Generalization Performance of Regularization Networks and Support Vector Machines *via* Entropy Numbers of Compact Operators

Robert C. Williamson, ANU [1] Alex J. Smola, GMD [2]
Bernhard Schölkopf, GMD [3]

NeuroCOLT2 Technical Report Series

NC2-TR-1998-019

July, 1998[4]

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] Department of Engineering, Australian National University, Canberra, ACT 0200, Australia Bob.Williamson@anu.edu.au

[2] {smola@first.gmd.de} GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany

[3] bs@first.gmd.de GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany

[4] Received 07-JUL-98, Revised 21-OCT-98

Abstract

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.

Index terms: ϵ-entropy, covering numbers, statistical learning theory, support vector machines, linear operators.

1 Introduction

In this paper 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. [1, 30]), construct regression functions or decision boundaries based on this mapping, and use constraints in feature space for capacity control. Support Vector (SV) machines, which have recently been proposed as a new class of learning algorithms solving problems of pattern recognition, regression estimation, and operator inversion [47] 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. SV machines, like most kernel based methods, possess the nice property of defining the feature map in a manner that allows its computation implicitly at little additional computational cost. Our reasoning also applies to similar algorithms such as regularization networks [14] or certain unsupervised learning algorithms [38]. Let us now take a closer look at SV machines. Central to them are two ideas: capacity control by maximizing margins, and the use of nonlinear kernel functions.

Capacity control. In order to perform pattern recognition using linear hyperplanes, often a maximum margin of separation between the classes is sought for, as this leads to good generalization ability independent of the dimensionality [48, 47, 41]. It can be shown that for separable training data

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \in \mathbb{R}^d \times \{\pm 1\}, \tag{1}$$

this is achieved by minimizing $\|\mathbf{w}\|_2$ subject to the constraints $y_j(\langle \mathbf{w}, \mathbf{x}_j \rangle + b) \geq 1$ for $j = 1, \ldots, m$, and some $b \in \mathbb{R}$. The decision function then takes the form

$$f(\mathbf{x}) = \operatorname{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle + b). \tag{2}$$

Similarly, a linear regression

$$f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b \tag{3}$$

can be estimated from data

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \in \mathbb{R}^d \times \mathbb{R} \tag{4}$$

by finding the flattest function which approximates the data within some margin of error: in this case, one minimizes $\|\mathbf{w}\|_2$ subject to $|f(\mathbf{x}_j) - y_j| \leq \varepsilon$, where the parameter $\varepsilon > 0$ plays the role of the margin, albeit not in the space of the inputs $\mathbf{x}$, but in that of the outputs y. The analogy to pattern recognition is somewhat loose, and there exist alternative ways of introducing a margin (e.g. in the space $\mathbb{R}^m$ of all outputs $y_1, \ldots, y_m$, [53]).

In both cases, generalizations for the nonseparable or nonrealizable case exist, using various types of cost functions [12, 47, 42].

Nonlinear kernels. In order to apply the above reasoning to a rather general class of *nonlinear* functions, one can use kernels computing dot products in high-dimensional spaces nonlinearly related to input space [1, 8]. Under certain conditions on a kernel k, to be stated below (Theorem 4), there exists a nonlinear map Φ into a reproducing kernel Hilbert space F (see e.g. [36]) such that k computes the dot product in F, i.e.

$$k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle_F. \tag{5}$$

Given any algorithm which can be expressed in terms of dot products exclusively, one can thus construct a nonlinear version of it by substituting a kernel for the dot product. Examples of such machines include SV pattern recognition [8], SV regression estimation [47], and kernel principal component analysis [38].

By using the kernel trick for SV machines, the maximum margin idea is thus extended to a large variety of nonlinear function classes (e.g. radial basis function networks, polynomial networks, neural networks), which in the case of regression estimation comprise functions written as kernel expansions

$$f(\mathbf{x}) = \sum_{i=1}^{m} \alpha_j k(\mathbf{x}_j, \mathbf{x}) + b, \tag{6}$$

with $\alpha_j \in \mathbb{R}$, $j = 1, \ldots, m$. It has been noticed that different kernels can be characterized by their regularization properties [43]: SV machines are regularization networks minimizing the regularized risk $R_{reg}[f] = R_{emp}[f] + \frac{\lambda}{2}\|Pf\|^2$, (with a regularization parameter $\lambda \geq 0$, and a regularization operator P) over the set of functions of the form (6), 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.

1.1 Outline of the paper

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.

A key feature of the present paper 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 3.

The remainder of the paper is organized as follows. We start by introducing notation and definitions (Section 2). Section 4 formulates generalization error bounds in terms of covering numbers. Section 5 contains the main result bounding entropy numbers in terms of the spectrum of a given kernel. The results in this paper 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 6 shows how to make use of kernels such as $k(x) = e^{-x^2}$ which do not have a discrete spectrum. Section 7 presents some results on the entropy numbers obtained for given rates of decay of eigenvalues and 8 shows how to extend the results to several dimensions. The concluding section (Section 9) indicates how the various results in the paper can be glued together in order to obtain overall bounds on the generalization error.

We do not present a single master generalization error theorem for three key reasons: 1) the only novelty in the paper 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 [40] to apply the bounds.

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) [49] 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 more tight) 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.

2 Definitions and Notation

For $d \in \mathbb{N}$, $\mathbb{R}^d$ denotes the d-dimensional space of vectors $\mathbf{x} = (x_1, \ldots, x_d)$. 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 the ℓ_p^d norm is used in some papers in learning theory (e.g. [45]).

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

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

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

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) unit ball

$$U_E := \{x \in E\colon \|x\|_E \leq 1\} \tag{7}$$

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

$$\|T\| := \sup_{x \in U_E} \|Tx\|_F. \tag{8}$$

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

$$\begin{aligned} \epsilon_n(M) := \inf\{\epsilon > 0 \ &: \ \text{there exists an } \epsilon\text{-cover for } M \text{ in } E \\ & \quad \text{containing } n \text{ or fewer points}\} \end{aligned} \tag{9}$$

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

$$\epsilon_n(T) := \epsilon_n(T(U_E)). \tag{10}$$

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

$$e_n(T) := \epsilon_{2^{n-1}}(T), \qquad n \in \mathbb{N}; \tag{11}$$

similarly, the dyadic entropy numbers of a set are defined from its entropy numbers. A very nice introduction to entropy numbers of operators is [11]. 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 paper, 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 $\|x\|_H = \sqrt{\langle x, x \rangle_H}$.

By log and ln, we denote the logarithms to base 2 and e, respectively. By i, we denote the imaginary unit $i = \sqrt{-1}$, k will always be a kernel, and d and m will be the input dimensionality and the number of examples

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \in \mathbb{R}^d \times \mathbb{R}, \tag{12}$$

respectively. We will map the input data into a feature space via a mapping Φ. We let $\tilde{\mathbf{x}} := \Phi(\mathbf{x})$.

3 Operator Theory Methods for Entropy Numbers

In this section we briefly explain the new viewpoint implicit in the present paper. With reference to Figure 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 $\mathcal{X}$, and $\mathbf{x}_i \in \mathcal{X}$ for $i = 1, \ldots, m$, define the ϵ-*Growth function* of the function class $\mathcal{F}$ on $\mathcal{X}$ as

$$\mathcal{N}^m(\epsilon, \mathcal{F}) := \sup_{\mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathcal{X}} \mathcal{N}(\epsilon, \mathcal{F}, \ell_\infty^{\mathbf{X}^m}), \tag{13}$$

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

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). 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 $\mathcal{N}^m(\epsilon, \mathcal{F})$ 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 [5].

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

$$F := \left\{ f : f = \sum_i \alpha_i k(x_i, \cdot) \text{ with } x_i \in C, \sum_{i,j} \alpha_i \alpha_j k(x_i, x_j) \leq r \right\} \tag{14}$$

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. Since [28] the matrix

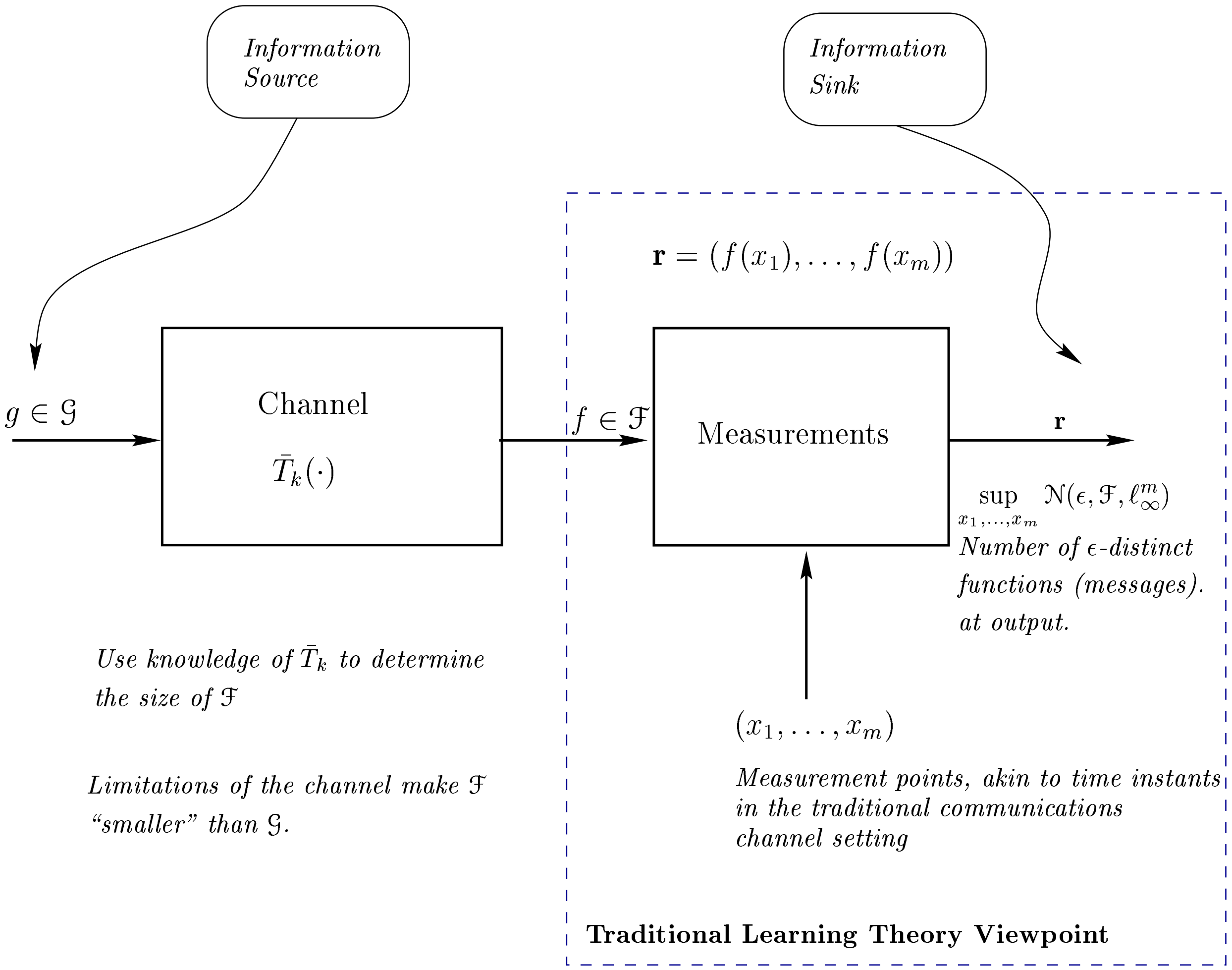
Information Source
Information Sink
$\mathbf{r} = (f(x_1), \ldots, f(x_m))$
$g \in \mathcal{G}$
Channel
$\bar{T}_k(\cdot)$
$f \in \mathcal{F}$
Measurements
$\mathbf{r}$
$\sup_{x_1,\ldots,x_m} \mathcal{N}(\epsilon, \mathcal{F}, \ell_\infty^m)$
Number of ϵ-distinct functions (messages). at output.
Use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

Figure 1 Schematic picture of the new viewpoint.

