9 Entropy Numbers, Operators and Support Vector Kernels

Robert C. Williamson
Australian National University, Department of Engineering
Canberra, ACT 0200, Australia
Robert.Williamson@anu.edu.au
http://spigot.anu.edu.au/people/williams/home.html

Alexander J. Smola, Bernhard Schölkopf
GMD FIRST
Rudower Chaussee 5, 12489 Berlin, Germany
smola,bs@first.gmd.de
http://www.first.gmd.de/∼smola,bs

We derive new bounds for the generalization error of kernel machines, such as support vector machines and related regularization networks by obtaining new bounds on their covering numbers. The proofs make use of a viewpoint that is apparently novel in the field of statistical learning theory. The hypothesis class is described in terms of a linear operator mapping from a possibly infinite dimensional unit ball in feature space into a finite dimensional space. The covering numbers of the class are then determined via the entropy numbers of the operator. These numbers, which characterize the degree of compactness of the operator, can be bounded in terms of the eigenvalues of an integral operator induced by the kernel function used by the machine. As a consequence we are able to theoretically explain the effect of the choice of kernel function on the generalization performance of support vector machines.

9.1 Introduction

In this chapter we give new bounds on the covering numbers for kernel machines. This leads to improved bounds on their generalization performance. Kernel machines perform a mapping from input space into a feature space (see e.g. Aizerman et al. (1964); Nilsson (1965)), construct regression functions or decision boundaries based on this mapping, and use constraints in feature space for capacity control. Support Vector machines are a well known example of this class. We will use SV machines as our model of choice to show how bounds on the covering numbers can be obtained. We outline the relatively standard methods one can then use to hence bound their generalization performance. Our reasoning also applies to similar algorithms such as regularization networks (Girosi et al., 1993) or certain unsupervised learning algorithms (chapter 20, and Schölkopf et al. (1998e)).

Regularization Networks and Kernels

It has been noticed that in SV machines different kernels can be characterized by their regularization properties (Smola et al., 1998c): SV machines are regularization networks minimizing the regularized risk $R_{reg}[f] = R_{emp}[f] + \frac{\lambda}{2}\|Pf\|^2$, (note the similarity to (1.42)) with a regularization parameter $\lambda = \frac{1}{C} \geq 0$, and a regularization operator P, over the set of functions of the form (1.49), provided that k and P are interrelated by $k(\mathbf{x}_s, \mathbf{x}_t) = \langle (Pk)(\mathbf{x}_s, \cdot), (Pk)(\mathbf{x}_t, \cdot) \rangle$. To this end, k is chosen as Green's function of P^*P, where P^* is the adjoint of P.

This provides insight into the regularization properties of SV kernels. However, it does not completely settle the issue of how to select a kernel for a given learning problem, and how using a specific kernel might influence the performance of a SV machine.

Outline of the Chapter.

In the present work, we show that properties of the spectrum of the kernel can be used to make statements about the generalization error of the associated class of learning machines. Unlike in previous SV learning studies, the kernel is no longer merely a means of broadening the class of functions used, e.g. by making a nonseparable dataset separable in a feature space nonlinearly related to input space. Rather, we now view it as a constructive handle by which we can control the generalization error.

Direct Bounds on Covering Number

A key feature is the manner in which we *directly* bound the covering numbers of interest rather than making use of a Combinatorial dimension (such as the VC dimension or the fat-shattering dimension) and subsequent application of a general result relating such dimensions to covering numbers. We bound covering numbers directly by viewing the relevant class of functions as the image of a unit ball under a particular compact operator. A general overview of the method is given in section 9.3.

The remainder of the chapter is organized as follows. We start by introducing notation and definitions (section 9.2). Section 9.4 formulates generalization error bounds in terms of covering numbers. Section 9.5 contains the main result bounding entropy numbers in terms of the spectrum of a given kernel. The results in this chapter rest on a connection between covering numbers of function classes and entropy numbers of suitably defined operators. In particular, we derive an upper bound on the entropy numbers in terms of the size of the weight vector in feature space and the eigenvalues of the kernel used. Section 9.6 shows how to make use of kernels such as $k(x) = e^{-x^2}$ which do not have a discrete spectrum. Section 9.7 presents some results on the entropy numbers obtained for given rates of decay of eigenvalues. The concluding section 9.8 indicates how the various results in the chapter can be glued together in order to obtain overall bounds on the generalization error. Lengthy proofs have been omitted wherever they were not crucial for the understanding of the basic idea — we refer the reader to Williamson et al. (1998a) for the missing details.

We do not present a single master generalization error theorem for three key reasons: 1) the only novelty in the chapter lies in the computation of covering numbers themselves; 2) the particular statistical result one needs to use depends on the specific problem situation; 3) many of the results obtained are in a form which, whilst quite amenable to ready computation on a computer, do not provide much direct insight by merely looking at them, except perhaps in the asymptotic sense, and finally, 4) some applications (such as classification) where further quantities like margins are estimated in a data dependent fashion, need an additional luckiness argument (Shawe-Taylor et al., 1998) to apply the bounds (see also chapter 4).

Thus although our goal has been theorems, we are ultimately forced to resort to a computer to make use of our results. This is not necessarily a disadvantage — it is a both a strength and a weakness of Structural Risk Minimization (SRM) (Vapnik, 1979) that a good generalization error bound is both necessary and sufficient to make the method work well. It is our expectation that the refined (and significantly tighter) covering number bounds obtainable by our methods will be exploitable in SRM algorithms — they could be used for example for model selection. If one is running a computer program anyway, there is little point in expending a large effort to make the generalization error bounds directly consumable in a pencil and paper sense.

9.2 Definitions and Notation

Norms in $\mathbb{R}^d$

We define spaces ℓ_p^d as follows: as vector spaces, they are identical to $\mathbb{R}^d$, in addition, they are endowed with p-norms: for $0 < p < \infty$, $\|\mathbf{x}\|_{\ell_p^d} := \|\mathbf{x}\|_p = \left(\sum_{j=1}^d |x_j|^p\right)^{1/p}$; for $p = \infty$, $\|\mathbf{x}\|_{\ell_\infty^d} := \|\mathbf{x}\|_\infty = \max_{j=1,\ldots,d} |x_j|$. Note that a different normalization of the ℓ_p^d norm is used in some papers in learning theory (e.g. Talagrand (1996)).

Given ℓ points $\mathbf{x}_1, \ldots, \mathbf{x}_\ell \in \ell_p^d$, we use the shorthand $\mathbf{X}^\ell = (\mathbf{x}_1^\top, \ldots, \mathbf{x}_\ell^\top)$.

Suppose $\mathcal{F}$ is a class of functions defined on $\mathbb{R}^d$. The ℓ_∞^d norm *with respect to* $\mathbf{X}^\ell$ of $f \in \mathcal{F}$ is defined as $\|f\|_{\ell_\infty^{\mathbf{X}^\ell}} := \max_{i=1,\ldots,\ell} |f(\mathbf{x}_i)|$.

Given some set $\mathcal{C}$, a measure μ on $\mathcal{C}$, some $1 \leq p < \infty$ and a function $f\colon \mathcal{C} \to \mathbb{K}$ we define $\|f\|_{L_p(\mathcal{C},\mathbb{K})} := \left(\int |f(x)|^p d\mu(x)\right)^{1/p}$ if the integral exists and $\|f\|_{L_\infty(\mathcal{C},\mathbb{K})} := \operatorname{ess\,sup}_{x \in \mathcal{C}} |f(x)|$. For $1 \leq p \leq \infty$, we let $L_p(\mathcal{C},\mathbb{K}) := \{f\colon \mathcal{C} \to \mathbb{K}\colon \|f\|_{L_p(\mathcal{C},\mathbb{K})} < \infty\}$. We let $L_p(\mathcal{C}) := L_p(\mathcal{C},\mathbb{R})$.

