Kernel-Dependent Support Vector Error Bounds*

Bernhard Schölkopf,[†‡] John Shawe-Taylor,[♯‡] Alex J. Smola,[†‡] Robert C. Williamson[‡]

bs@first.gmd.de jst@dcs.rhbnc.ac.uk smola@first.gmd.de bob.williamson@anu.edu.au

† GMD FIRST, Rudower Chaussee 5, 12489 Berlin Germany

♯ Dept. Computer Science Royal Holloway College, University of London Egham, TW20 0EX, UK

‡ Dept. of Engineering Australian Natl. University, 0200 Canberra Australia

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 eigenvalue distribution of the Gram matrix of the data. 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 and the width of the margin $1/\|w\|$ since this corresponds to a classifier with minimal fat shattering dimension [4]. Whilst in general capturing the correct scaling behaviour in terms of the weight vector w, this approach has the shortcoming that it completely ignores the information about the *distribution* of the data inside the sphere. 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. 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.

In this short paper, we shall summarize the main results. For lack of space, the proofs are relegated to [5].

*This work was supported by the Australian Research Council, the European Commission under the Working Group Nr. 27150 (NeuroCOLT2), and the DFG (Ja 379/7-1,9-1). Part of the motivation for this work originated in discussions with Mario Marchand in May 1996.

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. First we define the fat-shattering dimensions; cf. [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*

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

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.*

Definition 2.2 *For $M \in \mathbb{N}$, we define the spaces ℓ_p^M as follows: as vector spaces, they are identical to $\mathbb{R}^M$, in addition, they are endowed with p-norms: for $0 < p < \infty$, $\|x\|_{\ell_p^M} := \|x\|_p = \left(\sum_{j=1}^M |x_j|^p\right)^{1/p}$; for $p = \infty$, $\|x\|_{\ell_\infty^M} := \|x\|_\infty = \max_{j=1,\ldots,M} |x_j|$.*

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

Theorem 2.3 ([2]) *Let X be a ball of radius R in an inner product space and let*

$$\mathcal{F} = \{x \mapsto \langle w, x \rangle : \|w\|_2 \leq 1, x \in X\}. \quad (2)$$

Then for all $\gamma > 0$.

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

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

Definition 2.4 (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 $B \subseteq A$.*

Definition 2.5 (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*

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

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

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

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 2.6 *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*

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

Here and below, log denotes the logarithm to base 2.

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 2.7 (Evaluation Operator) *Let*

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

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 $\mathbb{R}^m$.*

Definition 2.8 *For $U \in \mathbb{N}$ and $\delta \in \mathbb{R}^+$, we define the function $\alpha(U, \delta) := 3.08\left(1 + \frac{1}{U} \ln \frac{1}{\delta}\right)$.*

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 2.9 ([10]) *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*

$$\begin{aligned} \epsilon(m, U, \delta) \quad = \quad & \tfrac{2}{m}\Big(U\Big(1 + \alpha\left(U, \tfrac{\delta}{2}\right) \log\left(\tfrac{5em}{U}\right) \\ & \log(17m)\Big) + \log\left(\tfrac{16m}{\delta}\right)\Big), \end{aligned} \tag{8}$$

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

Here, $\lfloor \cdot \rfloor$ denotes the floor function.

3 Main Result

In this section we will restrict consideration to linear functions of arbitrary dimension. By the standard kernel trick 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. 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 (cf. [12].

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

$$\mathcal{F} = \{x \mapsto \langle w, x \rangle : \|w\|_{\ell_2^M} \leq 1\}. \tag{9}$$

Note that if X is compact, then the sturdyness of $\mathcal{F}$ follows directly from 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 gives a bound on the generalization error in terms of the eigenvalues of the Gram matrix of the training points:

Theorem 3.1 *Let $\mathcal{F}$ be the set of linear functions in some feature space defined in (9). 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*

$$k(n) := \min\left\{j : j \in \{1, \ldots, m\}, \lambda_{j+1}^j \leq \left(\frac{\lambda_1 \cdots \lambda_j}{n^2}\right)\right\} \tag{10}$$