$[k(x_i, x_j)]_{ij}$ 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 $\mathcal{N}^m(\epsilon, \mathcal{F})$ 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 paper. A further exploitation of the new viewpoint can be found in [53]. 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 paper. A slightly different approach is taken in [53].

We conclude this section with some brief historical remarks.

The concept of the metric entropy of a set has been around for some time. It seems to have been introduced by Pontriagin and Schnirelmann [33] and was studied in detail by Kolmogorov and others [24]. The use of metric entropy to say something about linear operators was developed independently by several people. Prosser [34] appears to have been the first to make the idea explicit. He determined the effect of an operator's spectrum on its entropy numbers. In particular, he proved a number of results concerning the asymptotic rate of decrease of the entropy numbers in terms of the asymptotic behaviour of the eigenvalues. A similar result is actually implicit in section 22 of Shannon's famous paper [39], where he considered the effect of different convolution operators on the entropy of an ensemble. Prosser's paper [34] led to a handful of papers (see e.g. [35, 19, 3, 26]) which studied various convolutional operators. A connection between Prosser's ϵ-entropy of an operator and Kolmogorov's ϵ-entropy of a stochastic process was shown in [2]. Independently, another group of mathematicians including Carl and Stephani [11] studied covering numbers [46] and later entropy numbers [32] in the context of operator ideals. (They seem to be unaware of Prosser's work — see e.g. [9, p. 136].)

Connections between the local theory of Banach spaces and uniform convergence of empirical means has been noted before (e.g. [31]). More recently Gurvits [18] has obtained a result relating the Rademacher type of a Banach space to the fat-shattering dimension of linear functionals on that space and hence via the key result in [4] to the covering numbers of the induced class. We will make further remarks concerning the relationship between Gurvits' approach and ours in [53]; for now let us just note that the equivalence of the type of an operator (or of the space it maps to), and the rate of decay of its entropy numbers has been (independently) shown by Kolchinskiĭ [22, 23] and Defant and Junge [13, 20]. Note that the exact formulation of their results differs. Kolchinskiĭ was motivated by probabilistic problems not unlike ours.

4 Generalization Bounds via Uniform Convergence

The generalization performance of learning machines can be bounded via uniform convergence results as in [50, 49]. A recent review can be found in [5]. The key thing about these results is the role of the covering numbers of the hypothesis class — the focus of the present paper. 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 in [4]. Let $P_m(f) := \frac{1}{m}\sum_{i=1}^m f(\mathbf{x}_i)$ denote the *empirical mean* of f on the sample $\mathbf{x}_1, \ldots, \mathbf{x}_m$.

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

$$\Pr\left\{\sup_{f\in\mathcal{F}} |P_m(f) - P(f)| > \epsilon\right\} \leq 12m \cdot \mathbf{E}\left[\mathcal{N}\left(\frac{\epsilon}{6}, \mathcal{F}, \ell_\infty^{\bar{\mathbf{X}}^{2m}}\right)\right] e^{-\epsilon^2 m/36} \qquad (15)$$

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

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

$$\mathbf{E}\left[\mathcal{N}(\epsilon, \mathcal{F}, \ell_\infty^{\bar{\mathbf{X}}^m})\right] \leq \mathcal{N}^m(\epsilon, \mathcal{F}). \qquad (16)$$

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 [7, Lemma 17], is useful in this regard:

Lemma 2 *Denote $\mathcal{F}$ a set of functions from $\mathcal{X}$ 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. Let $\mathbf{z} := (\mathbf{x}_i, y_i)_{j=1}^m$, $l_f|_{\mathbf{z}_j} := l(f(\mathbf{x}_j) - y_j)$, $l_f|_{\mathbf{z}} := (l_f|_{\mathbf{z}_j})_{j=1}^m$, $l_{\mathcal{F}}|_{\mathbf{z}} := \{l_f|_{\mathbf{z}} \colon f \in \mathcal{F}\}$ and $\mathcal{N}(\epsilon, l|_{\mathbf{z}}) := \mathcal{N}(\epsilon, l_{\mathcal{F}}|_{\mathbf{z}}, \ell_\infty^{\mathbf{z}})$. Then the following two statements hold:*

1. *Suppose l satisfies the Lipschitz–condition*

$$l(\xi) - l(\xi') \leq C|\xi - \xi'| \text{ for all } \xi, \xi' \in [a-b, b-a]. \qquad (17)$$

Then for all $\epsilon > 0$

$$\max_{\mathbf{z}\in(\mathcal{X}\times[a,b])^m} \mathcal{N}(\epsilon, l|_{\mathbf{z}}) \leq \max_{\mathbf{x}\in\mathcal{X}^m} \mathcal{N}\left(\frac{\epsilon}{C}, \mathcal{F}|_{\mathbf{x}}\right) \qquad (18)$$

and

$$\max_{\mathbf{z}\in(\mathcal{X}\times[a,b])^m} \mathcal{N}(\epsilon, l|_{\mathbf{z}}) \leq \max_{\mathbf{x}\in\mathcal{X}^m} \mathcal{N}\left(\frac{\epsilon m}{C}, \mathcal{F}|_{\mathbf{x}}, l_1^{\mathbf{x}}\right). \qquad (19)$$

2. *Suppose that for some $C, \tilde{C} > 0$, l satisfies the "approximate Lipschitz–condition"*

$$l(\xi) - l(\xi') \leq \max(C|\xi - \xi'|, \tilde{C}) \text{ for all } \xi, \xi' \in [a-b, b-a] \tag{20}$$

then for all $\epsilon > \tilde{C}/C$

$$\max_{\mathbf{z} \in (\mathcal{X} \times [a,b])^m} \mathcal{N}(\epsilon, l|_{\mathbf{z}}) \leq \max_{\mathbf{x} \in \mathcal{X}^m} \mathcal{N}\left(\frac{\epsilon}{C}, \mathcal{F}|_{\mathbf{x}}\right). \tag{21}$$

Proof We show that, for any sequence $\mathbf{z}$ of $(\mathbf{x}, y)$ pairs in $\mathcal{X} \times [a,b]$ and any functions f and g, if the restrictions of f and g to $\mathbf{x}$ are close, then the restrictions of l_f and l_g to $\mathbf{z}$ are close. Thus, given a cover of $\mathcal{F}|_{\mathbf{x}}$ we can construct a cover of $l_{\mathcal{F}|\mathbf{z}}$ that is no bigger. For part 1 we get:

$$\begin{aligned}
\frac{1}{m}\left|\sum_{j=1}^m l(g(\mathbf{x}_j) - y_j) - l(f(\mathbf{x}_j) - y_j)\right| &\leq \frac{1}{m}\sum_{j=1}^m |l(g(\mathbf{x}_j - y_j)) - l(f(\mathbf{x}_j - y_j))| \\
&\leq \frac{1}{m}\sum_{j=1}^m C|g(\mathbf{x}_j) - f(\mathbf{x}_j)| \\
&= \frac{C}{m}\|g(\mathbf{X}^m) - f(\mathbf{X}^m)\|_{\ell_1^m} \\
&\leq C\|g(\mathbf{X}^m) - f(\mathbf{X}^m)\|_{\ell_\infty^m}.
\end{aligned}$$

In the second case we proceed similarly

$$\begin{aligned}
\frac{1}{m}\left|\sum_{j=1}^m l(g(\mathbf{x}_j - y_j)) - l(f(\mathbf{x}_j - y_j))\right| &\leq \frac{C}{m}\sum_{j=1}^m \max(|g(\mathbf{x}_j) - f(\mathbf{x}_j)|, \tilde{C}/C) \\
&\leq C\epsilon \quad \text{for } \epsilon \geq \tilde{C}/C.
\end{aligned}$$

∎

The second case can be useful, when the exact form of the cost function is not known, happens to be discontinuous or is badly behaved in some other way.[1] It shows how down to a scale $\tilde{C}/C$ statements about the covering numbers of the loss-function induced class can be made. Applying the result above to polynomial loss leads to the following corollary:

Corollary 3 *Let the assumptions be as above in lemma 2. Then for loss functions of type*

$$l(\eta) = \frac{1}{p}\eta^p \text{ with } p > 1 \tag{22}$$

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

$$\max_{\mathbf{z} \in (\mathcal{X} \times [a,b])^m} \mathcal{N}(\epsilon, l|_{\mathbf{z}}) \leq \max_{\mathbf{x} \in \mathcal{X}^m} \mathcal{N}\left(\frac{\epsilon}{(b-a)^{p-1} XC}, \mathcal{F}|_{\mathbf{x}}\right) \tag{23}$$

[1] The two cases could be combined into one by writing the conditions in terms of the modulus of continuity. For the sake of clarity , however, we refrained from doing so.

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

$$P^m\{\sup_f |R_{emp}(f) - R(f)| > \epsilon\} \leq c_1(m)\mathcal{N}^m(\epsilon, \mathcal{F})e^{-\epsilon^\beta m/c_2}. \tag{24}$$

Even the exponent in (24) depends on the setting: In regression β can be set to 1, however in agnostic learning [21] in general $\beta = 2$, except if the class is convex in which case it can be set to 1 [27]. Since our primary interest is in determining $\mathcal{N}^m(\epsilon, \mathcal{F})$ we will not try to summarize the large body of work now done on uniform convergence results and generalization error.

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

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

We note here that the determination of $\bar{\epsilon}(\delta, m)$ is quite convenient in terms of e_n, the dyadic entropy number associated with the covering number $\mathcal{N}^m(\epsilon, \mathcal{F})$ in (24). Setting the right hand side of (24) equal to δ, we have

$$\begin{aligned} & \delta && = c_1(m)\mathcal{N}^m(\epsilon, \mathcal{F})e^{-\epsilon^\beta m/c_2} \\ \Rightarrow\quad & \log\left(\frac{\delta}{c_1(m)}\right) + \frac{\epsilon^\beta m}{c_2 \ln 2} && = \log \mathcal{N}^m(\epsilon, \mathcal{F}) \\ \Rightarrow\quad & e_{\log\left(\frac{\delta}{c_1(m)}\right) + \frac{\epsilon^\beta m}{c_2 \ln 2} + 1} && = \epsilon. \end{aligned} \tag{25}$$

Thus $\bar{\epsilon}(\delta, m) = \{\epsilon \colon (25) \text{ holds}\}$. 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.

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. Nonetheless in Subsection 5.3 we will show how a more direct approach could be taken towards upper–bounding entropy numbers.

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 $\mathcal{X}$ under the feature map $\Phi(\cdot)$. We will make use of Mercer's theorem. The version stated below is a special case of the theorem proven in [25, p. 145]. In the following we will assume $(\mathcal{X}, \mu)$ to be a finite measure space, i.e. $\mu(\mathcal{X}) < \infty$. As usual, by "almost all" we mean for all elements of $\mathcal{X}^n$ except a set of μ^n-measure zero.

