasoning of the previous sections applies to these mapped data.

For a range of values of σ, we computed the SVM hyperplane and evaluated the bound (14). The Gram matrix, which in this case equals the kernel matrix $G_{ij} = k(x_i, x_j)$, was computed for centered data (cf. (39)). For general Φ (which could map into an infinite-dimensional space), the data cannot always be centered explicitly, nor can we always perform a singular value decomposition of the mapped data. We circumvent this problem by diagonalizing the kernel matrix of the centered data

$$\left\{ \Phi(x_i) - \frac{1}{m}\sum_{j=1}^{m} \Phi(x_j) \,\middle|\, i = 1, \ldots, m \right\}, \tag{44}$$

which (cf. [6]) equals $(\mathbf{1} - 1_m)G(\mathbf{1} - 1_m)$, where $(1_m)_{ij} := 1/m$ and $\mathbf{1}_{ij} = \delta_{ij}$ for $i, j \in [m]$, i.e. we carry out a rank–one update of G. Note that this is precisely the matrix used in kernel PCA.

The results are given in figures 1 – 3. For all three classification problems, the minimum of the test error occurs at the value of σ which minimizes the new bound. Moreover, the new bound even resembles the *shape* of the test error curve closely.

The previously known bound [11, 2], involving the fat-shattering dimension R^2/γ^2, led to worse predictions of the optimal σ. Essentially, in our situation

[4] it is known that this is not the case for the original USPS set. Nevertheless, similar results with error rates which are slightly worse are obtained if one does not permute.

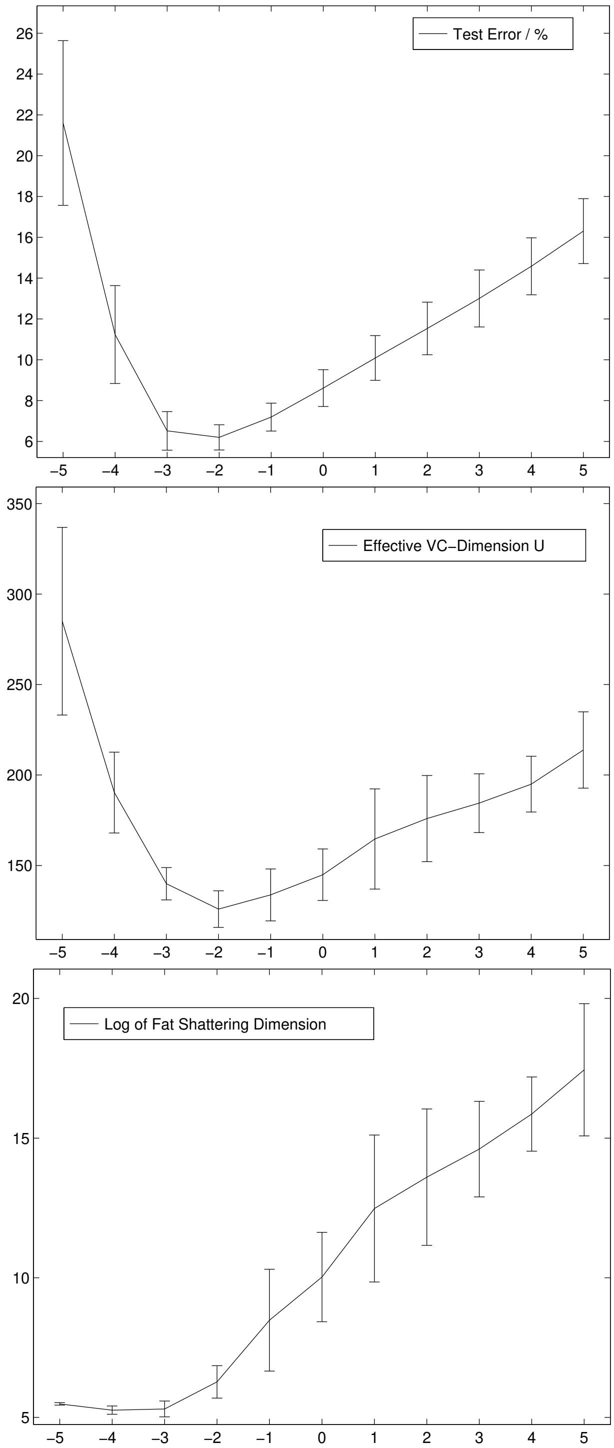

Figure 1 Task: separate digits 0 through 4 from 5 through 9. Shown are the test error, the new bound (more precisely, the "effective VC-Dimension U, cf. Theorem 4.1), and the log of the old R^2/γ^2 bound (cf. text).