Let $\mathfrak{L}(E, F)$ be the set of all bounded linear operators T between the normed spaces $(E, \|\cdot\|_E)$ and $(F, \|\cdot\|_F)$, i.e. operators such that the image of the (closed)

Operator Norms

unit ball

$$U_E := \{x \in E : \|x\|_E \leq 1\} \tag{9.1}$$

is bounded. The smallest such bound is called the *operator norm*,

$$\|T\| := \sup_{x \in U_E} \|Tx\|_F. \tag{9.2}$$

Entropy Numbers

The nth *entropy number of a set* $M \subset E$, for $n \in \mathbb{N}$, is

$$\epsilon_n(M) := \inf\{\epsilon > 0 : \text{there is an } \epsilon\text{-cover for } M \text{ in } E \text{ with } n \text{ or fewer points}\} \tag{9.3}$$

The *entropy numbers of an operator* $T \in \mathfrak{L}(E, F)$ are defined as

$$\epsilon_n(T) := \epsilon_n(T(U_E)). \tag{9.4}$$

Dyadic Entropy Numbers

Note that $\epsilon_1(T) = \|T\|$, and that $\epsilon_n(T)$ certainly is well defined for all $n \in \mathbb{N}$ if T is a *compact operator*, i.e. if $T(U_E)$ is compact. The *dyadic entropy numbers of an operator* are defined by

$$e_n(T) := \epsilon_{2^{n-1}}(T), \qquad n \in \mathbb{N}; \tag{9.5}$$

Covering Numbers

similarly, the dyadic entropy numbers of a set are defined from its entropy numbers. A very nice introduction to entropy numbers of operators is the book of Carl and Stephani (1990). The *ϵ-covering number of $\mathcal{F}$ with respect to the metric d* denoted $\mathcal{N}(\epsilon, \mathcal{F}, d)$ is the size of the smallest ϵ-cover for $\mathcal{F}$ using the metric d.

In this chapter, E and F will always be *Banach spaces*, i.e. complete normed spaces (for instance ℓ_p^d spaces). In some cases, they will be *Hilbert spaces* H, i.e. Banach spaces endowed with a dot product $\langle \cdot, \cdot \rangle_H$ giving rise to its norm via $\|x\|_H = \sqrt{\langle x, x \rangle_H}$. We will map the input data into a feature space via a mapping Φ and let $\tilde{\mathbf{x}} := \Phi(\mathbf{x})$.

9.3 Operator Theory Methods for Entropy Numbers

In this section we briefly explain the new viewpoint implicit in the present chapter. With reference to figure 9.1, consider the traditional viewpoint in statistical learning theory.

One is given a class of functions $\mathcal{F}$, and the generalization performance attainable using $\mathcal{F}$ is determined via the covering numbers of $\mathcal{F}$. More precisely, for some set $\mathcal{C}$, and $\mathbf{x}_i \in \mathcal{C}$ for $i = 1, \ldots, m$, define the *ϵ-Growth function* of the function class $\mathcal{F}$ on $\mathcal{C}$ as

$$\mathcal{N}^\ell(\epsilon, \mathcal{F}) := \sup_{\mathbf{x}_1, \ldots, \mathbf{x}_\ell \in \mathcal{C}} \mathcal{N}(\epsilon, \mathcal{F}, \ell_\infty^{\mathbf{X}^\ell}), \tag{9.6}$$

where $\mathcal{N}(\epsilon, \mathcal{F}, \ell_\infty^{\mathbf{X}^\ell})$ is the ϵ-covering number of $\mathcal{F}$ with respect to $\ell_\infty^{\mathbf{X}^\ell}$. Many generalization error bounds can be expressed in terms of $\mathcal{N}^\ell(\epsilon, \mathcal{F})$. An example is given in the following section.

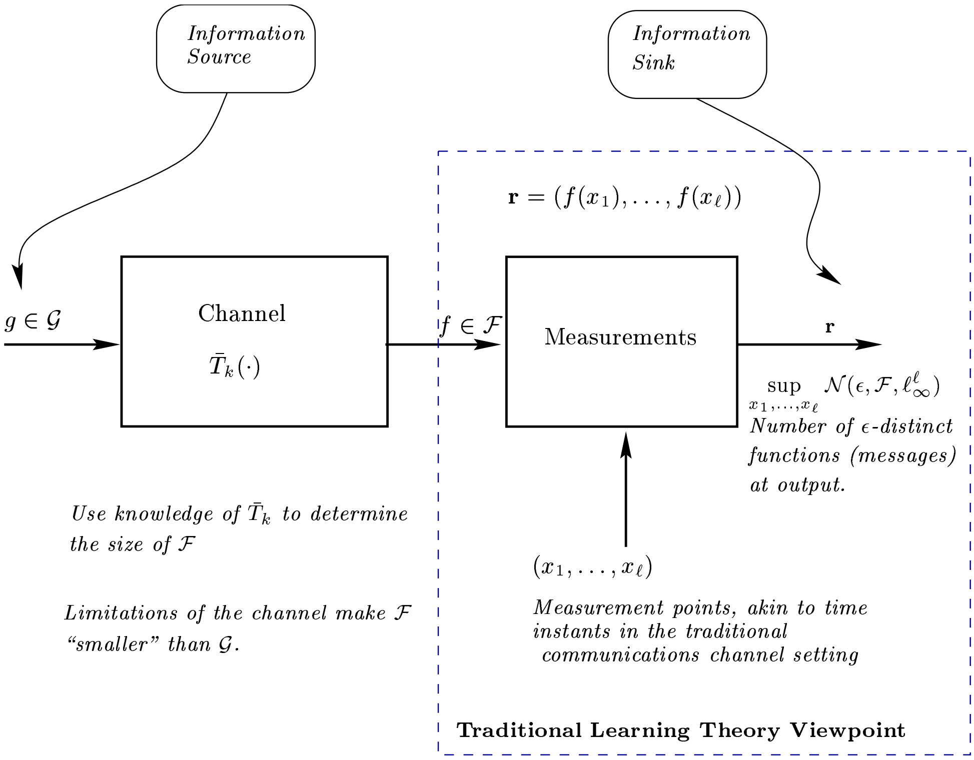
Information Source
Information Sink
$\mathbf{r} = (f(x_1), \ldots, f(x_\ell))$
$g \in \mathcal{G}$
Channel
$\bar{T}_k(\cdot)$
$f \in \mathcal{F}$
Measurements
$\mathbf{r}$
$\sup_{x_1,\ldots,x_\ell} \mathcal{N}(\epsilon, \mathcal{F}, \ell_\infty^\ell)$
Number of ϵ-distinct functions (messages) at output.
Use knowledge of $\bar{T}_k$ to determine the size of $\mathcal{F}$
Limitations of the channel make $\mathcal{F}$ "smaller" than $\mathcal{G}$.
$(x_1, \ldots, x_\ell)$
Measurement points, akin to time instants in the traditional communications channel setting
Traditional Learning Theory Viewpoint

Figure 9.1 Schematic picture of the new viewpoint.