Theorem 4 (Mercer) *Suppose $k \in L_\infty(\mathcal{X}^2)$ is a symmetric kernel (hence $k(x, x') = k(x', x)$) such that the integral operator $T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$,*

$$T_k f(\cdot) := \int_{\mathcal{X}} k(\cdot, \mathbf{y}) f(\mathbf{y}) d\mu(\mathbf{y}) \tag{26}$$

is positive. Let $\psi_j \in L_2(\mathcal{X})$ 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. *$\psi_j \in L_\infty(\mathcal{X})$ and $\sup_j \|\psi_j\|_{L_\infty} < \infty$.*
3. *$k(\mathbf{x}, \mathbf{y}) = \sum\limits_{j \in \mathbb{N}} \lambda_j \psi_j(\mathbf{x}) \psi_j(\mathbf{y})$ 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

$$|\psi_j(\mathbf{x})| \leq C_k \text{ for all } j \in \mathbb{N} \text{ and } \mathbf{x} \in \mathcal{X}. \tag{27}$$

(Actually (27) holds only for almost all $\mathbf{x} \in \mathcal{X}$, 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

$$\begin{array}{lcl} \Phi : \mathcal{X} & \to & \ell_2 \\ \mathbf{x} & \mapsto & (\phi_j(\mathbf{x}))_j := (\sqrt{\lambda_j} \psi_j(\mathbf{x}))_j \end{array} \tag{28}$$

for almost all $\mathbf{x} \in \mathcal{X}$. 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 $\Phi(\mathcal{X})$ 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 $\Phi(\mathcal{X})$ 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 5 (Mapping $\Phi(\mathbf{x})$ into ℓ_2) *Let S be the diagonal map*

$$\begin{array}{l} S : \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}} \\ S : (x_j)_j \mapsto S(x_j)_j = (s_j x_j)_j. \end{array} \tag{29}$$

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

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 $\mathbf{x} \in \mathcal{X}$,

$$\|S\Phi(\mathbf{x})\|_{\ell_2}^2 = \sum_{j\in\mathbb{N}} s_j^2\lambda_j|\psi_j(\mathbf{x})|^2 \le \sum_{j\in\mathbb{N}} s_j^2\lambda_j C_k^2 = R_S^2. \tag{30}$$

Hence $S\Phi(\mathcal{X}) \subseteq \ell_2$.

($\Rightarrow$) Suppose $(s_j\sqrt{\lambda_j})_j$ is not in ℓ_2. Hence the sequence $(A_n)_n$ with $A_n := \sum\limits_{j=1}^{n} s_j^2\lambda_j$ is unbounded. Now define

$$a_n(\mathbf{x}) := \sum_{j=1}^{n} s_j^2\lambda_j|\psi_j(\mathbf{x})|^2. \tag{31}$$

Then $\|a_n(\cdot)\|_{L_1(\mathcal{X})} = A_n$ due to the normalization condition on ψ_j. However, as $\mu(\mathcal{X}) < \infty$ there exists a set $\tilde{\mathcal{X}}$ of nonzero measure such that

$$a_n(\mathbf{x}) \ge \frac{A_n}{\mu(\mathcal{X})} \quad \text{for all } \mathbf{x} \in \tilde{\mathcal{X}}. \tag{32}$$

Combining the left side of (30) with (31) we obtain $\|S\Phi(\mathbf{x})\|_{\ell_2}^2 \ge a_n(\mathbf{x})$ for all $n \in \mathbb{N}$ and almost all $\mathbf{x}$. Since $a_n(\mathbf{x})$ is unbounded for a set $\tilde{\mathcal{X}}$ with nonzero measure in $\mathcal{X}$, we can see that $S\Phi(\mathcal{X}) \not\subset \ell_2$. ∎

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 $\Phi(\mathcal{X})$. More formally

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

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

Once we know that $\Phi(\mathcal{X})$ "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 $\Phi(\mathcal{X}) \subset \mathcal{E}$ as in the following diagram.

$$\begin{tikzcd}
\mathcal{X} \arrow[r, "\Phi"] & \Phi(\mathcal{X}) \arrow[r, "A^{-1}"] \arrow[d, phantom, "\cap"] & U_{\ell_2} \arrow[dl, "A"] \\
& \mathcal{E} &
\end{tikzcd} \tag{33}$$

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

$$A(U_{\ell_2}) \subseteq \mathcal{E}. \tag{34}$$

We can ensure this by constructing A such that

$$A\colon (x_j)_j \mapsto (R_A a_j x_j)_j \tag{35}$$

with $R_A := C_k \|(\sqrt{\lambda_j}/a_j)_j\|_{\ell_2}$. From Proposition 5 it follows that all those operators A for which $R_A < \infty$ will satisfy (34). We call such scaling (inverse) operators *admissible.*

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, König and Schütt [15, p. 226] (stated in the present form in [11, p. 17]).

Theorem 6 *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*

$$D\mathbf{x} = (\sigma_1 x_1, \sigma_2 x_2, \ldots, \sigma_j x_j, \ldots) \tag{36}$$

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 $n \in \mathbb{N}$,

$$\sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}} \leq \epsilon_n(D) \leq 6 \sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{j}}. \tag{37}$$

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 7 (Entropy numbers for $\Phi(\mathcal{X})$) *Let $k\colon \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel and let A be defined by (35). Then*

$$\epsilon_n(A\colon \ell_2 \to \ell_2) \leq \inf_{(a_s)_s : \left(\sqrt{\lambda_s}/a_s\right)_s \in \ell_2} \sup_{j \in \mathbb{N}} 6 C_k \left\| \left(\sqrt{\lambda_s}/a_s\right)_s \right\|_{\ell_2} n^{-\frac{1}{j}} (a_1 a_2 \cdots a_j)^{\frac{1}{j}}. \tag{38}$$

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 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 $\mathcal{S} = \operatorname{span}(\Phi(\mathcal{X}))$ and $b \in \mathbb{R}$. The kernel trick as introduced by [1] was then successfully employed in [8] and [12] 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; for now we consider the class

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

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 [48, 47] 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 $T = S_{\tilde{\mathbf{X}}^m} R_{\mathbf{w}}$ where $R_{\mathbf{w}} \in \mathbb{R}$ and the operator $S_{\tilde{\mathbf{X}}^m}$ is defined by

$$\begin{array}{lcl} S_{\tilde{\mathbf{X}}^m} : \ell_2 & \to & \ell_\infty^m \\ S_{\tilde{\mathbf{X}}^m} : \mathbf{w} & \mapsto & (\langle \tilde{\mathbf{x}}_1, \mathbf{w}\rangle, \ldots, \langle \tilde{\mathbf{x}}_m, \mathbf{w}\rangle) . \end{array} \tag{39}$$

with $\tilde{\mathbf{x}}_j \in \Phi(\mathcal{X})$ for all j. The following theorem is useful when computing entropy numbers in terms of T and A. It is originally due to Maurey, and was extended by Carl [10]. See [53] for some extensions and historical remarks.

Theorem 8 (Carl and Stephani [11, p. 246]) *Let $S \in \mathfrak{L}(H, \ell_\infty^m)$ where H is a Hilbert space. Then there exists a constant $c > 0$ such that for all $m \in \mathbb{N}$, and $1 \leq j \leq m$*

$$e_n(S) \leq c\|S\| \left(n^{-1} \log\left(1 + \frac{m}{n}\right)\right)^{1/2} .$$

An alternative proof of this result (given in [53]) provides a small explicit value for the constant: $c = 2(\frac{6}{2-\log 3})^{1/2} \leq 5.3771$.

The restatement of Theorem 8 in terms of $\epsilon_{2^{n-1}} = e_n$ will be useful in the following. Under the assumptions above we have

$$\epsilon_n(S) \leq c\|S\| \left((\log n + 1)^{-1} \log\left(1 + \frac{m}{\log n + 1}\right)\right)^{1/2} . \tag{40}$$

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

Lemma 9 (Carl and Stephani [11, p. 11]) *Let E, F, G be Banach spaces, $R \in \mathfrak{L}(F, G)$, and $S \in \mathfrak{L}(E, F)$. Then, for $n, t \in \mathbb{N}$,*

$$\begin{aligned} \epsilon_{nt}(RS) &\leq \epsilon_n(R)\epsilon_t(S) && (41) \\ \epsilon_n(RS) &\leq \epsilon_n(R)\|S\| && (42) \\ \epsilon_n(RS) &\leq \epsilon_n(S)\|R\|. && (43) \end{aligned}$$

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 10 (Bounds for SV classes) *Let k be a Mercer kernel, let Φ be induced via (28) and let $T := S_{\tilde{\mathbf{X}}^m} R_{\mathbf{w}}$ where $S_{\tilde{\mathbf{X}}^m}$ is given by (39) and $R_{\mathbf{w}} \in \mathbb{R}^+$. Let A be defined by (35) and suppose $\tilde{\mathbf{x}}_j = \Phi(\mathbf{x}_j)$ for $j = 1, \ldots, m$. Then the entropy numbers of T satisfy the following inequalities:*

$$\begin{aligned} \epsilon_n(T) &\leq c\|A\|R_{\mathbf{w}} \log^{-1/2} n \log^{1/2}\left(1 + \tfrac{m}{\log n}\right) && (44) \\ \epsilon_n(T) &\leq 6R_{\mathbf{w}}\epsilon_n(A) && (45) \\ \epsilon_{nt}(T) &\leq 6cR_{\mathbf{w}} \log^{-1/2} n \log^{1/2}\left(1 + \tfrac{m}{\log n}\right) \epsilon_t(A) \end{aligned}$$

where C_k and c are defined as in Corollary 7 and Lemma 8.

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 $\mathcal{N}^m(\epsilon, \mathcal{F}_{R_{\mathbf{w}}})$ since

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

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

$$\begin{tikzcd}
U_{\ell_2} \arrow[r, "T"] \arrow[d, "R_{\mathbf{w}}"'] & \ell_\infty^m \\
R_{\mathbf{w}} U_{\ell_2} \arrow[r, "A"] \arrow[ur, "S_{\tilde{\mathbf{X}}^m}"] & R_{\mathbf{w}} \mathcal{E} \arrow[u, "S_{(A^{-1}\tilde{\mathbf{X}}^m)}"']
\end{tikzcd} \tag{46}$$

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

$$\langle \mathbf{w}, \tilde{\mathbf{x}} \rangle = \langle \mathbf{w}, AA^{-1}\tilde{\mathbf{x}} \rangle = \langle A\mathbf{w}, A^{-1}\tilde{\mathbf{x}} \rangle. \tag{47}$$

Instead of computing the covering number of $T = S_{\tilde{\mathbf{X}}^m} R_{\mathbf{w}}$ directly, which is difficult or wasteful, as the the bound on $S_{\tilde{\mathbf{X}}^m}$ 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 $S_{(A^{-1}\tilde{\mathbf{X}}^m)} A R_{\mathbf{w}}$. This is more efficient as we constructed A such that $A^{-1}\Phi(\mathcal{X}) \in U_{\ell_2}$ filling a larger proportion of it than just $\frac{1}{\rho}\Phi(\mathcal{X})$.

By construction of A and due to the Cauchy-Schwarz inequality we have $\|S_{A^{-1}\tilde{\mathbf{X}}^m}\| = 1$. Thus applying lemma 9 to the factorization of T and using Theorem 8 proves the theorem. ∎

As we shall see in Section 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 $\epsilon_n(A)$.

Lemma 11 (Rate bounds on ϵ_n) *Let k be a Mercer kernel and suppose A is the scaling operator associated with it as defined by (35).*

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

$$\epsilon_n(T) = O(\log^{-(\alpha+1/2)} n). \tag{48}$$

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

$$\log \epsilon_n(T) = O(\log^{-\beta} n). \tag{49}$$

This Lemma shows that in the first case, Maurey's result (theorem 8) 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 8 we know that $\epsilon_n(S) = O(\log^{-2} n)$. Now use (41), splitting the index n in the following way:

$$n = n^{\tau} n^{(1-\tau)} \text{ with } \tau \in (0,1). \tag{50}$$

For the first case this yields

$$\begin{aligned} \epsilon_n(T) &= O(\log^{-1/2} n^{\tau}) O(\log^{-\alpha} j^{\tau-1}) \\ &= \tau^{-1/2}(1-\tau)^{-\alpha} O(\log^{-(\alpha+1/2)} n) = O(\log^{-(\alpha+1/2)} n). \end{aligned} \tag{51}$$

In the second case we have

$$\log \epsilon_n(T) = \log\left((\tau^{-1/2}) O(\log^{-1/2} n)\right) + (1-\tau)^{-\beta} O(\log^{-\beta} n) = O(\log^{-\beta} n). \tag{52}$$

∎

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 16 and also Section 6 for the discretization question.