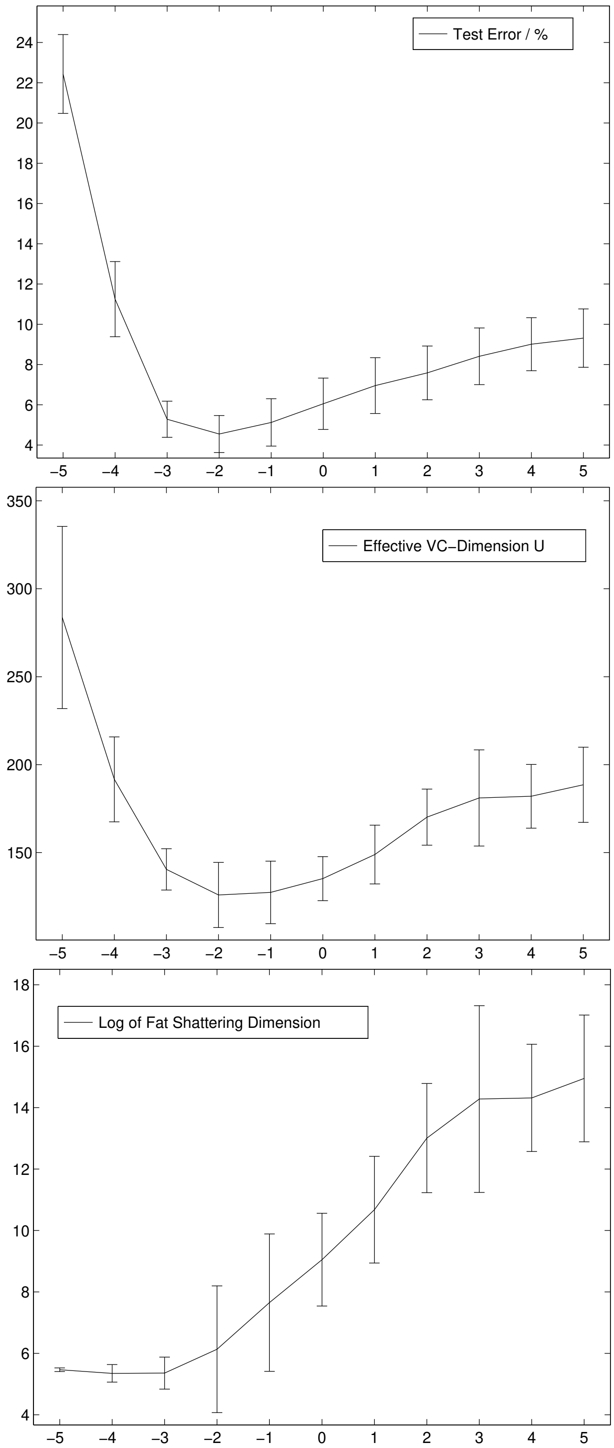

Figure 2 Task: separate even from odd digits.