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

$$\Upsilon(n) := 6 \min_{j \in \{1, \ldots n-1\}} 2^{\frac{1-j}{k(j)}} (\lambda_1 \cdots \lambda_{k(j)})^{\frac{1}{2k(j)}} \cdot \min\left(1, 1.86\sqrt{\frac{\log(\frac{m}{n-j}+1)}{n-j}}\right) \tag{11}$$

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.

$$\gamma := \min_{i=1,\ldots,m} |f^*(x_i) - \theta|. \tag{12}$$

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

$$\epsilon(m, U, \delta) = \frac{2}{m}\left(U\left(1 + \alpha\left(U, \tfrac{\delta}{2}\right)\log\left(\tfrac{5em}{U}\right)\log(17m)\right) + \log\left(\tfrac{16m}{\delta}\right)\right), \tag{13}$$

where α is as in Definition 2.8.

The proof of this result follows from Theorem 3.10 below and Theorem 2.9 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.

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 2.7. 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.

Definition 3.2 (Entropy Numbers of Sets) *The nth* entropy number *of a set $S \subset X$ is defined by $\epsilon_n(S) = \epsilon_n(S, d) := \inf\{\epsilon > 0 : \mathcal{N}_d(\epsilon, S) \leq n\}$.*

We denote by U_d the (closed) unit ball: $U_d := \{x \in X : d(x, 0) \leq 1\}$. 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 $\|T\| := \sup\{\|Tx\|_Y : \|x\|_X \leq 1\}$. 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.

Definition 3.3 (Entropy Numbers of Operators) *If $T \in \mathcal{L}(X, Y)$ the* entropy numbers of the operator T *are defined by $\epsilon_n(T) := \epsilon_n(T(U_X))$. The* dyadic entropy numbers $e_n(T)$ *are defined by $e_n(T) := \epsilon_{2^{n-1}}(T)$ with $n \in \mathbb{N}$.*

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

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

Lemma 3.4 ([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.

Theorem 3.5 ([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*

$$D : \ell_p \to \ell_p \text{ with } D\mathbf{x} = (\sigma_1 x_1, \sigma_2 x_2, \ldots, \sigma_j x_j, \ldots) \tag{14}$$

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 $n \in \mathbb{N}$,

$$\sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}} \leq \epsilon_n(D) \leq 6 \sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}}. \tag{15}$$

An alternative form of this result in terms of the dyadic entropy numbers is convenient for us. The upper bound becomes: for all $n \in \mathbb{N}$, $e_n(D) \leq 6 \sup_{j \in \mathbb{N}} 2^{-\frac{n-1}{j}} (\sigma_1 \cdots \sigma_j)^{1/j}$.

The bound of Theorem 3.5 is worth analysing more closely. The following lemma shows for which value of j the sup is obtained.

Lemma 3.6 ([5]) *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*

$$n_j := \frac{\prod_{i=1}^{j-1} \sigma_i}{\sigma_j^{j-1}} = \frac{\prod_{i=1}^{j} \sigma_i}{\sigma_j^{j}}. \tag{16}$$

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

$$\begin{aligned}\sigma_{k+1} &\le \sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}} \\ &= n^{-\frac{1}{k}} (\sigma_1 \sigma_2 \cdots \sigma_k)^{\frac{1}{k}} \le \sigma_k. \quad (17)\end{aligned}$$

In order to compute the bound in (15), we need only find a k such that $n_k \le n \le n_{k+1}$. To obtain the best bound we want the smallest k such that $n \le n_{k+1} = (\sigma_1 \cdots \sigma_k)/\sigma_{k+1}^k$ or equivalently $\sigma_{k+1} \le ((\sigma_1 \cdots \sigma_k)/n)^{1/k}$. In [5], it is shown that for any $n \in \mathbb{N}$ such a k always exists if $\lim_{j \to \infty} \sigma_j = 0$. This justifies the definition of $k(n)$ in (10).

The other operator whose entropy numbers we must bound is $\mathrm{id}_{2,\infty}^m \colon \ell_2^m \to \ell_\infty^m, x \mapsto x$.