5.3 Empirical Bounds

Instead of theoretically determining the shape of $\Phi(\mathfrak{X})$ *a priori* one could use the training and/or test data to empirically estimate its shape and use this quantity to compute an operator B_{emp} analogously to (33) which performs the mapping described above. In this subsection we will sketch a possible approach — the full development of these ideas would requires considerable further work and will be deferred to a subsequent paper.

For instance assume that we are given an m–sample of datapoints $\mathbf{X} := \{\mathbf{x}_1, \ldots, \mathbf{x}_m\} \subset \mathfrak{X}$, not necessarily only from the training set but perhaps also comprising unlabelled test samples, drawn from the same distribution P. Now suppose we could estimate the first j radii (e.g. in a manner similar to the radius estimate in [37]) $\{r_1, \ldots, r_j\}$ of an ellipsoid enclosing $\Phi(\mathbf{X})$ with probability say $1-\eta$.[2] Denote by $\{\mathbf{e}_1, \ldots, \mathbf{e}_{j-1}\} \subset \ell_2$ a set of orthogonal vectors pointing in the directions given by the radii and P_j the projector onto $(\text{span}\{\mathbf{e}_1, \ldots, \mathbf{e}_{j-1}\})$. Note that we only have to be sure that with probability $1-\eta$ the data lies inside the ellipsoid and that we need no statement on the precision of the estimate

[2] This for instance could be done in a way similar to Kernel–PCA [43] by computing the eigensystem (λ_i, α_{ij}) of the Gram matrix $k_{ij} = k(x_i, x_j)$. Then, possibly after some ordering, $\sqrt{\lambda_i} \geq r_i = \sqrt{\lambda_i} \max_j |\alpha_{ij}|$.

of the radii — this makes a big difference in terms of the volume. Then with probability $1-\eta$ we could upper bound the covering numbers of a scaling operator B_{emp} by making use of corollary 7. Due to the ellipsoid condition the following inequality holds for all $\mathbf{x}_s$:

$$\sum_{t=1}^{j-1} \frac{\langle \mathbf{e}_t, \Phi(\mathbf{x}_t)\rangle^2}{r_t^2} \leq 1 \tag{53}$$

and moreover $\|P_j \Phi(\mathbf{x}_t)\|_{\ell_2} \leq r_j$ for all $1 \leq t \leq M$. Hence for an operator ${B_{\text{emp}}}^{-1}$ scaling the first $j-1$ directions $\mathbf{e}_1, \ldots, \mathbf{e}_{j-1}$ by r_t^{-1} and the rest by r_j^{-1}, ${B_{\text{emp}}}^{-1}\Phi(\mathbf{X})$ would still be enclosed in $\sqrt{2}U_{\ell_2}$. Hence we have a similar situation as in the case where we explicitly computed all eigenvalues analytically. Setting $r_t := r_j$ for $t > j$ and applying corollary 7 leads to the following upper bound on the entropy of B_{emp}

$$\epsilon_n(B_{\text{emp}}) \leq 6\sqrt{2} \sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} (r_1 r_2 \cdots r_j)^{\frac{1}{j}} \tag{54}$$

The aim is now to find the maximum n for which the estimate will not break down yet (for we have the liberty of distributing the covering numbers arbitrarily between the shrinkage operator B_{emp} and the actual evaluation operator $S_{\mathbf{X}^m}$ as shown in section 5). In other words we are looking for that particular value of j where $\sup_j$ is taken on for the smallest radius estimated. Ignoring the fact that $n \in \mathbb{N}$ for a moment we arrive at the following equation:

$$n^{-\frac{1}{j}} (r_1 r_2 \cdots r_j)^{\frac{1}{j}} = n^{-\frac{1}{j+1}} (r_1 r_2 \cdots r_j \cdot r_j)^{\frac{1}{j+1}} \tag{55}$$

Solving for n yields and taking $n \in \mathbb{N}$ into account yields

$$n_j = \left\lfloor \frac{r_1 r_2 \cdots r_j}{r_j^j} \right\rfloor \text{ and therefore } \epsilon_{n_j+1}(B_{\text{emp}}) \leq 6\sqrt{2} r_j \tag{56}$$

This calculation is valid as n_j is a nondecreasing function of j: $\frac{n_{j+1}}{n_j} = \left(\frac{r_j}{r_{j+1}}\right)^j$. If this assumption failed to be true one would have to redefine B_{emp} to scale only the first $\tilde{j}$ directions for which this happened to hold — it gains one little to scale in directions where the decay rate is too slow.

Instead of taking real data (which may be expensive to get) we also could upper bound the first j radii by a Monte–Carlo method, once we can bound the set $\mathcal{X}$. This is also useful when no analytic expansion in terms of eigenvalues of the operator can be obtained or where it would be too tedious to obtain explicitly. In cases with a sufficient amount of computational power available this may even be a more practical and faster way than computing the spectrum given by k analytically. The latter, at least in order to obtain optimal bounds, would have to be done for each learning problem anew. The method proposed here would obviate the need for such detailed theoretical calculations which may be impractical to carry out in some instances.

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

$$(T_k f)(x) = \int_{-\infty}^{\infty} k(x,y) f(y) dy \tag{57}$$

then has a continuous spectrum (a nondenumerable infinity of eigenvalues) and thus T_k is not compact [6, 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 [52], 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

$$h_k(x) := \sum_{j=1}^{m} \alpha_j k(x,\mathbf{x}_j) = \sum_{j=1}^{m} \alpha_j k_v(x,\mathbf{x}_j) =: h_{k_v}(x) \tag{58}$$

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

$$k_v(x) := \sum_{j=-\infty}^{\infty} k(x-jv). \tag{59}$$

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 12 *Let $k\colon \mathbb{R} \to \mathbb{R}$ be a symmetric convolution kernel, let $K(\omega) = F[k(x)](\omega)$ denote the Fourier transform of $k(\cdot)$ (see (69)) 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*

$$\begin{aligned} k_v(x-y) &= \sum_{j=-\infty}^{\infty} \tfrac{\sqrt{2\pi}}{v} K(j\omega_0) e^{ij\omega_0 x} \\ &= \tfrac{\sqrt{2\pi}}{v} K(0) + \textstyle\sum_{j=1}^{\infty} \tfrac{2}{v} \sqrt{2\pi} K(j\omega_0) \cos(j\omega_0(x-y)). \end{aligned} \tag{60}$$

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

Proof Clearly the Fourier series coefficients K_j of k_v exist (as k_v exists) with

$$K_j := \frac{1}{\sqrt{v}} \int_{v/2}^{v/2} e^{-ij\omega_0 x} k_v(x) dx$$

and therefore by the definition of k_v and the existence of $K(\omega)$ we conclude

$$\begin{aligned} K_j &= \tfrac{1}{\sqrt{v}} \int_{v/2}^{v/2} \textstyle\sum_{j=-\infty}^{\infty} e^{-ij\omega_0 x} k(x - jv) \\ &= \tfrac{1}{\sqrt{v}} \textstyle\sum_{j=-\infty}^{\infty} \int_{v/2}^{v/2} e^{-ij\omega_0 x} k(x - jv) = \sqrt{\tfrac{2\pi}{v}} K(j\omega_0). \end{aligned}$$

This and the fact that $\{x \mapsto v^{-1/2} e^{ij\omega_0 x} : j \in \mathbb{Z}\}$ forms an orthogonal basis in $L_2([-\frac{v}{2}, \frac{v}{2}], \mathbb{C})$ proves (60). (Note that from $k(x) = k(-x)$ we conclude $\overline{K(\omega)} = K(-\omega)$). Furthermore, we are interested in real valued basis functions for $k(x - y)$. The functions

$$\begin{aligned} \psi_0(x) &:= \tfrac{1}{\sqrt{v}} \\ \psi_j(x) &:= \sqrt{\tfrac{2}{v}} \cos(j\omega_0 x) \text{ and } \psi_{-j}(x) := \sqrt{\tfrac{2}{v}} \sin(j\omega_0 x) \text{ for all } j \in \mathbb{N} \end{aligned} \tag{61}$$

form an eigensystem of the integral operator defined by k_v with the corresponding eigenvalues $\sqrt{2\pi} K(j\omega_0)$. Finally one can see that $C_k = \sqrt{\frac{2}{v}}$ by computing the max over $j \in \mathbb{N}$ and $x \in [-v/2, v/2]$. ∎

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 (59) 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 12, $\lambda_j = \sqrt{2\pi} K(2\pi j/v)$ 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 $|k(x) - k_v(x)| \le \tilde{\epsilon}$ 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 $\sum_{j=1}^{m} |\alpha_j| \le C$.

Remark 13 *The above Lemma can be readily extended to d dimensions. Assume $k(\mathbf{x})$ is v-periodic in each direction ($\mathbf{x} = (x_1, \ldots, x_d)$), we get*

$$\lambda_{\mathbf{j}} = (2\pi)^{\frac{d}{2}} K(\omega_0 \mathbf{j}) = (2\pi)^{\frac{d}{2}} K(\omega_0 \|\mathbf{j}\|) \tag{62}$$

for radially symmetric k and finally for the eigenfunctions $C_k = (2/v)^{\frac{d}{2}}$.

Finally it is worth explicitly noting how the choice of a different bandwidth of the kernel, i.e. letting $k^{(\sigma)}(\mathbf{x}) := \sigma^d k(\sigma\mathbf{x})$, 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 $T_{k^{(\sigma)}}$.

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 [34], in order to compute the entropy numbers of integral operators. However all of his operators mapped into $L_2(\mathcal{X}, \mathbb{C})$. Furthermore, whilst our propositions are stated as asympotic results as his were, the proofs actually give non-asympototic information with explicit constants.

Note that we need to sort the eigenvalues in a nonincreasing manner because of the requirements in corollary 7. 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. In section 8.2 we will show how to systematically deal with that case.

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 (38) explicitly. By the eigenvalues of a kernel k we mean the eigenvalues of the the induced integral operator T_k.

Proposition 14 (Polynomial Decay) *Let k be a Mercer kernel with eigenvalues satisfying $\lambda_j = \beta^2 i^{-(\alpha+1)}$ for some $\alpha > 0$. Then*

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

An example of such a kernel is $k(x) = e^{-x}$. The proof is in the appendix.

Proposition 15 (Exponential Decay) *Suppose k is a Mercer kernel with eigenvalues $\lambda_j = \beta^2 e^{-\alpha(j-1)}$ for some $\alpha, \beta > 0$. Then*

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

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

Proposition 16 (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*

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

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

Proposition 17 (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*

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

See the appendix for a proof sketch. 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 6.)

Proposition 18 *The rates given in propositions 14, 15, 16, and 17 are tight.*

Proof We start with proposition 14. Carl and Stephani [11, Proposition 1.5.1] show that the dyadic entropy numbers e_k and the eigenvalues of the corresponding diagonal operators on ℓ_p scale in an identical manner (in the sense that they are members of the same Lorentz sequence space $\ell_{s,t}$). In our case this means that they have the same polynomial rate of decay. Hence the bound in Proposition 14 is tight for the operator A we assumed. Moreover from proposition 5 we conclude that it is impossible to use another operator, say A' that would have a faster rate of decay than A.