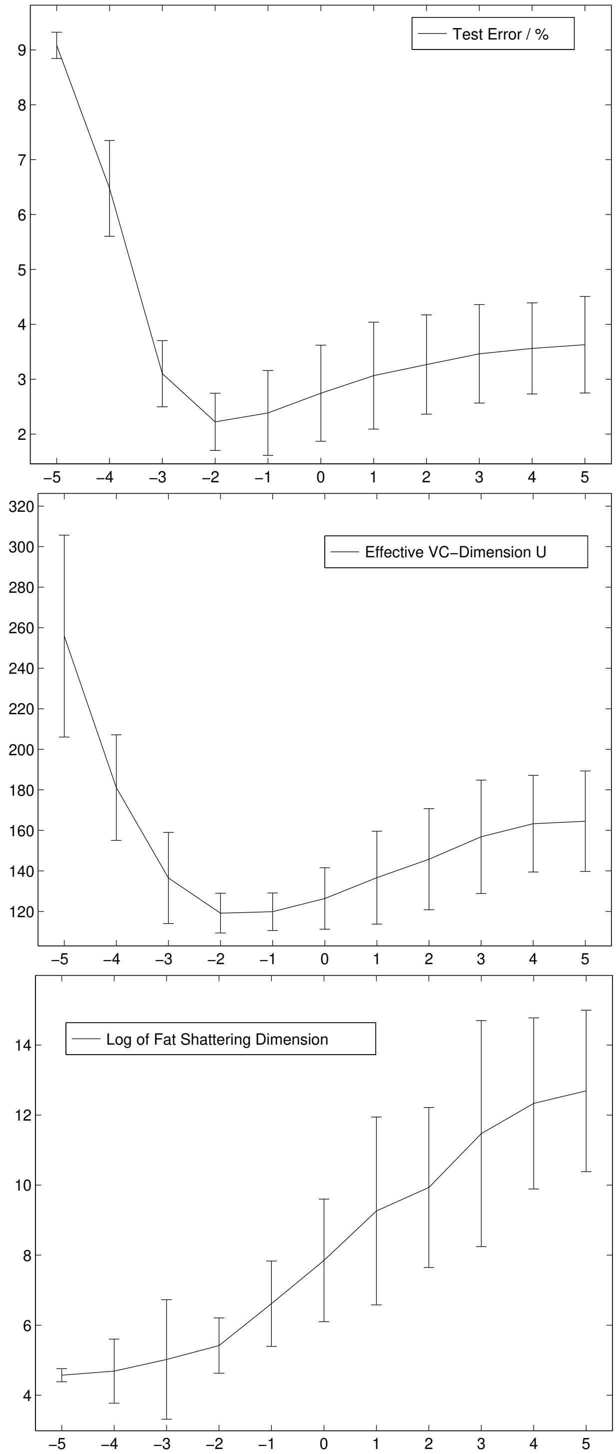

Figure 3 Task: separate digit 4 from the rest.

this bounds states that we should select the σ which leads to the maximum margin: the value of R was estimated as 1 since in the case of RBF-kernels, $k(x,x) = 1$ for all x.

The figures show that the most important shortcoming of the old bound, and the strength of the new one, is the case of large σ. This makes sense: since the eigenvalues of a translation-invariant kernel are obtained by a Fourier transform, the case of large σ corresponds to the fastest decay in the eigenvalues, which is exactly what is taken into account by the new bound.

Finally we point out that in a sense we made it hard for ourselves in this example. Let $\tilde{x} = \Phi(x)$ be the point in feature space corresponding to an input x. Observe that $\|\tilde{x}\|^2 = \langle\Phi(x), \Phi(x)\rangle = k(x,x) = 1$ for kernels of the form (42). Thus *all* points in feature space here have length 1, and so our new bound is only exploiting their direction. One could readily construct example where the length of vectors varied greatly, in which case our bound should be much better than the R^2/γ^2 one.

6 Conclusions

This paper has presented a method by which the eigenvalues of the Gram matrix of the training set can be used to bound the generalization error of a linear classifier. We presented an example which illustrates that the minimum of the bound is a good predictor of the true error. Our result can be taken to justify the folklore heuristic of taking some account of these eigenvalues in tuning a classifier.

Naturally much remains to be done. Whilst the behaviour of the bound (in our example) seems good, the numerical value is rather unsatisfying. There are several reasons why the bound is not tight. We believe the constant 6 in Theorem 4.3 can be reduced, perhaps to 3. The $\log^2(m)$ factor in the main result comes from the result in [1]. We believe this can be reduced to $\log(m)$. More speculatively, we believe the $\epsilon/4$ in Corollary 2.5 can be improved to $\epsilon/2$. Apart from tightening the results, there are some obvious generalizations to be done: the case of non-zero training error being the most pressing.

The new results should be straight-forwardly applicable to bounding the error on an independent test set (as in transduction). This would avoid any of the looseness incurred by the empirical covering number statistical result. Finally, all of the new results should be compared with the *a priori* bounds in terms of the eigenvalues of the integral operator induced by the kernel k [13]; in some cases it is likely that the *a priori* bounds could be better since they do not need Theorem 3.2.

Acknowledgements This work was supported by the Australian Research Council, the European Commission under the Working Group Nr. 27150 (NeuroCOLT2), and the DFG (Ja 379/71, 379/51, and 379/9-1). Part of the motivation for this work originated in discussions with Mario Marchand in May 1996.

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] P. Bartlett and J. Shawe-Taylor. Generalization performance of support vector machines and other pattern classifiers. In B. Schölkopf, C. J. C. Burges, and A. J. Smola, editors, *Advances in Kernel Methods — Support Vector Learning*, pages 43–54, Cambridge, MA, 1999. MIT Press.

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

[4] 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.

[5] B. Schölkopf, C. Burges, and V. Vapnik. Extracting support data for a given task. In U. M. Fayyad and R. Uthurusamy, editors, *Proceedings, First International Conference on Knowledge Discovery & Data Mining*, Menlo Park, 1995. AAAI Press.

[6] B. Schölkopf, A. Smola, and K.-R. Müller. Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation*, 10:1299–1319, 1998.

[7] B. Schölkopf, A. J. Smola, P. Bartlett, and R. Williamson. Shrinking the tube — a new support vector regression algorithm. In *Neural Information Processing Systems 1998*, Boston, MA, 1999. MIT Press. forthcoming.

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

[9] 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.

[10] J. Shawe-Taylor and R. C. Williamson. Generalization performance of classifiers in terms of observed covering numbers. In *Proceedings of EUROCOLT'99*, 1999.

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

[12] R. C. Williamson, B. Schölkopf, and A. J. Smola. A Maximum Margin Miscellany. Typescript, 1999.

[13] R. C. Williamson, A. J. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines via entropy numbers of compact operators. NeuroCOLT NC-TR-98-019, Royal Holloway College, 1998.