The key novelty in the present work solely concerns the manner in which the covering numbers are computed. Traditionally, appeal has been made to a result such as the so-called Sauer's lemma (originally due to Vapnik and Chervonenkis (1968)). In the case of function learning, a generalization due to Pollard (called the pseudo-dimension), or Vapnik and Chervonenkis (called the VC dimension of real valued functions), or a scale-sensitive generalization of that (called the fat-shattering dimension) is used to bound the covering numbers. These results reduce the computation of $\mathcal{N}^\ell(\epsilon, \mathcal{F})$ to the computation of a single "dimension-like" quantity. An overview of these various dimensions, some details of their history, and some examples of their computation can be found in (Anthony, 1997).

RBF-Networks With Infinite VC Dimension

Note that the 'plain' VC dimension is not appropriate in SV regression at all, as can be seen in the following: Denote r an arbitrary positive number and $C \in \mathbb{R}^n$ a compact set. Consider the class of functions

$$F := \left\{ f : f = \sum_i \alpha_i k(x_i, \cdot) \text{ with } x_i \in C, \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \leq r \right\} \tag{9.7}$$

We show that F has infinite VC dimension by showing that any arbitrary set $X =$

$\{x_1, \ldots, x_\ell\} \subset C$ of size ℓ can be shattered. Micchelli (1986) showed that the matrix $(k(x_i, x_j))_{ij}$ has full rank for Gaussian rbf-kernels. For arbitrary $\{y_1, \ldots y_\ell\} \in \{-1, 1\}$ there exists a function $f(\cdot) = \sum_i \alpha_i k(x_i, \cdot)$ with $f(x_i) = y_i$. Rescaling f finally yields some $\tilde{f} \in F$ which proves the statement.

In the present work, we view the class $\mathcal{F}$ as being induced by an operator $\bar{T}_k$ depending on some kernel function k. Thus $\mathcal{F}$ is the image of a "base class" $\mathcal{G}$ under $\bar{T}_k$. The analogy implicit in the picture is that the quantity that matters is the number of ϵ-distinguishable messages obtainable at the information sink. (Recall the equivalence up to a constant factor of packing and covering numbers.) In a typical communications problem, one tries to maximize the number of distinguisable messages (per unit time), in order to maximize the information transmission rate. But from the point of view of the receiver, the job is made easier the *smaller* the number of distinct messages that one needs to be concerned with decoding. The significance of the picture is that the kernel in question is exactly the kernel that is used, for example, in support vector machines. As a consequence, the determination of $\mathcal{N}^\ell(\epsilon, \mathcal{F})$ can be done in terms of properties of the operator $\bar{T}_k$. The latter thus plays a constructive role in controlling the complexity of $\mathcal{F}$ and hence the difficulty of the learning task. We believe that the new viewpoint in itself is potentially very valuable, perhaps more so than the specific results in the chapter. A further exploitation of the new viewpoint can be found in (Williamson et al., 1998b). There are in fact a variety of ways to define exactly what is meant by $\bar{T}_k$, and we have deliberately not been explicit in the picture. We make use of one particular $\bar{T}_k$ in this chapter.

Communication Theory Viewpoint

9.4 Generalization Bounds via Uniform Convergence

The generalization performance of learning machines can be bounded via uniform convergence results as in (Vapnik and Chervonenkis, 1981; Vapnik, 1979). A recent review can be found in (Anthony, 1997). The key thing about these results is the role of the covering numbers of the hypothesis class — the focus of the present chapter. Results for both classification and regression are now known. For the sake of concreteness, we quote below a result suitable for regression which was proved by Alon et al. (1997). See also chapter 4 for results on classification and pattern recognition. Let $P_m(f) := \frac{1}{\ell}\sum_{i=1}^{\ell} f(\mathbf{x}_i)$ denote the *empirical mean* of f on the sample $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$.

Uniform Convergence Bounds

Lemma 9.1 (Alon, Ben-David, Cesa-Bianchi, and Haussler, 1997) Let $\mathcal{F}$ be a class of functions from $\mathcal{C}$ into $[0, 1]$ and let P be a distribution over $\mathcal{C}$. then, for all $\epsilon > 0$ and all $m \geq \frac{2}{\epsilon^2}$,

$$\Pr\left\{\sup_{f \in \mathcal{F}} |P_\ell(f) - P(f)| > \epsilon\right\} \leq 12\ell \cdot \mathbf{E}\left[\mathcal{N}\left(\tfrac{\epsilon}{6}, \mathcal{F}, \ell_\infty^{\mathbf{X}^{2\ell}}\right)\right] e^{-\epsilon^2 \ell/36} \tag{9.8}$$

where Pr denotes the probability w.r.t. the sample $\mathbf{x}_1, \ldots, \mathbf{x}_\ell$ drawn i.i.d. from P, and $\mathbf{E}$ the expectation w.r.t. a second sample $\bar{\mathbf{X}}^\ell = (\bar{\mathbf{x}}_1^\top, \ldots, \bar{\mathbf{x}}_{2\ell}^\top)$ also drawn i.i.d. from P.

In order to use this lemma one usually makes use of the fact that for any P,

$$\mathbf{E}\left[\mathcal{N}(\epsilon, \mathcal{F}, \ell_\infty^{\bar{\mathbf{X}}^\ell})\right] \leq \mathcal{N}^\ell(\epsilon, \mathcal{F}). \tag{9.9}$$

The above result can be used to give a generalization error result by applying it to the loss-function induced class. The following Lemma, which is an improved version of (Bartlett et al., 1996, Lemma 17), is useful in this regard:

Lemma 9.2

Loss Functions

Denote $\mathcal{F}$ a set of functions from $\mathcal{C}$ to $[a, b]$ with $a < b$, $a, b \in \mathbb{R} \cup \pm\infty$ and $l : \mathbb{R} \to \mathbb{R}_0^+$ a loss function satisfying a Lipschitz-condition

$$l(\xi) - l(\xi') \leq C|\xi - \xi'| \text{ for all } \xi, \xi' \in [a - b, b - a]. \tag{9.10}$$

Moreover let $\mathbf{z} := (\mathbf{x}_i, y_i)_{j=1}^\ell$, $l_f|_{\mathbf{z}_j} := l(f(\mathbf{x}_j) - y_j)$, $l_f|_{\mathbf{z}} := (l_f|_{\mathbf{z}_j})_{j=1}^\ell$, $l_\mathcal{F}|_\mathbf{z} := \{l_f|_\mathbf{z} \colon f \in \mathcal{F}\}$ and $\mathcal{N}(\epsilon, l|_\mathbf{z}) := \mathcal{N}(\epsilon, l_\mathcal{F}|_\mathbf{z}, \ell_\infty^\mathbf{z})$. Then the following equation holds:

$$\max_{\mathbf{z} \in (\mathcal{C} \times [a,b])^\ell} \mathcal{N}(\epsilon, l|_\mathbf{z}) \leq \max_{\mathbf{x} \in \mathcal{C}^\ell} \mathcal{N}\left(\frac{\epsilon}{C}, \mathcal{F}|_\mathbf{x}\right) \tag{9.11}$$

Applying the result above to polynomial loss leads to the following corollary:

Corollary 9.1

Polynomial Loss

Let the assumptions be as above in lemma 9.2. Then for loss functions of type

$$l(\eta) = \frac{1}{p}\eta^p \text{ with } p > 1 \tag{9.12}$$

we have $C = (b - a)^{(p-1)}$, in particular $C = (b - a)$ for $p = 2$ and therefore

$$\max_{\mathbf{z} \in (\mathcal{C} \times [a,b])^\ell} \mathcal{N}(\epsilon, l|_\mathbf{z}) \leq \max_{\mathbf{x} \in \mathcal{C}^\ell} \mathcal{N}\left(\frac{\epsilon}{(b-a)^{p-1}}, \mathcal{F}|_\mathbf{x}\right) \tag{9.13}$$

One can readily combine the uniform convergence results with the above results to get overall bounds on generalization performance. We do not explicitly state such a result here since the particular uniform convergence result needed depends on the exact set-up of the learning problem. A typical uniform convergence result takes the form

$$P^\ell\{\sup_f |R_{emp}(f) - R(f)| > \epsilon\} \leq c_1(\ell)\mathcal{N}^\ell(\epsilon, \mathcal{F})e^{-\epsilon^\beta \ell / c_2}. \tag{9.14}$$

Even the exponent in (9.14) depends on the setting: In regression β can be set to 1, however in agnostic learning (Kearns et al., 1994) show that in general $\beta = 2$, except if the class is convex in which case it can be set to 1 (Lee et al., 1998). Since our primary interest is in determining $\mathcal{N}^\ell(\epsilon, \mathcal{F})$ we will not try to summarize the large body of work now done on uniform convergence results and generalization error.

Learning Curves

These generalization bounds are typically used by setting the right hand side equal to δ and solving for $\ell = \ell(\epsilon, \delta)$ (which is called the sample complexity). Another way to use these results is as a learning curve bound $\bar{\epsilon}(\delta, \ell)$ where

$$P^\ell\{\sup_f |R_{emp}(f) - R(f)| > \bar{\epsilon}(\delta, \ell)\} \leq \delta. \tag{9.15}$$

We note here that the determination of $\bar{\epsilon}(\delta, \ell)$ is quite convenient in terms of e_n, the dyadic entropy number associated with the covering number $\mathcal{N}^\ell(\epsilon, \mathcal{F})$ in (9.14). Setting the right hand side of (9.14) equal to δ, we have

$$\begin{aligned} \delta &= c_1(\ell)\mathcal{N}^\ell(\epsilon, \mathcal{F})e^{-\epsilon^\beta \ell / c_2} \\ \Rightarrow \log_2\left(\tfrac{\delta}{c_1(\ell)}\right) + \tfrac{\epsilon^\beta \ell}{c_2 \ln 2} &= \log_2 \mathcal{N}^\ell(\epsilon, \mathcal{F}) \\ \Rightarrow e_{\log_2\left(\frac{\delta}{c_1(\ell)}\right) + \frac{\epsilon^\beta \ell}{c_2 \ln 2} + 1} &= \epsilon. \end{aligned} \tag{9.16}$$

Thus $\bar{\epsilon}(\delta, \ell) = \{\epsilon \colon (9.16) \text{ holds}\}$. Thus the use of ϵ_n or e_n (which will arise naturally from our techniques) is in fact a convenient thing to do for finding learning curves.

9.5 Entropy Numbers for Kernel Machines

In the following we will mainly consider machines where the mapping into feature space is defined by Mercer kernels $k(\mathbf{x}, \mathbf{y})$ as they are easier to deal with using functional analytic methods. Such machines have become very popular due to the success of SV machines.

9.5.1 Mercer's Theorem, Feature Spaces and Scaling

Our goal is to make statements about the shape of the image of the input space $\mathcal{C}$ under the feature map $\Phi(\cdot)$. We will make use of Mercer's theorem. For various reasons we need a somewhat stronger statement than theorem 1.1, thus we repeat it as a whole for the sake of completeness. The version stated below is a special case of the theorem proven in (König, 1986, p. 145). In the following we will assume $(\mathcal{C}, \mu)$ to be a finite measure space, i.e. $\mu(\mathcal{C}) < \infty$. As usual, by "almost all" we mean for all elements of $\mathcal{C}^n$ except a set of μ^n-measure zero.

Theorem 9.1 (Mercer)
Suppose $k \in L_\infty(\mathcal{C}^2)$ to be symmetric, i.e. $k(x, x') = k(x', x)$, such that the integral operator $T_k : L_2(\mathcal{C}) \to L_2(\mathcal{C})$,

$$T_k f(\cdot) := \int_{\mathcal{C}} k(\cdot, \mathbf{y}) f(\mathbf{y}) d\mu(\mathbf{y}) \tag{9.17}$$

Bounded and Uniformly Convergent

is positive. Let $\psi_j \in L_2(\mathcal{C})$ be the eigenfunction of T_k associated with the eigenvalue $\lambda_j \neq 0$ and normalized such that $\|\psi_j\|_{L_2} = 1$.

1. $(\lambda_j(T))_j \in \ell_1$.

2. $\psi_j \in L_\infty(\mathcal{C})$ and $\sup_j \|\psi_j\|_{L_\infty} < \infty$.
3. $k(\mathbf{x}, \mathbf{y}) = \sum_{j \in \mathbb{N}} \lambda_j \psi_j(\mathbf{x}) \psi_j(\mathbf{y})$ holds for almost all $(\mathbf{x}, \mathbf{y})$, where the series converges absolutely and uniformly for almost all $(\mathbf{x}, \mathbf{y})$.

We will call a kernel satisfying the conditions of this theorem a *Mercer kernel.* From statement 2 of Mercer's theorem there exists some constant $C_k \in \mathbb{R}^+$ depending on $k(\cdot, \cdot)$ such that

$$|\psi_j(\mathbf{x})| \leq C_k \text{ for all } j \in \mathbb{N} \text{ and } \mathbf{x} \in \mathcal{C}. \tag{9.18}$$

Choice of Coordinate System

(Actually (9.18) holds only for almost all $\mathbf{x} \in \mathcal{C}$, but from here on we gloss over these measure-theoretic niceties in the exposition.) Moreover from statement 3 it follows that $k(\mathbf{x}, \mathbf{y})$ corresponds to a dot product in ℓ_2 i.e. $k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle_{\ell_2}$ with

$$\begin{array}{lcl} \Phi : \mathcal{C} & \to & \ell_2 \\ \mathbf{x} & \mapsto & (\phi_j(\mathbf{x}))_j := (\sqrt{\lambda_j} \psi_j(\mathbf{x}))_j \end{array} \tag{9.19}$$

for almost all $\mathbf{x} \in \mathcal{C}$. In the following we will (without loss of generality) assume the sequence of $(\lambda_j)_j$ be sorted in nonincreasing order. From the argument above one can see that $\Phi(\mathcal{C})$ lives not only in ℓ_2 but in an axis parallel parallelepiped with lengths $2C_k \sqrt{\lambda_j}$.

It will be useful to consider maps that map $\Phi(\mathcal{C})$ into balls of some radius R centered at the origin. The following proposition shows that the class of all these maps is determined by elements of ℓ_2 and the sequence of eigenvalues $(\lambda_j)_j$.

Proposition 9.1 (Mapping $\Phi(\mathbf{x})$ into ℓ_2)
Let S be the diagonal map

$$\begin{array}{l} S : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}} \\ S : (x_j)_j \mapsto S(x_j)_j = (s_j x_j)_j. \end{array} \tag{9.20}$$

Then S maps $\Phi(\mathcal{C})$ into a ball of finite radius R_S centered at the origin if and only if $(s_j \sqrt{\lambda_j})_j \in \ell_2$.