For the other propositions we have to do some more work, however it suffices to show tightness for proposition 17 as the other cases are just a special case thereof. Our proof relies on Equation 37 of theorem 6 as this also provides a lower bound on $\epsilon_n(A)$ in terms of the eigenvalues of A. Analogously to theorem 7 one can show that

$$\epsilon_n(A\colon \ell_2 \to \ell_2) \geq \inf_{(a_s)_s\colon\left(\sqrt{\lambda_s}/a_s\right)_s \in \ell_2} \sup_{j \in \mathbb{N}} C_k \left\| \left(\sqrt{\lambda_s}/a_s\right)_s \right\|_{\ell_2} n^{-\frac{1}{j}} (a_1 a_2 \cdots a_j)^{\frac{1}{j}} \tag{63}$$

The ℓ_2–norm can be bounded below by λ_1/a_1^2 which we can set to β, without loss of generality (as choosing the first scaling factors does not influence the rate at all). For any operator with diagonal scaling coefficients $a_j = e^{-\frac{\tau}{2} j^p}$ one can find a constant ϕ_1 such that

$$(a_1 a_2 \ldots a_j)^{\frac{1}{j}} = e^{-\frac{1}{2j}\tau \sum_{s=1}^{j} s^p} \geq e^{-\tau \phi_1 j^p}. \tag{64}$$

Now computing the $\sup_j$ in (63) yields $j = \phi_2 \tau^{-\frac{1}{p+1}} \ln^{\frac{1}{p+1}} n$ for some positive constant ϕ_2 and finally for some $\phi_3 > 0$

$$\epsilon_n(A\colon \ell_2 \to \ell_2) \geq \lambda_1 C_k \inf_{\tau \in [0, \alpha/2]} e^{-\phi_3 \tau^{\frac{1}{p+1}} \ln^{\frac{p}{p+1}} n}. \tag{65}$$

The $\inf_\tau$ is obtained for $\tau = \alpha/2$ and consequently for some $\phi_4, \phi_5 > 0$

$$\epsilon_n(A{:}\, \ell_2 \to \ell_2) \geq \phi_4 e^{-\phi_4 \ln^{\frac{p}{p+1}} n} \tag{66}$$

which gives the claimed rate. Note that due to proposition 5 it is impossible to get any operator A with a faster rate of decay than the one for $\tau = \alpha/2$. This shows it was sufficient to consider only this specific parametric family of operators A and therefore the rates are tight for arbitrary A. ∎

8 Higher Dimensions

Things get a little bit more complicated in higher dimensions. There are basically two ways that can be pursued for constructing kernels in $\mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ with $d > 1$ if no particular assumptions on the data we are dealing with are made. Firstly one could construct kernels by

$$k(\mathbf{x} - \mathbf{y}) = k(x_1 - y_1) \times \cdots \times k(x_d - y_d). \tag{67}$$

This choice will usually lead to preferred directions in input space as the kernels are not rotationally invariant in general. The second approach consists in setting

$$k(\mathbf{x} - \mathbf{y}) = k(\|\mathbf{x} - \mathbf{y}\|_{\ell_2}). \tag{68}$$

This approach also leads to translationally invariant kernels which are also rotationally invariant. In the following we will exploit this approach to compute regularization operators and corresponding Green's functions. It is quite straightforward, however, to generalize our exposition to the rotational asymmetric case. Now let us define the basic ingredients needed for the further calculations.

8.1 Basic Tools

The d-dimensional Fourier transform is defined by

$$F : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d) \text{ with } F[f](\boldsymbol{\omega}) := \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{-i\langle \boldsymbol{\omega}, \mathbf{x}\rangle} f(\mathbf{x}) d\mathbf{x}. \tag{69}$$

Then its inverse transform is given by

$$F^{-1} : L_2(\mathbb{R}^d) \to L_2(\mathbb{R}^d) \text{ with } F^{-1}[f](\mathbf{x}) = \frac{1}{(2\pi)^{d/2}} \int_{\mathbb{R}^d} e^{i\langle \boldsymbol{\omega}, \mathbf{x}\rangle} f(\boldsymbol{\omega}) d\boldsymbol{\omega}. \tag{70}$$

F can be shown to be an isometry on $L_2(\mathbb{R}^d)$.

Now introduce regularization operators P defined by

$$\langle Pf, Pg\rangle := \int_{\operatorname{supp} P(\boldsymbol{\omega})} \frac{\overline{F[f](\boldsymbol{\omega})} F[g](\boldsymbol{\omega})}{P(\boldsymbol{\omega})} d\boldsymbol{\omega} \tag{71}$$

for some nonnegative function $P(\boldsymbol{\omega})$ converging to 0 for $\|\boldsymbol{\omega}\| \to \infty$. It can be shown [43] that for a kernel to be a Green's function of P^*P, i.e.

$$\langle Pk(\mathbf{x}), Pk(\mathbf{x} - \mathbf{x}_0)\rangle = k(\mathbf{x}_0), \tag{72}$$

we need $F[k](\boldsymbol{\omega}) = P(\boldsymbol{\omega})$. For radially symmetric functions, i.e. $f(\mathbf{x}) = f(\|\mathbf{x}\|_2)$, we can explicitly carry out the integration on the sphere to obtain Fourier transform which is also radially symmetric (see e.g. [44, 29]), namely

$$F[f](\|\omega\|) = \omega^{-\nu} H_\nu[r^\nu f(r)](\|\omega\|), \tag{73}$$

where $\nu := \frac{1}{2}d - 1$ and H_ν is the Hankel transform over the positive real line. The latter is defined by

$$H_\nu[f](\omega) := \int_0^\infty r f(r) J_\nu(\omega r) dr. \tag{74}$$

Here J_ν is the Bessel function of the first kind defined by

$$J_\nu(r) := r^\nu 2^{-\nu} \sum_{j=0}^{\infty} \frac{(-1)^j r^{2j}}{2^{2j} j! \Gamma(j + \nu + 1)}. \tag{75}$$

Note that $H_\nu = H_\nu^{-1}$, i.e. $f = H_\nu[H_\nu[f]]$ (in L_2) due to the Hankel inversion theorem [44].

8.2 Degenerate Systems

Computing the Fourier transform for a given kernel k gives us the continuous spectrum. As pointed out in Section 6, we are interested in the discrete spectrum of integral kernels defined on $\mathcal{X}$. This means that the eigenvalues are defined on the grid $\omega_0 \mathbb{Z}^d$ with $\omega_0 = 2\pi/v$. Assuming $k(\mathbf{x})$ is rotationally invariant, so is $K(\boldsymbol{\omega})$ and therefore also the eigenvalues $\lambda_{\mathbf{j}} = (2\pi)^{\frac{d}{2}} K(\mathbf{j}\omega_0)$ as shown in Lemma 12. Consequently we have degeneracies in the point spectrum of the integral operator given by k (or k_v respectively) as all $\mathbf{j}\omega_0$ with equal length will have the same eigenvalue. In order to deal with this case efficiently we slightly modify Theorem 6 for our purposes. The following theorem allows proper account to be taken of the multiplicity of eigenvalues, and thus allows the straight-forward calculation of the sought for entropy numbers.

Theorem 19 *Let $(s_t)_t \in \mathbb{N}^{\mathbb{N}_0}$ be an increasing sequence with $s_0 = 1$ and $(\sigma_j)_j \in \mathbb{R}^{\mathbb{N}}$ be a non–increasing sequence of non–negative numbers with*

$$\sigma_j < \sigma_{\bar{j}} \text{ for } j < \bar{j} \text{ and } \sigma_j = \sigma_{s_t} \text{ for } s_{t-1} < j \leq s_t$$

and let

$$D\mathbf{x} = (\sigma_1 x_1, \sigma_2 x_2, \ldots, \sigma_j x_j, \ldots) \tag{76}$$

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 $n \in \mathbb{N}$,

$$\sup_{t \in \mathbb{N}} n^{-\frac{1}{s_t}} (\sigma_1 \sigma_2 \cdots \sigma_{s_t})^{\frac{1}{s_t}} \leq \epsilon_n(D) \leq 6 \sup_{t \in \mathbb{N}} n^{-\frac{1}{s_t}} (\sigma_1 \sigma_2 \cdots \sigma_{s_t})^{\frac{1}{s_t}}. \tag{77}$$

See the appendix for a proof.

This theorem allows us to obtain a similar result to corollary 7.

Corollary 20 (Entropy numbers for degenerate systems) *Let $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel and let A be defined by (35) with the additional restriction that the coefficients a_j have to match the degeneracy of λ_j. Then*

$$\epsilon_n(A: \ell_2 \to \ell_2) \leq \inf_{(a_j)_j : \left(\sqrt{\lambda_j}/a_j\right)_j \in \ell_2} \sup_{t \in \mathbb{N}} 6C_k \left\| \left(\sqrt{\lambda_j}/a_j\right)_j \right\|_{\ell_2} n^{-\frac{1}{s_t}} (a_1 a_2 \dots a_{s_t})^{\frac{1}{s_t}} \tag{78}$$

This result by itself may not appear too useful. However this is exactly what we need for the degenerate case (it is slightly tighter than the original statement, as the sup effectively has to be carried out only over a subset of $\mathbb{N}$). Finally we have to compute the degree of multiplicity that occurs for different indices $\mathbf{j}$. For this purpose consider shells of radius r in $\mathbb{R}^d$ centered at the origin, i.e. rS^{d-1}, which contain a nonzero number of elements of $\mathbb{Z}^d$. Denote the corresponding radii by r_j and let $n(r_j, d)$ be the number of elements on these shells. Observe that $n(r, d) \neq 0$ only when $r^2 \in \mathbb{N}$. Thus

$$\begin{array}{lcl} n(r,d) & := & |\mathbb{Z}^d \cap rS^{d-1}| \\ N(r,d) & := & \sum_{\{0 \leq \rho \leq r : \rho^2 \in \mathbb{N}\}} n(\rho, d). \end{array} \tag{79}$$

The determination of $n(r, d)$ is a classical problem which is completely solved by the use of the θ-series. (see e.g. [17]):

Theorem 21 (Occupation numbers of shells) *Let the formal power series $\theta(x)$ be defined by*

$$\theta(x) := \sum_{j=-\infty}^{\infty} x^{j^2} = 1 + 2\sum_{j=1}^{\infty} x^{j^2}. \tag{80}$$

Then

$$(\theta(x))^d = \sum_{j=1}^{\infty} n(\sqrt{j}, d) x^j. \tag{81}$$

This theorem allows one to readily compute $n(r, d)$ exactly; see the appendix for some Maple code to do so. (Note that whilst there do exist closed form asymptotic approximate formulae for $n(r, d)$ [17, p. 155], they are inordinately complicated and of little use for our purposes.)

We can now construct an index of the eigenvalues which satisfies the required ordering (at least for nonincreasing functions $K(\omega)$) and we get the following result:

Corollary 22 (Radially Symmetric Systems on a Lattice) *Let $k: \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a Mercer kernel with eigenvalues given by a radially symmetric nonincreasing function on a lattice, i.e. $\lambda_{\mathbf{j}} = \lambda(\|\mathbf{j}\|)$ with $\mathbf{j} \in \mathbb{Z}^d$ and* *let A be defined by (35) with the additional restriction that the coefficients $a_{\mathbf{i}}$ have to match the degeneracy of $\lambda_{\mathbf{j}}$. Then*

$$\begin{aligned}&\epsilon_n(A\colon \ell_2 \to \ell_2) \leq \\ &\inf_{(a_{\mathbf{j}})_{\mathbf{j}}:\left(\frac{\sqrt{\lambda_{\mathbf{j}}}}{a_{\mathbf{j}}}\right)_{\mathbf{j}} \in (\ell_2)^d} \sup\nolimits_{t\in\mathbb{N}} 6C_k \left\| \left(\frac{\sqrt{\lambda_{\mathbf{j}}}}{a_{\mathbf{j}}}\right)_{\mathbf{j}} \right\|_{(\ell_2)^d} n^{-\frac{1}{N(r_t,d)}} \left(\prod\nolimits_{q=1}^{j} a(r_q)^{n(r_q,d)}\right)^{\frac{1}{N(r_t,d)}}.\end{aligned} \tag{82}$$

Note that this result, although it may seem straightforward, cannot be obtained from corollary 7 directly as there the sup would have to be carried out over $\mathbb{N}$ instead of $(N(r_t,d))_t$. The different formulation allows us to compute bounds on the entropy numbers more easily.