Theorem 3.7 (Entropy Numbers of $\mathrm{id}_{2,\infty}^m$ [8, 13]) *For all* $m \in \mathbb{N}$ *and* $k \in \mathbb{N}$,

$$e_{k+1}(\mathrm{id}_{2,\infty}^m) \le \min\left(1, 1.86\sqrt{\frac{\log(\frac{m}{k}+1)}{k}}\right). \tag{18}$$

We first decompose the multiple evaluation operator $S_{\mathbf{x}}$ (cf. Definition 2.7 and (9)) 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.

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 3.5. The situation is summarized in the following diagram.

$$\begin{tikzcd} \ell_2^M \arrow[rr, "S_{\mathbf{x}}"] \arrow[d, "W_m^\top"'] \arrow[drr, "\tilde{S}_{\mathbf{x}}"] & & \ell_\infty^m \\ \ell_2^m \arrow[r, "\Sigma"] & \ell_2^m \arrow[r, "V"] & \ell_2^m \arrow[u, "\mathrm{id}"'] \end{tikzcd} \tag{19}$$

Lemma 3.8 ([5]) *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*

$$\tilde{S}_{\mathbf{x}} = V \circ \Sigma \circ W_m^\top, \tag{20}$$

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 (19) commutes.*

Corollary 3.9 ([5]) *For any* $n \in \mathbb{N}$ *and* $j = 1, \ldots, n$, *the dyadic entropy numbers of the multiple evaluation map satisfy*

$$e_n(S_{\mathbf{x}}) \le e_j(\Sigma) e_{n-j+1}(\mathrm{id}_{2,\infty}^m). \tag{21}$$

The proof uses Lemma 3.4.

Lemma 3.6 allows us to prove the following bound on the covering numbers of the function class $\mathcal{F}$.

Theorem 3.10 ([5]) *Let* $\mathcal{F}$ *be as in (9). 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 3.1. Then for all* $n \in \mathbb{N}$, *and for all* $\epsilon > \Upsilon(n)$,

$$\log \mathcal{N}(\epsilon, \mathcal{F}, \mathbf{X}) \le n. \tag{22}$$

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

$$-t = \bar{x} = \frac{1}{m} \sum_{i=1}^m x_i. \tag{23}$$

This will not only choose an origin guaranteed to be in the centre of any affine subspace containing the points, but will also minimise $\sum_{i=1}^m \|x_i + t\|^2 = \sum_{i=1}^m \lambda_i$, a fact which is well known from the study of PCA. For a discussion of how the choice of the origin is related to the constant offset used in SVMs, cf. [5].

4 Experiments and Discussion

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 · 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

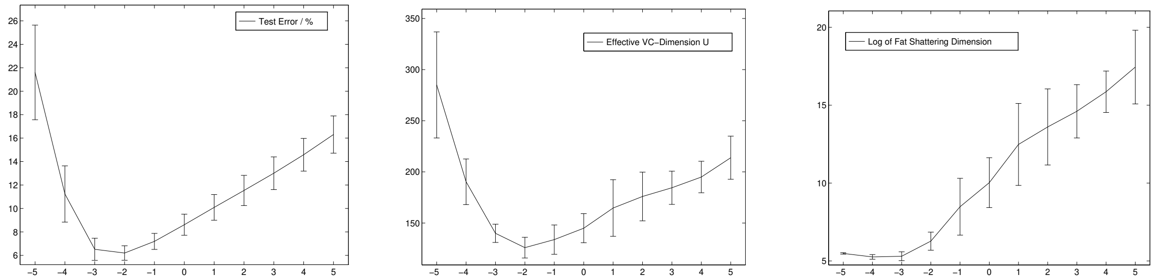

Figure 1: Task: separate digits 0 through 4 from 5 through 9. Shown are the test error, the new bound (more precisely, the "effective VC-Dimension U, cf. Theorem 3.1), and the log of the old R^2/γ^2 bound (cf. text), all as functions of $\log(\sigma)$.