Proof ($\Leftarrow$) Suppose $(s_j \sqrt{\lambda_j})_j \in \ell_2$ and let $R_S^2 := C_k^2 \|(s_j \sqrt{\lambda_j})_j\|_{\ell_2}^2 < \infty$. For any $\mathbf{x} \in \mathcal{C}$,

$$\|S\Phi(\mathbf{x})\|_{\ell_2}^2 = \sum_{j \in \mathbb{N}} s_j^2 \lambda_j |\psi_j(\mathbf{x})|^2 \leq \sum_{j \in \mathbb{N}} s_j^2 \lambda_j C_k^2 = R_S^2. \tag{9.21}$$

Hence $S\Phi(\mathcal{C}) \subseteq \ell_2$.
($\Rightarrow$) Suppose $(s_j \sqrt{\lambda_j})_j$ is not in ℓ_2. Hence the sequence $(A_n)_n$ with $A_n := \sum_{j=1}^{n} s_j^2 \lambda_j$ is unbounded. Now define

$$a_n(\mathbf{x}) := \sum_{j=1}^{n} s_j^2 \lambda_j |\psi_j(\mathbf{x})|^2. \tag{9.22}$$

Then $\|a_n(\cdot)\|_{L_1(\mathcal{C})} = A_n$ due to the normalization condition on ψ_j. However, as

$\mu(\mathcal{C}) < \infty$, there exists a set $\tilde{\mathcal{C}}$ of nonzero measure such that

$$a_n(\mathbf{x}) \geq \frac{A_n}{\mu(\mathcal{C})} \quad \text{for all } \mathbf{x} \in \tilde{\mathcal{C}}. \tag{9.23}$$

Shape Bound on $\Phi(\mathcal{C})$

Combining the left side of (9.21) with (9.22) we obtain $\|S\Phi(\mathbf{x})\|_{\ell_2}^2 \geq a_n(\mathbf{x})$ for all $n \in \mathbb{N}$ and almost all $\mathbf{x}$. Since $a_n(\mathbf{x})$ is unbounded for a set $\tilde{\mathcal{C}}$ with nonzero measure in $\mathcal{C}$, we can see that $S\Phi(\mathcal{C}) \not\subset \ell_2$. ■

The consequence of this result is that there exists no *axis parallel* ellipsoid $\mathcal{E}$ not completely containing the (also) axis parallel parallelepiped $\mathcal{B}$ of sidelength $(2C_k\sqrt{\lambda_j})_j$, such that $\mathcal{E}$ would contain $\Phi(\mathcal{C})$. More formally

$$\mathcal{B} \subset \mathcal{E} \text{ if and only if } \Phi(\mathcal{C}) \subset \mathcal{E}. \tag{9.24}$$

Hence $\Phi(\mathcal{C})$ contains a set of nonzero measure of elements near the corners of the parallelepiped.

Once we know that $\Phi(\mathcal{C})$ "fills" the parallelepiped described above we can use this result to construct an inverse mapping A from the unit ball in ℓ_2 to an ellipsoid $\mathcal{E}$ such that $\Phi(\mathcal{C}) \subset \mathcal{E}$ as in the following diagram.

$$\begin{tikzcd}
\mathcal{C} \arrow[r, "\Phi"] & \Phi(\mathcal{C}) \arrow[r, "A^{-1}"] \arrow[d, phantom, "\cap"] & U_{\ell_2} \arrow[dl, "A"'] \\
 & \mathcal{E} &
\end{tikzcd} \tag{9.25}$$

Shrinkage Operator

The operator A will be useful for computing the entropy numbers of concatenations of operators. (Knowing the inverse will allow us to compute the forward operator, and that can be used to bound the covering numbers of the class of functions, as shown in the next subsection.) We thus seek an operator $A : \ell_2 \to \ell_2$ such that

$$A(U_{\ell_2}) \subseteq \mathcal{E}. \tag{9.26}$$

We can ensure this by constructing A such that

$$A\colon (x_j)_j \mapsto (R_A a_j x_j)_j \tag{9.27}$$

with $R_A := C_k \|(\sqrt{\lambda_j}/a_j)_j\|_{\ell_2}$. From Proposition 9.1 it follows that all those operators A for which $R_A < \infty$ will satisfy (9.26). We call such scaling (inverse) operators *admissible.*

9.5.2 Entropy Numbers

The next step is to compute the entropy numbers of the operator A and use this to obtain bounds on the entropy numbers for kernel machines like SV machines. We will make use of the following theorem due to Gordon et al. (1987), p. 226, stated in the present form in (Carl and Stephani, 1990, p. 17).

Theorem 9.2
Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_j \geq \cdots \geq 0$ be a non-increasing sequence of non-negative numbers and let

$$D\mathbf{x} = (\sigma_1 x_1, \sigma_2 x_2, \ldots, \sigma_j x_j, \ldots) \tag{9.28}$$

Diagonal Operator

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 $n \in \mathbb{N}$,

$$\sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}} \leq \epsilon_n(D) \leq 6 \sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}}. \tag{9.29}$$

We can exploit the freedom in choosing A to minimize an entropy number as the following corollary shows. This will be a key ingredient of our calculation of the covering numbers for SV classes, as shown below.

Corollary 9.2 (Entropy numbers for $\Phi(\mathcal{C})$)

Application to $\Phi(\mathcal{C})$

Let $k: \mathcal{C} \times \mathcal{C} \to \mathbb{R}$ be a Mercer kernel and let A be defined by (9.27). Then

$$\epsilon_n(A: \ell_2 \to \ell_2) \leq \inf_{(a_s)_s : \left(\sqrt{\lambda_s}/a_s\right)_s \in \ell_2} \sup_{j \in \mathbb{N}} 6 C_k \left\| \left(\sqrt{\lambda_s}/a_s\right)_s \right\|_{\ell_2} n^{-\frac{1}{j}} (a_1 a_2 \cdots a_j)^{\frac{1}{j}}. \tag{9.30}$$

This result follows immediately by identifying D and A and exploiting the freedom that we still have in choosing a particular operator A among the class of admissible ones.

As already described in section 9.1 the hypotheses that a SV machine generates can be expressed as $\langle \mathbf{w}, \tilde{\mathbf{x}} \rangle + b$ where both $\mathbf{w}$ and $\tilde{\mathbf{x}}$ are defined in the feature space $\mathcal{S} = \operatorname{span}(\Phi(\mathcal{C}))$ and $b \in \mathbb{R}$. The kernel trick as introduced by Aizerman et al. (1964) was then successfully employed by Boser et al. (1992) and Cortes and Vapnik (1995) to extend the Optimal Margin Hyperplane classifier to what is now known as the SV machine. We deal with the "$+b$" term in section 9.8; for now we consider the class

$$\mathcal{F}_{R_\mathbf{w}} := \{\langle \mathbf{w}, \tilde{\mathbf{x}} \rangle : \tilde{\mathbf{x}} \in \mathcal{S}, \|\mathbf{w}\| \leq R_\mathbf{w}\} \subseteq \mathbb{R}^{\mathcal{S}}. \tag{9.31}$$

Note that $\mathcal{F}_{R_\mathbf{w}}$ depends implicitly on k since $\mathcal{S}$ does.