8.3 Bounds for Kernels in $\mathbb{R}^d$

Let us conclude this section with some examples of the eigenvalue sequences for kernels typically used in SV machines. These can then be used to evaluate the right hand side in corollary 22. Recall that $\nu = \frac{d}{2} - 1$. First we have to compute the Fourier/Hankel transform for the kernels.

Example 23 (Gaussian RBFs) *For Gaussian rbfs in d dimensions we have $k(r) = \sigma^{-d} e^{-\frac{r^2}{2\sigma^2}}$ and correspondingly*

$$\begin{aligned} F[k](\omega) &= \omega^{-\nu}\sigma^{-d} H_\nu \left[r^\nu e^{-\frac{r^2}{2\sigma^2}} \right](\omega) \\ &= \omega^{-\nu}\sigma^{2(\nu+1)-d}\omega^\nu e^{-\frac{\omega^2\sigma^2}{2}} \\ &= e^{-\frac{\omega^2\sigma^2}{2}} \end{aligned}$$

or in other words — the Fourier transform of a Gaussian is a Gaussian.

Example 24 (Exponential RBFs) *In the case of $k(r) = e^{-ar}$ we get*

$$\begin{aligned} F[k](\omega) &= \omega^{-\nu} H_\nu \left[r^\nu e^{-ar} \right](\omega) \\ &= \omega^{-\nu} 2^{\nu+1} \omega^\nu a \pi^{-\frac{1}{2}} \Gamma\left(\nu + \tfrac{3}{2}\right)\left(a^2 + \omega^2\right)^{-\nu - \frac{3}{2}} \\ &= 2^{\frac{d}{2}} a \pi^{-\frac{1}{2}} \Gamma\left(\tfrac{d}{2} + 1\right) \frac{1}{\left(a^2 + \omega^2\right)^{\frac{d+1}{2}}} \end{aligned}$$

i.e. in the case of $d = 1$ we recover the damped harmonic oscillator (in frequency domain). In general we get a decay in terms of the eigenvalues like $\omega^{-(d+1)}$. Moreover we can conclude from this that the Fourier transform of k, viewed itself as a kernel, i.e. $k(r) = \left(1 + r^2\right)^{-\frac{d+1}{2}}$, yields the initial kernel as its corresponding power spectrum in Fourier domain.

Example 25 (Damped Harmonic Oscillator) *Another way to generalize the harmonic oscillator, this time in a way, that k does not depend on the* *dimensionality d is to set $k(r) = \frac{1}{a^2+r^2}$. Following [51] (section 13.6) we get*

$$\begin{aligned} F[k](\omega) &= \omega^{-\nu} H_\nu \left[\frac{r^\nu}{a^2+r^2}\right](\omega) \\ &= \omega^{-\nu} a^\nu K_\nu(\omega a) \end{aligned}$$

where K_ν is the Bessel function of the second kind, defined by (see [44])

$$K_\nu(x) = \int_0^\infty e^{-x\cosh t}\cosh(\nu t)dt. \tag{83}$$

It is possible to upper bound $F[k]$ via

$$K_\nu(x) = \sqrt{\frac{\pi}{2x}}e^{-x}\left[\sum_{j=0}^{p-1}(2x)^{-j}\frac{\Gamma\left(\nu+j+\frac{1}{2}\right)}{j!\Gamma\left(\nu-j+\frac{1}{2}\right)} + \theta\cdot(2x)^{-p}\frac{\Gamma\left(\nu+p+\frac{1}{2}\right)}{j!\Gamma\left(\nu-p+\frac{1}{2}\right)}\right] \tag{84}$$

with $p > \nu - \frac{1}{2}$ and $\theta \in [0,1]$ [16, eq. 8.451.6]). As one can see the term in the brackets $[\cdot]$ converges to 1 for $x \to \infty$ and we get exponential decay of the eigenvalues.

Using Theorem 21, Corollary 22 and Remark 13 one may compute the entropy numbers numerically for a particular kernel and a particular set of parameters. This may seem unsatisfactory from a theoretician's point of view. However, as the ultimate goal is to use the obtained bounds for model selection, it is desirable to obtain as tight bounds (especially in the constants) as possible. Hence if much more precise bounds can be obtained by some not too expensive numerical calculation it is definitely worth while to use those instead of a theoretically nice but not sufficiently tight upper bound. The computational effort to calculate these quantities is typically negligible in comparison to training the actual learning machine.

Notwithstanding the above, in order to give a feeling for the effect of the decay of the Fourier transform of the kernel on the entropy numbers of the A operator, we conclude with the following general result, the proof of which is relegated to the appendix.

Proposition 26 (Polynomial exponential decay in $\mathbb{R}^d$) *For kernels $k(\cdot,\cdot)$ in $\mathbb{R}^d \times \mathbb{R}^d$ with $\lambda(\omega) = \beta^2 e^{-\alpha\|\omega\|^p}$ with $\alpha, \beta, p > 0$ we have*

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

9 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 paper 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.

9.1 A Possible Procedure to use the Results of this Paper

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 6.

Choose the period v of the kernel One suggested procedure is outlined in Section 6.

Bound $\epsilon_n(A)$ This can be done using Corollary 7 (for the case $d = 1$) or Corollary 20 or 22 for the case $d > 1$. Some examples of this sort of calculation are given in Section 7.

Bound $\epsilon_n(T)$ Using Theorem 10.

Take account of the "$+b$" The key observation is that given a class $\mathcal{F}$ with known $\mathcal{N}^m(\epsilon, \mathcal{F})$, one can bound $\mathcal{N}^m(\epsilon, \mathcal{F}^+)$ as follows. (Here $\mathcal{F}^+ := \{f + b\colon f \in \mathcal{F}, b \in \mathbb{R}\}$.) 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

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

is an ϵ-cover for $\mathcal{F}^+$ and thus $\mathcal{N}^m(\epsilon, \mathcal{F}^+) \leq \frac{2B}{\epsilon}\mathcal{N}^m(\epsilon, \mathcal{F})$. 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 2 for example.

Plug into a uniform convergence result See the pointers to the literature and the example in Section 4.

9.2 Future Work

One might think of similar algorithms (e.g. weight decay) which place a similar constraint not on the weight vector in feature space but, say, in input space. It seems promising to explore this direction in more detail within the framework presented here.

The results of the present paper hinge on the measurement of the size of the weight vector **w** by a ℓ_2 norm. In [53] we show the effect of different norms for measuring the size of **w**, as well as presenting a number of related results.

Acknowledgements

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

A Proofs of Results in Section 7

Proof (Proposition 14) We will make use of Corollary 7. In this case all sequences $(a_j)_j = (j^{-\frac{\tau}{2}})_j$ with $0 < \tau < \alpha$ lead to an admissible scaling property. Here we have

$$\left\| \left(\frac{\sqrt{\lambda_j}}{a_j} \right)_j \right\|_{\ell_2} = \beta \left\| \left(j^{\frac{\tau - \alpha - 2}{2}} \right)_j \right\|_{\ell_2} = \beta \zeta(\alpha - \tau + 1) \tag{85}$$

where $\zeta(\cdot)$ is Riemann's zeta function. Moreover we can upper bound $\zeta(\cdot)$ by $\zeta(1 + \frac{1}{x}) \leq x + \gamma$ where γ is Euler's constant. The next step is to evaluate the expression

$$(a_1 a_2 \cdots a_j)^{\frac{1}{j}} = \left(\prod_{s=1}^{j} s^{-\frac{\tau}{2}} \right)^{\frac{1}{j}} = (j!)^{-\frac{\tau}{2j}} \tag{86}$$

Hence we get

$$\epsilon_n \leq 6 C_k \beta \inf_{\tau \in (0,\alpha)} \sup_{j \in \mathbb{N}} \left(\frac{1}{\alpha - \tau} + \gamma \right) n^{-\frac{1}{j}} (j!)^{-\frac{\tau}{2j}} \tag{87}$$

Replacing $\sup_{j \in \mathbb{N}}$ in (87) by $\sup_{j \in [1,\infty)}$ and rewriting the expressions as exponentials we obtain

$$\epsilon_n \leq 6 C_k \beta \inf_{\tau \in (0,\alpha)} \sup_{j \in [1,\infty)} \left(\frac{1}{\alpha - \tau} + \gamma \right) e^{-\frac{1}{j} \ln n - \frac{\tau}{2j} \ln \Gamma(j+1)} \tag{88}$$

One can check (by making use of an asymptotic expansion of $\Gamma(x)$) that

$$\frac{\tau}{2j} \ln \Gamma(j+1) \geq \frac{\tau}{2} (\ln j - 1) - \frac{\tau \ln 2\pi}{4j}. \tag{89}$$

Hence we can upper–bound the exponent (as we have to compute the sup) in (88) by

$$-\frac{1}{j} \ln n - \frac{1}{2} \tau (\ln j - 1) - \frac{\tau (\ln 2\pi)}{4j}. \tag{90}$$

The maximum is obtained for $j = \frac{2 \ln n}{\tau} + \frac{\ln(2\pi)}{2}$ and hence

$$\epsilon_n \leq 6 C_k \beta \inf_{\tau \in (0,\alpha)} \left(\frac{1}{\alpha - \tau} + \gamma \right) \left(\frac{2 \ln n}{\tau} + \frac{1}{2} \ln(2\pi) \right)^{-\frac{\tau}{2}}. \tag{91}$$

The inf is approximately obtained for

$$\tau = \alpha - \frac{4}{\gamma \ln^2 Z} \text{ with } Z = \frac{2 \ln n}{\alpha} + \frac{1}{2} \ln 2\pi$$

and consequently $\tau \to \alpha$ for $n \to \infty$. Substitution concludes the proof. ■

Proof (Proposition 15) By using the same argument as in the example above, for sequences of exponential decay, i.e. $(a_j)_j = (e^{-\frac{\tau}{2}(j-1)})_j$ with $\tau \in (0, \alpha)$ we get

$$\left\| \left(\frac{\sqrt{\lambda_j}}{a_j} \right)_j \right\|_{\ell_2} = \beta \left\| \left(e^{\frac{\tau-\alpha}{2}}(j-1) \right)_j \right\|_{\ell_2} = \beta \frac{1}{1 - e^{\tau-\alpha}} \tag{92}$$

and

$$(a_1 a_2 \dots a_j)^{\frac{1}{j}} = e^{-\frac{1}{2j}\tau \sum\limits_{s=1}^{j}(s-1)} = e^{-\frac{j-1}{2}\tau} \tag{93}$$

Hence we get

$$\epsilon_n \leq 6 C_k \beta \inf_{\tau \in (0,\alpha)} \sup_{j \in \mathbb{N}} n^{-\frac{1}{j}} \frac{1}{1 - e^{\tau-\alpha}} e^{-\frac{j-1}{2}\tau} \tag{94}$$

In order to compute the inf sup we will make some approximations. The first step to replace $\sup\limits_{j \in \mathbb{N}}$ by $\sup\limits_{j \in \mathbb{R}^+}$. As we only want to compute an upper bound this is a useful assumption. The maximum is achieved at $j = \sqrt{\frac{2 \ln n}{\tau}}$. Plugging j back in and rearranging terms yields

$$\epsilon_n \leq 6 C_k \beta \inf_{\tau \in (0,\alpha)} \frac{e^{-\sqrt{2 \ln n \tau} + \frac{\tau}{2}}}{1 - e^{\tau-\alpha}} \tag{95}$$

This concludes the proof as the proposition holds for any $\tau \in (0, \alpha)$. ∎

Remark 27 (Computing $\inf\limits_{\tau \in (0,\alpha)}$) *Basically there are two options — either to numerically compute* inf *of (95), or to use approximations. For small* $\ln n$*, i.e.* $\ln n < 2\alpha$ *setting* $\tau = \alpha/2$ *is a good estimate. For* $\ln n \gg 2\alpha$ *we can find the approximate minimum of (95) for*

$$e^{\tau-\alpha} = \frac{\sqrt{2 \ln n} - \sqrt{\alpha}}{\sqrt{2 \ln n} + \sqrt{\alpha}} \approx 1 - \sqrt{\frac{2\alpha}{\ln n}} \tag{96}$$

This will determine our particular choice of τ *(although it may not be optimal). From this it follows that* $\tau \geq \alpha - \sqrt{\frac{2\alpha}{\ln n}}$*, i.e.* $\tau \to \alpha$ *for* $n \to \infty$*. This leads to*

$$\epsilon_n \leq 6 C \beta \sqrt{\frac{\ln n}{2\alpha}} e^{\frac{\alpha}{2} - \sqrt{2\alpha \ln n - \sqrt{8\alpha \ln n}}} \tag{97}$$

This bound has the same rate but possibly better constant factors.