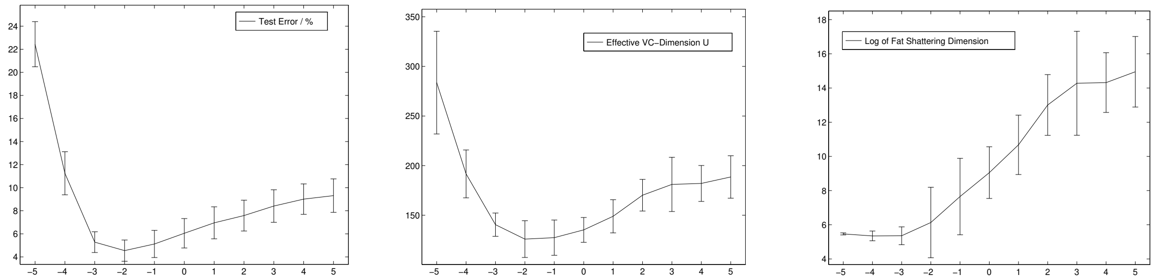

Figure 2: Task: separate even from odd digits.

training and test data is the same. 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

$$k(x,y) = \exp\left(-\|x-y\|^2/\sigma\right). \qquad (24)$$

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, $k(x,y) = \langle\Phi(x), \Phi(y)\rangle$, 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 (13). The kernel Gram matrix $G_{ij} = k(x_i, x_j)$ was computed for centered data (cf. (23)). 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

$$\left\{\Phi(x_i) - \frac{1}{m}\sum_{j=1}^{m}\Phi(x_j)\middle| i = 1,\ldots,m\right\}, \qquad (25)$$

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 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

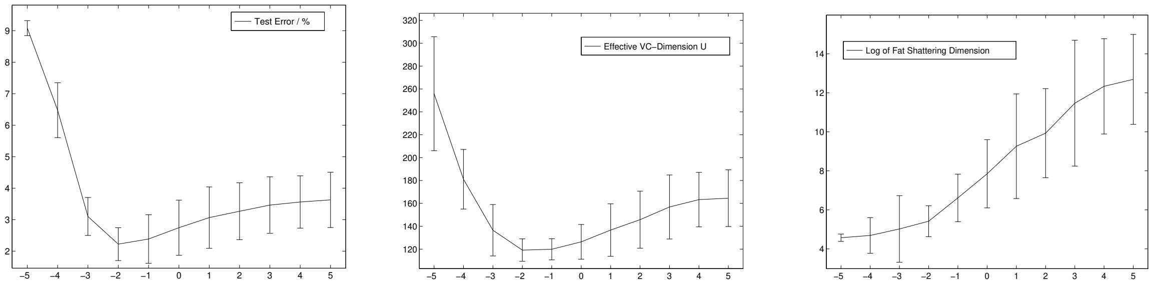

Figure 3: Task: separate digit 4 from the rest.

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 (24). Thus *all* points in feature space have length 1, and so our new bound is only exploiting their direction. One could readily construct an example where the length of vectors varied greatly, in which case our bound should be much better than the R^2/γ^2 one.

We conclude that the new method by which the eigenvalues of the Gram matrix of the training set are used to bound the generalization error of an SVM is a promising and astonishingly accurate substitute for the previous bound. Our result can be taken to justify the heuristic of taking some account of these eigenvalues in tuning a classifier.

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. Burges, and A. Smola, editors, *Advances in Kernel Methods*, pages 43 – 54. MIT Press, Cambridge, MA, 1999.

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

[4] 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*. AAAI Press, Menlo Park, CA, 1995.

[5] B. Schölkopf, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson. Generalization bounds via eigenvalues of the Gram matrix. NeuroColt TR 035, 1999. Available from http://svm.first.gmd.de.

[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. Smola, R. C. Williamson, and P. L. Bartlett. New support vector algorithms. *To appear in: Neural Computation*, 1999. Also: NeuroColt2-TR 1998-031.

[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 *Computational Learning Theory: 4th European Conference*, volume 1572 of *Lecture Notes in Artificial Intelligence*, pages 274 – 284. Springer, 1999.

[11] V. Vapnik. *The Nature of Statistical Learning Theory*. Springer Verlag, New York, 1995.

[12] 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 TR 19, http://www.neurocolt.com, 1998.

[13] R. C. Williamson, A. J. Smola, and B. Schölkopf. A Maximum Margin Miscellany. Typescript, March 1998.