What we seek are the ℓ_∞^m covering numbers for the class $\mathcal{F}_{R_\mathbf{w}}$ induced by the kernel in terms of the parameter $R_\mathbf{w}$ which is the inverse of the size of the margin in feature space, or equivalently, the size of the weight vector in feature space as defined by the dot product in $\mathcal{S}$ (see Vapnik and Chervonenkis (1974); Vapnik (1995) for details). In the following we will call such hypothesis classes with length constraint on the weight vectors in feature space *SV classes*. Let T be the operator $T = S_{\tilde{\mathbf{X}}^\ell} R_\mathbf{w}$ where $R_\mathbf{w} \in \mathbb{R}$ and the operator $S_{\tilde{\mathbf{X}}^\ell}$ is defined by

SV Classes

$$\begin{array}{lcl} S_{\tilde{\mathbf{X}}^\ell} : \ell_2 & \to & \ell_\infty^\ell \\ S_{\tilde{\mathbf{X}}^\ell} : \mathbf{w} & \mapsto & (\langle \tilde{\mathbf{x}}_1, \mathbf{w} \rangle, \ldots, \langle \tilde{\mathbf{x}}_\ell, \mathbf{w} \rangle). \end{array} \tag{9.32}$$

with $\tilde{\mathbf{x}}_j \in \Phi(\mathcal{C})$ for all j. The following theorem is useful to compute entropy numbers in terms of T and A. Originally due to Maurey it was extended by Carl (1985). See (Williamson et al., 1998b) for some extensions and historical remarks.

Maurey's theorem

Theorem 9.3 (Carl and Stephani, 1990, p. 246)
Let $S \in \mathfrak{L}(H, \ell_\infty^\ell)$ where H is a Hilbert space. Then there exists a constant $c > 0$ such that for all $\ell \in \mathbb{N}$, and $1 \le j \le \ell$

$$e_n(S) \le c\|S\| \left(n^{-1} \log_2 \left(1 + \frac{\ell}{n}\right)\right)^{1/2}. \tag{9.33}$$

An alternative proof of this result (given by Williamson et al. (1998b)) provides a small explicit value for the constant: $c = 2(\frac{6}{2-\log_2 3})^{1/2} \le 5.3771$.

The restatement of Theorem 9.3 in terms of $\epsilon_{2^{n-1}} = e_n$ will be useful in the following. Under the assumptions above we have

$$\epsilon_n(S) \le c\|S\| \left((\log_2 n + 1)^{-1} \log_2 \left(1 + \frac{\ell}{\log_2 n + 1}\right)\right). \tag{9.34}$$

Now we can combine the bounds on entropy numbers of A and $S_{\mathbf{X}^\ell}$ to obtain bounds for SV classes. First we need the following lemma.

Product Bounds

Lemma 9.3 (Carl and Stephani, 1990, p. 11)
Let E, F, G be Banach spaces, $R \in \mathfrak{L}(F, G)$, and $S \in \mathfrak{L}(E, F)$. Then, for $n, t \in \mathbb{N}$,

$$\epsilon_{nt}(RS) \le \epsilon_n(R)\epsilon_t(S) \tag{9.35}$$

$$\epsilon_n(RS) \le \epsilon_n(R)\|S\| \tag{9.36}$$

$$\epsilon_n(RS) \le \epsilon_n(S)\|R\|. \tag{9.37}$$

Note that the latter two inequalities follow directly from the fact that $\epsilon_1(R) = \|R\|$ for all $R \in \mathfrak{L}(F, G)$.

Theorem 9.4 Bounds for SV classes
Let k be a Mercer kernel, let Φ be induced via (9.19) and let $T := S_{\tilde{\mathbf{X}}^\ell} R_{\mathbf{w}}$ where $S_{\tilde{\mathbf{X}}^\ell}$ is given by (9.32) and $R_{\mathbf{w}} \in \mathbb{R}^+$. Let A be defined by (9.27) and suppose $\tilde{\mathbf{x}}_j = \Phi(\mathbf{x}_j)$ for $j = 1, \ldots, \ell$. Then the entropy numbers of T satisfy the following inequalities:

$$\epsilon_n(T) \le c\|A\|R_{\mathbf{w}} \log_2^{-1/2} n \log_2^{-1/2} \left(1 + \tfrac{\ell}{\log_2 n}\right) \tag{9.38}$$

$$\epsilon_n(T) \le 6 R_{\mathbf{w}} C_k \epsilon_n(A) \tag{9.39}$$

$$\epsilon_{nt}(T) \le 6 c C_k R_{\mathbf{w}} \log_2^{-1/2} n \log_2^{-1/2} \left(1 + \tfrac{\ell}{\log_2 n}\right) \epsilon_t(A)$$

where C_k and c are defined as in Corollary 9.2 and Lemma 9.3.

This result gives several options for bounding $\epsilon_n(T)$. The reason for using ϵ_n instead of e_n is that the index only may be integer in the former case (whereas it can be in $[1, \infty)$ in the latter), thus making it easier to obtain tighter bounds. We shall see in examples later that the best inequality to use depends on the rate of decay of the eigenvalues of k. The result gives effective bounds on $\mathcal{N}^\ell(\epsilon, \mathcal{F}_{R_{\mathbf{w}}})$ since

$$\epsilon_n(T\colon \ell_2 \to \ell_\infty^\ell) \le \epsilon_0 \;\Rightarrow\; \mathcal{N}^\ell(\epsilon_0, \mathcal{F}_{R_{\mathbf{w}}}) \le n. \tag{9.40}$$

Factorization

Proof We will use the following factorization of T to upper bound $\epsilon_n(T)$.

$$\begin{tikzcd}
U_{\ell_2} \arrow[r, "\mathsf{T}"] \arrow[d, "R_{\mathbf{w}}"'] & \ell_\infty^\ell \\
R_{\mathbf{w}} U_{\ell_2} \arrow[r, "A"] \arrow[ur, "S_{\tilde{\mathbf{X}}^\ell}"] & R_{\mathbf{w}} \mathcal{E} \arrow[u, "S_{(A^{-1}\tilde{\mathbf{X}}^\ell)}"']
\end{tikzcd} \tag{9.41}$$

The top left part of the diagram follows from the definition of T. The fact that the diagram commutes stems from the fact that since A is diagonal, it is self-adjoint and so

$$\langle \mathbf{w}, \tilde{\mathbf{x}} \rangle = \langle \mathbf{w}, AA^{-1}\tilde{\mathbf{x}} \rangle = \langle A\mathbf{w}, A^{-1}\tilde{\mathbf{x}} \rangle. \tag{9.42}$$

Instead of computing the covering number of $T = S_{\tilde{\mathbf{X}}^\ell} R_{\mathbf{w}}$ directly, which is difficult or wasteful, as the bound on $S_{\tilde{\mathbf{X}}^\ell}$ does not take into account that $\tilde{\mathbf{x}} \in \mathcal{E}$ but just makes the assumption of $\tilde{\mathbf{x}} \in \rho U_{\ell_2}$ for some $\rho > 0$, we will represent T as $S_{(A^{-1}\tilde{\mathbf{X}}^\ell)} A R_{\mathbf{w}}$. This is more efficient as we constructed A such that $\Phi(\mathcal{C})A^{-1} \in U_{\ell_2}$ filling a larger proportion of it than just $\frac{1}{\rho}\Phi(\mathcal{C})$.