For the proof of Proposition 16 we need the following Lemma:

Lemma 28 (Summation and Integration in $\mathbb{R}^1$) *Suppose* $f : \mathbb{R} \to \mathbb{R}$ *is a nonincreasing function. Then the following inequality holds for any* $a \in \mathbb{Z}$

$$\int_a^\infty f(x)dx \leq \sum_{n=a}^{\infty} f(n) \leq \int_{a-1}^\infty f(x)dx. \tag{98}$$

Proof The proof relies on the fact that

$$f(n) \geq \int_{n}^{n+1} f(n)dn \geq f(n+1)$$

due to the monotonicity of f and a decomposition of the integral $\int\limits_0^\infty = \sum\limits_{n=0}^{\infty} \int\limits_n^{n+1}$. The lemma is a direct consequence thereof. ∎

Proof (Proposition 16) Choose $(a_j)_j = (e^{-\frac{\tau}{2}(j-1)^2})_j$ with $\tau \in (0,\alpha)$ analogously to the two cases above. This leads to

$$\left\| \left(\frac{\sqrt{\lambda_j}}{a_j} \right)_j \right\|_{\ell_2} = \beta \left(\sum_{j=0}^{\infty} e^{(\tau-\alpha)j^2} \right)^{\frac{1}{2}} \leq \beta \sqrt{1 + \sqrt{\frac{\pi}{2(\alpha-\tau)}}} \tag{99}$$

due to Lemma 28 and

$$(a_1 a_2 \dots a_j)^{\frac{1}{j}} = e^{-\frac{1}{2j}\tau \sum\limits_{s=1}^{j}(s-1)^2} = e^{-\frac{\tau}{12}(j-1)(2j-1)}. \tag{100}$$

Putting everything together yields

$$\epsilon_n \leq 6C\beta \inf_{\tau\in(0,\alpha)} \sup_{j\in\mathbb{N}} n^{-\frac{1}{j}} \sqrt{1 + \sqrt{\frac{\pi}{2(\alpha-\tau)}}} e^{-\frac{\tau}{12}(j-1)(2j-1)}. \tag{101}$$

Now we will upper bound (101). For evaluating the sup let us distinguish two cases.

For the sup being obtained at $j = 1$ we get

$$\epsilon_n \leq 6C\beta n^{-1} \left(1 + \sqrt{\frac{\pi}{2\alpha}} \right), \tag{102}$$

thus $\ln \epsilon^{-1} = O(\ln n)$ for $j = 1$. For $j \geq 2$ the following expression is an upper bound for ϵ_n

$$\epsilon_n \leq 6C\beta \inf_{\tau\in(0,\alpha)} \sup_{j\in[2,\infty)} n^{-\frac{1}{j}} \sqrt{1 + \sqrt{\frac{\pi}{2(\alpha-\tau)}}} e^{-\frac{\phi\tau}{6}j^2}. \tag{103}$$

with $\phi = 3/8$ as $\frac{1}{12}(j-1)(2j-1) \geq \frac{1}{16}j^2$ for all $j \in [2,\infty)$. Computing the $\sup_{j\in[2,\infty]}$ leads to

$$j = \left(\frac{3 \ln n}{\phi\tau} \right)^{\frac{1}{3}} \tag{104}$$

and therefore

$$\epsilon_n \leq 6C\beta \inf_{\tau\in(0,\alpha)} \sqrt{1 + \sqrt{\frac{\pi}{2(\alpha-\tau)}}} e^{-\frac{1}{2}(9\phi\tau)^{\frac{1}{3}} \ln^{\frac{2}{3}} n} \tag{105}$$

The overall bound on ϵ_n is bounded by the maximum of the bounds obtained for $j = 1$ and $j \geq 2$. As $\ln n^{\frac{2}{3}}$ decays more slowly than $\ln n$ the case of $j \geq 2$ dominates for large n and thus

$$\ln \epsilon_n^{-1} = O(\ln^{2/3} n) \tag{106}$$

This proves the scaling behaviour for quadratic polynomial decay as the given rate holds for any particular τ ∎

Note that for sufficiently large $\ln n$ we can let ϕ get arbitrarily close to 1 and thus obtain a better rate.

Remark 29 (Computing $\inf\limits_{\tau \in (0,\alpha)}$) *Again we have the choice of either numerically evaluating (105) or computing an approximate solution by either setting $\tau = \alpha/2$ or making the assumption of large* $\ln n$. *In the latter case we obtain an approximately optimal solution for*

$$\tau = \alpha - \frac{1}{2}\left(\frac{9\pi^2\alpha^4}{\phi^2 \ln^4 n}\right)^{1/9}$$

which converges to α for $\ln n \to \infty$.

Proof (Proposition 17)(sketch only) Analogously to before we use a series $(a_j)_j = e^{-\tau/2j^p}$. Then we bound

$$\left\| \left(\frac{\sqrt{\lambda_j}}{a_j}\right)_j \right\|_{\ell_2} = \beta \left(\sum_{j=0}^{\infty} e^{(\tau-\alpha)j^p}\right)^{\frac{1}{2}} \leq \beta\sqrt{1 + \frac{\Gamma(1/p)}{p(\alpha-\tau)^{1/p}}} \tag{107}$$

and

$$(a_1 a_2 \dots a_j)^{\frac{1}{j}} = e^{-\frac{1}{2j}\tau \sum\limits_{s=1}^{j} s^p} \leq e^{-\tau\phi j^p} \tag{108}$$

for some positive number ϕ. And again computing the $\sup_{k \in [1,\infty]}$ we get $k = \phi' \ln^{\frac{1}{p+1}} n$ for some ϕ', and then finally resubstitution yields the claimed rate of convergence for any $\tau \in (0, \alpha)$ which proves the theorem. ∎

B Proof of Theorem 19

Proof The first part of the inequality follows directly from theorem 6 as it is a weaker statement than the original one. We prove the second part by closely mimicking the proof in [11, p. 17]. We define

$$\delta(n) := 8 \sup_{t \in \mathbb{N}} n^{-\frac{1}{s_t}} (\sigma_1 \sigma_2 \cdots \sigma_{s_t})^{\frac{1}{s_t}} \tag{109}$$

and show that for all n there is an index s_j with $\sigma_{s_j+1} \leq \frac{\delta(n)}{4}$. For this purpose choose an index r such that $n \leq 2^{s_j+1}$ and thus $1 \leq 2n^{-1/(s_j+1)}$. Moreover we have

$$\sigma_{s_j+1} \leq (\sigma_1 \sigma_2 \cdots \sigma_{s_j+1})^{\frac{1}{s_j+1}} \tag{110}$$

because of the monotonicity of $(\sigma_j)_j$ and finally

$$\sigma_{s_j+1} \leq 2n^{-1/(s_j+1)}(\sigma_1\sigma_2\cdots\sigma_{s_j+1})^{\frac{1}{s_j+1}}. \tag{111}$$

Using the definition of $\delta(n)$ we thus conclude $\sigma_{s_j+1} \leq \delta(n)/4$. If this happens to be the case for σ_1 we have $\epsilon_n(D) \leq \sigma_1$ which proves the theorem.

If this is not the case there exists an index s_j such that $\sigma_{s_j+1} \leq \delta(n)/4 < \sigma_{s_j}$. Hence the corresponding sectional operator

$$\begin{aligned}&D_{s_j} : \ell_p \to \ell_p \text{ with}\\ &D_{s_j}(x_1, x_2, \ldots, x_{s_j}, x_{s_j+1}, \ldots) = (\sigma_1 x_1, \sigma_2 x_2, \ldots, \sigma_{s_j} x_{s_j}, 0, 0, \ldots)\end{aligned} \tag{112}$$

is of rank s_j and the image $D_{s_j}(U_p)$ of the closed unit ball U_p of ℓ_p is isometric to the subset $D^{(s_j)}(U_p^{(s_j)})$ of $\ell_p^{s_j}$. In any case $D_{s_j}(U_p)$ is a precompact subset of ℓ_p. So let $y_1, y_2, \ldots, y_N$ be a maximal system of elements in $D_{s_j}(U_p)$ with

$$\|y_j - y_{\bar{\jmath}}\| > \delta(n)/2 \text{ for } j \neq \bar{\jmath}. \tag{113}$$

The maximality of this system guarantees that

$$D_{s_j}(U_p) \subseteq \bigcup_{j=1}^{N} \left\{ y_j + \frac{\delta(n)}{2} U_p \right\} \tag{114}$$

and thus $\epsilon_N(D_{s_j}) \leq \delta(n)/2$. In order to get an estimate for $\epsilon_N(D)$ we split the operator D into two parts $D = (D - D_{s_j}) + D_{s_j}$ which allows us to bound

$$\epsilon_N(D) \leq \|D - D_{s_j}\| + \epsilon_N(D_{s_j}). \tag{115}$$

Using $\|D - D_{s_j}\| = \sigma_{s_j+1} \leq \delta(n)/4$ and the bound on $\epsilon_N(D_{s_j})$ we arrive at

$$\epsilon_N(D) \leq \frac{3}{4}\delta(n). \tag{116}$$

The final step is to show that $N \leq n$ as then by substituting in the definition of $\delta(n)$ into (116) yields the result. This is again achieved by a comparison of volumes. Consider the sets $\{y_j + (\delta(n)/4)U_p^{s_j}\}$ as subsets of the space $\ell_p^{s_j}$ which is possible since $y_j \in D_{s_j}(U_p)$ and $D_{s_j}(U_p) = D^{(s_j)}(U_p^{s_j})$. These sets are obviously pairwise disjoint. On the other hand we have

$$\bigcup_{j=1}^{N} \left\{ y_j + \frac{\delta(n)}{4} U_p^{s_j} \right\} \subseteq D^{(s_j)}(U_p^{s_j}) + \tfrac{\delta(n)}{4} U_p^{s_j}) \subseteq 2D^{(s_j)}(U_p^{s_j}) \tag{117}$$

as $\delta(n)/4 < \sigma_1$. Now a comparison of the d-dimensional Euclidean volumes vol_d provides

$$N\left(\frac{\delta(n)}{4}\right)^{s_j} \mathrm{vol}_{s_j}(U_p^{s_j}) \leq 2^{s_j}\sigma_1\sigma_2\cdots\sigma_{s_j}\mathrm{vol}_{s_j}(U_p^{s_j}) \tag{118}$$

and therefore $N \leq (8/\delta(n))^{s_j}\sigma_1\sigma_2\cdots\sigma_{s_j}$. Using the definition of $\delta(n)$ this yields $N \leq n$. ∎

C Proof of Proposition 26

Proof We will completely ignore the fact that we are actually dealing with a countable set of eigenvalues on a lattice and replace all summations by integrals without further worry. Of course this is not accurate but still will give us the correct rates for the entropy numbers.

Denote $1/\Lambda := (2\pi/v)^{\frac{d}{2}}$ the size of a unit cell, i.e. $\Lambda = (v/(2\pi))^{\frac{d}{2}}$ the density of lattice points in frequency space as given in section 6. Then we get for infinitesimal volumes dV and numbers of points dN in frequency space

$$dV = S_{d-1} r^{d-1} dr \text{ and therefore } dN = \Lambda S_{d-1} r^{d-1} dr \tag{119}$$

(here S_{d-1} denotes the volume of the $d-1$ dimensional unit sphere) leading to

$$N(r) = \frac{1}{d} \Lambda S_{d-1} r^d. \tag{120}$$

We introduce a scaling operator whose eigenvalues decay like $a(\omega) = e^{-\frac{\tau}{2}\|\omega\|^p}$ for $\tau \in [0, \alpha)$. It is straightforward to check that all these values lead to both useful and admissible scaling operators. Now we will estimate the separate terms in (82).

$$\begin{aligned} \left\| \left(\frac{\sqrt{\lambda_{\mathbf{i}}}}{a_{\mathbf{i}}} \right)_{\mathbf{i}} \right\|_{\ell_2}^2 &\,\tilde{=}\, \int dN(\omega) \frac{\lambda(\omega)}{a^2(\omega)} = S_{d-1} \Lambda \int_0^\infty r^{d-1} \beta^2 e^{-(\alpha-\tau)\|\omega\|^p} \\ &= S_{d-1} \Lambda \beta^2 (\alpha - \tau)^{-\frac{d}{p}} \tau \left(\frac{d}{p} \right) p^{-1} \end{aligned} \tag{121}$$

Next we have

$$\ln \left(n^{-\frac{1}{N(r)}} \right) = -\frac{d}{\Lambda S_{d-1} r_0^d} \ln n \tag{122}$$

and

$$\begin{aligned} \ln \left(a_1 \cdot a_2 \cdots a_{N(r)} \right)^{\frac{1}{N(r)}} &= -\frac{d}{\Lambda S_{d-1} r_0^d} \sum_{j=1}^{N(r)} \ln a_j \\ &\approx d r^{-d} \int_0^r \omega^{d-1} \ln a(\omega) d\omega \qquad (123) \\ &= -d r^{-d} \int_0^r \omega^{d-1} \frac{\tau}{2} \omega^p d\omega = -\frac{\tau}{2} \frac{d}{d+p} r^p. \qquad (124) \end{aligned}$$

This leads to

$$\epsilon_n \leq 6 C_k \beta \sqrt{\frac{S_{d-1} \Lambda \Gamma\left(\frac{d}{p}\right)}{p}} \inf_{\tau \in [0,\alpha)} (\alpha - \tau)^{-\frac{d}{2p}} \sup_{r \in \mathbb{R}^+} \exp \left(-\frac{d}{\Lambda S_{d-1} r^d} \ln n - \frac{\tau}{2} \tfrac{d}{d+p} r^p \right). \tag{125}$$

Computing the $\sup_{r \in \mathbb{R}^+}$ yields

$$r = \left(\frac{2}{\tau \Lambda S_{d-1}} \frac{(d+p)d}{p} \ln n \right)^{\frac{1}{d+p}} \tag{126}$$

and therefore

$$\epsilon_n \leq 6C_k\beta\sqrt{\frac{S_{d-1}\Lambda\Gamma\left(\frac{d}{p}\right)}{p}} \inf_{\tau\in[0,\alpha)} (\alpha-\tau)^{-\frac{d}{2p}} \exp\left(-\left(\frac{\tau}{2}\right)^{\frac{d}{d+p}}\left(\frac{(d+p)d}{p}\frac{\ln n}{\Lambda S_{d-1}}\right)^{\frac{p}{d+p}}\right). \tag{127}$$

Already from this expression one can observe the rate bounds on ϵ_n. What remains to be done is to compute the $\inf_\tau$. This can be done by differentiating (127) w.r.t. τ. For increased clarity of exposition define

$$T_n := \left(\frac{(d+p)d}{p}\frac{\ln n}{\Lambda S_{d-1}}\right)^{\frac{p}{d+p}} \tag{128}$$

which leads to the optimality condition on τ

$$(\alpha-\tau)\tau^{-\frac{p}{d+p}} = \frac{d+p}{2T_n p}2^{\frac{d}{d+p}} \text{ with } \tau \in (0,\alpha] \tag{129}$$

which can be solved numerically.

∎

D Maple code to compute n(r,d)

```
# This code defines a function t where
# t(m,d) is number of points on a sphere of radius^2=m from Z^d
h:=n->eval(`if`(isolve(m^2=n,m)=NULL,0,`if`(n=0,1,2)),1):
powseries[powcreate](theta(n)=h(n)):
t:=(m,d)->
   coeff(convert(powseries[tpsform](powseries[evalpow](theta^d),
        x,m+1),polynom),x,m):
```

References

[1] M. A. Aizerman, E. M. Braverman, and L. I. Rozonoér. Theoretical foundations of the potential function method in pattern recognition learning. *Automation and Remote Control*, 25:821–837, 1964.

[2] S. Akashi. Characterization of ε-entropy in Gaussian processes. *Kodai Mathematical Journal*, 9:58–67, 1986.

[3] S. Akashi. The asymptotic behaviour of ε-entropy of a compact positive operator. *Journal of Mathematical Analysis and Applications*, 153:250–257, 1990.

[4] N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale–sensitive Dimensions, Uniform Convergence, and Learnability. *J. of the ACM*, 44(4):615–631, 1997.

[5] M. Anthony. Probabilistic analysis of learning in artificial neural networks: The PAC model and its variants. *Neural Computing Surveys*, 1:1–47, 1997. http://www.icsi.berkeley.edu/˜jagota/NCS.

[6] R. Ash. *Information Theory.* Interscience Publishers, New York, 1965.

[7] P. Bartlett, P. Long, and R. Williamson. Fat–Shattering and the Learnability of Real–Valued Functions. *Journal of Computer and System Sciences*, 52(3):434–452, 1996.

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

[9] B. Carl. Entropy numbers of diagonal operators with an application to eigenvalue problems. *Journal of Approximation Theory*, 32:135–150, 1981.

[10] B. Carl. Inequalities of Bernstein-Jackson-type and the degree of compactness of operators in Banach spaces. *Annales de l'Institut Fourier*, 35(3):79–118, 1985.

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

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

[13] M. Defant and M. Junge. Characterization of weak type by the entropy distribution of r-nuclear operators. *Studia Mathematica*, 107(1):1–14, 1993.

[14] F. Girosi, M. Jones, and T. Poggio. Priors, stabilizers and basis functions: From regularization to radial, tensor and additive splines. A.I. Memo No. 1430, MIT, 1993.

[15] Y. Gordon, H. König, and C. Schütt. Geometric and probabilistic estimates for entropy and approximation numbers of operators. *Journal of Approximation Theory*, 49:219–239, 1987.

[16] I.S. Gradshteyn and I.M. Ryzhik. *Table of integrals, series, and products.* Academic Press, New York, 1981.

[17] E. Grosswald. *Representation of Integers as Sums of Squares.* Springer, N. Y., 1985.

[18] L. Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. In M. Li and A. Maruoka, editors, *Algorithmic Learning Theory ALT-97*, LNAI-1316, pages 352–363, Berlin, 1997. Springer.

[19] D. Jagerman. ε-entropy and approximation of bandlimited functions. *SIAM Journal on Applied Mathematics*, 17(2):362–377, 1969.

[20] M. Junge and M. Defant. Some estimates of entropy numbers. *Israel Journal of Mathematics*, 84:417–433, 1993.

[21] M. J. Kearns, R. E. Schapire, and L. M. Sellie. Toward efficient agnostic learning. *Machine Learning*, 17(2):115–141, 1994.

[22] V.I. Kolchinskiĭ. Operators of type p and metric entropy. *Teoriya Veroyatnosteĭ Matematicheskaya Statistika*, 38:69–76, 135, 1988. (In Russian. MR 89j:60007).

[23] V.I. Kolchinskiĭ. Entropic order of operators in banach spaces and the central limit theorem. *Theory of Probability and its Applications*, 36(2):303–315, 1991.

[24] A.N. Kolmogorov and V.M. Tihomirov. ε-entropy and ε-capacity of sets in functional spaces. *American Mathematical Society Translations, Series 2*, 17:277–364, 1961.

[25] H. König. *Eigenvalue Distribution of Compact Operators.* Birkhäuser, Basel, 1986.

[26] T. Koski, L.-E. Persson, and J. Peetre. ε-entropy ε-rate, and interpolation spaces revisited with an application to linear communication channels. *Journal of Mathematical Analysis and Applications*, 186:265–276, 1994.

[27] W.S. Lee, P.L. Bartlett, and R.C. Williamson. The importance of convexity in learning with squared loss. *IEEE Transactions on Information Theory*, 1998. to appear.

[28] C.A. Micchelli. Interpolation of scattered data: distance matrices and conditionally positive definite functions. *Constructive Approximation*, 2:11–22, 1986.

[29] C. Müller. *Analysis of Spherical Symmetries in Euclidean Spaces*, volume 129 of *Applied Mathematical Sciences.* Springer, New York, 1997.

[30] N.J. Nilsson. *Learning machines: Foundations of Trainable Pattern Classifying Systems.* McGraw–Hill, 1965.

[31] A. Pajor. *Sous-espaces ℓ_n^1 des espaces de Banach.* Hermann, Paris, 1985.

[32] A. Pietsch. *Operator ideals.* North-Holland, Amsterdam, 1980.

[33] L.S. Pontriagin and L.G. Schnirelmann. Sur une propriété métrique de la dimension. *Annals of Mathematics*, 33:156–162, 1932.

[34] R.T. Prosser. The ε–Entropy and ε–Capacity of Certain Time–Varying Channels. *Journal of Mathematical Analysis and Applications*, 16:553–573, 1966.

[35] R.T. Prosser and W.L. Root. The ε-entropy and ε-capacity of certain time-invariant channels. *Journal of Mathematical Analysis and its Applications*, 21:233–241, 1968.

[36] S. Saitoh. *Theory of Reproducing Kernels and its Applications.* Longman Scientific & Technical, Harlow, England, 1988.

[37] 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.

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

[39] C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[40] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 1998. To appear. Also: NeuroCOLT TR-96-053, 1996, ftp.dcs.rhbnc.ac.uk/pub/neurocolt/tech_reports.

[41] J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. A framework for structural risk minimisation. In *Proceedings of the 9th Annual Conference on Computational Learning Theory*, pages 68–76. ACM Press, 1996.

[42] A. J. Smola and B. Schölkopf. On a kernel–based method for pattern recognition, regression, approximation and operator inversion. *Algorithmica*, 22:211–231, 1998.

[43] A.J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 11:637–649, 1998.

[44] I. H. Sneddon. *The Use of Integral Transforms.* McGraw–Hill, New York, 1972.

[45] M. Talagrand. The Glivenko–Cantelli problem, ten years later. *Journal of Theoretical Probability*, 9(2):371–384, 1996.

[46] H. Triebel. Interpolationseigenschaften von Entropie- und Durchmesseridealen kompakter Operatoren. *Studia Mathematica*, 34:89–107, 1970.

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

[48] V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition [in Russian].* Nauka, Moscow, 1974. (German Translation: W. Wapnik & A. Tscherwonenkis, *Theorie der Zeichenerkennung*, Akademie–Verlag, Berlin, 1979).

[49] V.N. Vapnik. *Estimation of Dependences from Empirical Data.* Springer, New York, 1982.

[50] V.N. Vapnik and A.Ya. Chervonenkis. Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and its Applications*, 26(3):532–553, 1981.

[51] G.N. Watson. *A Treatise on the Theory of Bessel Functions.* Cambridge University Press, Cambridge, UK, 2 edition, 1958.

[52] H. Widom. Asymptotic behaviour of eigenvalues of certain integral operators. *Archive for Rational Mechanics and Analysis*, 17:215–229, 1964.

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