By construction of A and the Cauchy-Schwarz inequality we have $\|S_{A^{-1}\tilde{\mathbf{X}}^\ell}\| = 1$. Thus applying lemma 9.3 to the factorization of T and using Theorem 9.3 proves the theorem. ■

As we shall see in section 9.7, one can give asymptotic rates of decay for $\epsilon_n(A)$. (In fact we give non-asymptotic results with explicitly evaluable constants.) It is thus of some interest to give overall asymptotic rates of decay of $\epsilon_n(T)$ in terms of the order of $\epsilon_n(A)$.

Overall Asymptotic Rates

Lemma 9.4 (Rate bounds on ϵ_n)
Let k be a Mercer kernel and suppose A is the scaling operator associated with it as defined by (9.27).

1. If $\epsilon_n(A) = O(\log_2^{-\alpha} n)$ for some $\alpha > 0$ then

$$\epsilon_n(T) = O(\log_2^{-(\alpha+2)} n). \tag{9.43}$$

2. If $\log_2 \epsilon_n(A) = O(\log_2^{-\beta} n)$ for some $\beta > 0$ then

$$\log_2 \epsilon_n(T) = O(\log_2^{-\beta} n). \tag{9.44}$$

This Lemma shows that in the first case, Maurey's result (theorem 9.3) allows an improvement in the exponent of the entropy number of T, whereas in the second, it affords none (since the entropy numbers decay so fast anyway). The Maurey result may still help in that case though for nonasymptotic n.

Proof From theorem 9.3 we know that $\epsilon_n(S) = O(\log_2^{-2} n)$. Now use (9.35), splitting the index n in the following way:

$$n = n^\tau n^{(1-\tau)} \text{ with } \tau \in (0,1). \tag{9.45}$$

For the first case this yields

Dominant Rates

$$\epsilon_n(T) = O(\log_2^{-2} n^\tau) O(\log_2^{-\alpha} j^{\tau-1}) = \tau^{-2}(1-\tau)^{-\alpha} O(\log_2^{-(\alpha+2)} n). \tag{9.46}$$

In the second case we have

$$\log_2 \epsilon_n(T) = \log_2 \left((\tau^{-2}) O(\log_2^{-2} n)\right) + (1-\tau)^{-\beta} O(\log_2^{-\beta} n) = O(\log_2^{-\beta} n). \tag{9.47}$$

■

In a nutshell we can always obtain rates of convergence better than those due to Maurey's theorem, because we are not dealing with *arbitrary* mappings into infinite dimensional spaces. In fact, for logarithmic dependency of $\epsilon_n(T)$ on n, the effect of the kernel is so strong that it completely dominates the $1/\epsilon^2$ behaviour for arbitrary Hilbert spaces. An example of such a kernel is $k(x, y) = \exp(-(x-y)^2)$; see Proposition 9.4 and also section 9.6 for the discretization question.

9.6 Discrete Spectra of Convolution Operators

The results presented above show that if one knows the eigenvalue sequence $(\lambda_i)_i$ of a compact operator, one can bound its entropy numbers. Whilst it is always possible to assume that the *data* fed into a SV machine have bounded support, the same can not be said of the kernel $k(\cdot,\cdot)$; a commonly used kernel is $k(x, y) = \exp(-(x-y)^2)$ which has noncompact support. The induced integral operator

Integral Operator

$$(T_k f)(x) = \int_{-\infty}^{\infty} k(x, y) f(y) dy \tag{9.48}$$

then has a continuous spectrum (a nondenumerable infinity of eigenvalues) and thus T_k is not compact (Ash, 1965, p.267). The question arises: can we make use of such kernels in SV machines and still obtain generalization error bounds of the form developed above? A further motivation stems from the fact that by a theorem of Widom (1964), the eigenvalue decay of any convolution operator defined on a a compact set via a kernel having compact support can decay no faster than $\lambda_j = O(e^{-j^2})$ and thus if one seeks very rapid decay of eigenvalues (with concomitantly small entropy numbers), one must use convolution kernels with noncompact support.

We will resolve these issues in the present section. Before doing so, let us first consider the case that $\operatorname{supp} k \subseteq [-a, a]$ for some $a < \infty$. Suppose further that the data points $\mathbf{x}_j$ satisfy $\mathbf{x}_j \in [-b, b]$ for all j. If $k(\cdot,\cdot)$ is a convolution kernel (i.e. $k(x,y) = k(x-y)$), then the SV hypothesis $h_k(\cdot)$ can be written

$$h_k(x) := \sum_{j=1}^{m} \alpha_j k(x, \mathbf{x}_j) = \sum_{j=1}^{m} \alpha_j k_v(x, \mathbf{x}_j) =: h_{k_v}(x) \tag{9.49}$$

for $v \geq 2(a+b)$ where $k_v(\cdot)$ is the v-periodic extension of $k(\cdot)$:

$$k_v(x) := \sum_{j=-\infty}^{\infty} k(x - jv). \tag{9.50}$$

We now relate the eigenvalues of T_{k_v} to the Fourier transform of $k(\cdot)$. We do so for the case of $d=1$ and then state the general case later.

Lemma 9.5

Connection between Spectrum and Fourier Transform

Let $k: \mathbb{R} \to \mathbb{R}$ be a symmetric convolution kernel, let $K(\omega) = F[k(x)](\omega)$ denote the Fourier transform of $k(\cdot)$ and k_v denote the v-periodical kernel derived from k (also assume that k_v exists). Then k_v has a representation as a Fourier series with $\omega_0 := \frac{2\pi}{v}$ and

$$\begin{aligned} k_v(x-y) &= \sum_{j=-\infty}^{\infty} \frac{\sqrt{2\pi}}{v} K(j\omega_0) e^{ij\omega_0 x} \\ &= \frac{\sqrt{2\pi}}{v} K(0) + \sum_{j=1}^{\infty} \frac{2}{v} \sqrt{2\pi} K(j\omega_0) \cos(j\omega_0(x-y)). \end{aligned} \tag{9.51}$$

Moreover $\lambda_j = \sqrt{2\pi} K(j\omega_0)$ for $j \in \mathbb{Z}$ and $C_k = \sqrt{\frac{2}{v}}$.

For a proof see (Williamson et al., 1998a). Thus even though T_k may not be compact, T_{k_v} may be (if $(K(j\omega_0))_{j \in \mathbb{N}} \subset \ell_2$ for example). The above lemma can be applied whenever we can form $k_v(\cdot)$ from $k(\cdot)$. Clearly $k(x) = O(x^{-(1+\epsilon)})$ for some $\epsilon > 0$ suffices to ensure the sum in (9.50) converges.

Let us now consider how to choose v. Note that the Riemann-Lebesgue lemma tells us that for integrable $k(\cdot)$ of bounded variation (surely any kernel one would use would satisfy that assumption), one has $K(\omega) = O(1/\omega)$. There is an tradeoff in choosing v in that for large enough ω, $K(\omega)$ is a decreasing function of ω (at least as fast as $1/\omega$) and thus by Lemma 9.5, $\lambda_j = \sqrt{2\pi} K(2\pi j/v)$ is an increasing function of v. This suggests one should choose a small value of v. But a small v will lead to high empirical error (as the kernel "wraps around" and its localization properties are lost) and large C_k. There are several approaches to picking a value of v. One obvious one is to *a priori* pick some $\tilde{\epsilon} > 0$ and choose the smallest v such that $|k(x) - k_v(x)| \leq \tilde{\epsilon}$ for all $x \in [-v/2, v/2]$. Thus one would obtain a hypothesis $h_{k_v}(x)$ uniformly within $C\tilde{\epsilon}$ of $h_k(x)$ where $\sum_{j=1}^{m} |\alpha_j| \leq C$.

Influence of Bandwidth

Finally it is worth explicitly noting how the choice of a different bandwidth of the kernel, i.e. letting $k^{(\sigma)}(\mathbf{x}) := \sigma k(\sigma \mathbf{x})$, affects the eigenspectrum of the corresponding operator. We have $K^{(\sigma)}(\boldsymbol{\omega}) = K(\boldsymbol{\omega}/\sigma)$, hence scaling a kernel by σ means more densely spaced eigenvalues in the spectrum of the integral operator $T_{k^{(\sigma)}}$.

9.7 Covering Numbers for Given Decay Rates

In this section we will show how the asymptotic behaviour of $\epsilon_n(A\colon \ell_2 \to \ell_2)$, where A is the scaling operator introduced before, depends on the eigenvalues of T_k.

A similar analysis has been carried out by Prosser (1966), in order to compute the entropy numbers of integral operators. However all of his operators mapped into $L_2(\mathcal{C}, \mathbb{C})$. Furthermore, whilst our propositions are stated as asympotic results as his were, the proofs actually give non-asympototic information with explicit constants. See (Williamson et al., 1998a) for details.

Note that we need to sort the eigenvalues in a nonincreasing manner because of the requirements in corollary 9.2. If the eigenvalues were unsorted one could obtain far too small numbers in the geometrical mean of $\lambda_1, \ldots, \lambda_j$. Many one-dimensional kernels have nondegenerate systems of eigenvalues in which case it is straightforward to explicitly compute the geometrical means of the eigenvalues as will be shown below. Note that whilst all of the examples below are for convolution kernels, i.e. $k(x,y) = k(x-y)$, there is nothing in the formulations of the propositions themselves that requires this. When we consider the d-dimensional case we shall see that with rotationally invariant kernels, degenerate systems of eigenvalues are generic. This can be dealt with by a slight modification of theorem 9.2 — see Williamson et al. (1998a) for details.

Let us consider the special case where $(\lambda_j)_j$ decays asymptotically with some polynomial or exponential degree. In this case we can choose a sequence $(a_j)_j$ for which we can evaluate (9.30) explicitly. By the eigenvalues of a kernel k we mean the eigenvalues of the induced integral operator T_k.

Laplacian Kernel

Proposition 9.2 (Polynomial Decay)
Let k be a Mercer kernel with eigenvalues satisfying $\lambda_j = \beta^2 i^{-(\alpha+1)}$ for some $\alpha > 0$. Then

$$\epsilon_n(A\colon \ell_2 \to \ell_2) = O\left((\ln n)^{-\frac{\alpha}{2}+O(\ln^{-2}\ln n)}\right) = O(\ln^{-\frac{\alpha}{2}} n). \tag{9.52}$$

An example of such a kernel is $k(x) = e^{-x}$.

Proposition 9.3 (Exponential Decay)
Suppose k is a Mercer kernel with eigenvalues $\lambda_j = \beta^2 e^{-\alpha(j-1)}$ for some $\alpha, \beta > 0$. Then

$$\ln \epsilon_n^{-1}(A\colon \ell_2 \to \ell_2) = O(\ln^{\frac{1}{2}} n) \tag{9.53}$$

An example of such a kernel is $k(x) = \frac{1}{1+x^2}$.

Gaussian Kernel

Proposition 9.4 (Exponential Quadratic Decay)
Suppose k is a Mercer kernel with $\lambda_j = \beta^2 e^{-\alpha(j-1)^2}$ for some $\alpha, \beta > 0$. Then

$$\ln \epsilon_n^{-1}(A\colon \ell_2 \to \ell_2) = O(\ln^{\frac{2}{3}} n). \tag{9.54}$$

An example of such a kernel is the Gaussian $k(x) = e^{-x^2}$. We conclude this section with a general relation between exponential-polynomial decay rates and orders of bounds on $\epsilon_n(A)$.

Proposition 9.5 (Exponential-Polynomial decay)
Suppose k is a Mercer kernel with $\lambda_j = \beta^2 e^{-\alpha j^p}$ for some $\alpha, \beta, p > 0$. Then

$$\ln \epsilon_n^{-1}(A: \ell_2 \to \ell_2) = O(\ln^{\frac{p}{p+1}} n) \tag{9.55}$$

This result is interesting but probably of little theoretical relevance as most practical kernels do not exhibit these rapid decay properties. (Recall the remarks at the beginning of section 9.6.)

Proposition 9.6
The rates given in propositions 9.2, 9.3, 9.4, and 9.5 are tight.

9.8 Conclusions

We have shown how to connect properties known about mappings into feature spaces with bounds on the covering numbers. Our reasoning relied on the fact that this mapping exhibits certain decay properties to ensure rapid convergence and a constraint on the size of the weight vector in feature space. This means that the corresponding algorithms have to restrict exactly this quantity to ensure good generalization performance. This is exactly what is done in Support Vector machines.

The actual application of our results, perhaps for model selection using structural risk minimization, is somewhat involved. Below we outline one possible path. As said before, the viewpoint in this chapter is new, and perhaps there will be refinements soon forthcoming which would make the codification of our existing results into a single generalization bound premature.

9.8.1 A Possible Procedure to use the Results of this Chapter

Choose k and σ The kernel k may be chosen for a variety of reasons, which we have nothing additional to say about here. The choice of σ should take account of the discussion in section 9.6.

Choose the period v of the kernel One suggested procedure is outlined in section 9.6.

Bound $\epsilon_n(A)$ This can be done using Corollary 9.2. Some examples of this sort of calculation are given in section 9.7.

Bound $\epsilon_n(T)$ Using Theorem 9.4.

Standard
SV Case

Take account of the "$+b$" The key observation is that given a class $\mathcal{F}$ with known $\mathcal{N}^\ell(\epsilon, \mathcal{F})$, one can bound $\mathcal{N}^\ell(\epsilon, \mathcal{F}^+)$ as follows. (Here $\mathcal{F}^+ := \{f + b: f \in \mathcal{F}, b \in \mathbb{R}\}$.) Suppose V_ϵ is an ϵ-cover for $\mathcal{F}$ and elements of $\mathcal{F}+$ are uniformly bounded by B (this implies a limit on $|b|$ as well as a uniform bound on elements of $\mathcal{F}$). Then

$$V_\epsilon^+ := \bigcup_{j=-B/\epsilon}^{B/\epsilon} V_\epsilon + j\epsilon \tag{9.56}$$

is an ϵ-cover for $\mathcal{F}^+$ and thus $\mathcal{N}^\ell(\epsilon, \mathcal{F}^+) \leq \frac{2B}{\epsilon}\mathcal{N}^\ell(\epsilon, \mathcal{F})$. Observe that this will only be "noticeable" for classes $\mathcal{F}$ with very slowly growing covering numbers (polynomial in $1/\epsilon$).

Take account of the loss function using Lemma 9.2 for example.

Plug into a uniform convergence result See the pointers to the literature and the example in section 9.4.

Classification and Pattern Recognition Together with a stratification of the hypothesis classes in terms of the margin in a data dependent fashion (Shawe-Taylor et al., 1998) the bounds could be used for classification.

Acknowledgements

This work was supported by the Australian Research Council, a grant of the DFG (# Ja 379/71), and NeuroColt II.