Entropy Numbers, Operators and Support Vector Kernels[⋆]

Robert C. Williamson[1], Alex J. Smola[2], and Bernhard Schölkopf[2]

[1] Department of Engineering,
Australian National University,
Canberra, ACT 0200, Australia
Bob.Williamson@anu.edu.au
[2] GMD FIRST, Rudower Chaussee 5,
12489 Berlin, Germany
{smola,bs}@first.gmd.de

Abstract. We derive new bounds for the generalization error of feature space machines, such as support vector machines and related regularization networks by obtaining new bounds on their covering numbers. The proofs are based on 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 functions on the generalization performance of support vector machines.

1 Introduction, Definitions and Notation

In this paper we give new bounds on the covering numbers for feature space machines. This leads to improved bounds on their generalization performance. Feature space machines perform a mapping from input space into a feature space 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 [32] are a well known example of this class.
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

[⋆] Supported by the Australian Research Council and the DFG (# Ja 379/71).

P. Fischer and H.U. Simon (Eds.): EuroCOLT'99, LNAI 1572, pp. 285–298, 1999.

bound covering numbers by viewing the relevant class of functions as the image of a unit ball under a particular compact operator. The results can be applied to bound the generalization performance of SV regression machines, although we do not explictly indicate the results so obtained in this brief paper.

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 [28]. It can be shown that for separable training data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \in \mathbb{R}^d \times \{\pm 1\}$, 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 $f(\mathbf{x}) = \operatorname{sgn}(\langle \mathbf{w}, \mathbf{x} \rangle + b)$. Similarly, a linear regression $f(\mathbf{x}) = \langle \mathbf{w}, \mathbf{x} \rangle + b$ can be estimated from data $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \in \mathbb{R}^d \times \mathbb{R}$ 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.

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 [1,7]. Under certain conditions on a kernel k, to be stated below (Theorem 1), there exists a nonlinear map Φ into a reproducing kernel Hilbert space F such that k computes the dot product in F, i.e. $k(\mathbf{x}, \mathbf{y}) = \langle \Phi(\mathbf{x}), \Phi(\mathbf{y}) \rangle_F$. 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.

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 $f(\mathbf{x}) = \sum_{j=1}^{m} \alpha_j k(\mathbf{x}_j, \mathbf{x}) + b$, with $\alpha_j \in \mathbb{R}$, $j = 1, \ldots, m$. It has been noticed that different kernels can be characterized by their regularization properties [30]. This provides insight into the regularization properties of SV kernels. However, it does not give us a comprehensive understanding 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.

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 $p = \infty$, $\|\mathbf{x}\|_{\ell_\infty^d} := \|\mathbf{x}\|_\infty = \max_{j=1,\ldots,d} |x_j|$. Analogously ℓ_p is the space of infinite sequences with the obvious definition of the norm. Given m points $\mathbf{x}_1, \ldots, \mathbf{x}_m \in \ell_p^d$, we use the shorthand $\mathbf{X}^m = (\mathbf{x}_1^T, \ldots, \mathbf{x}_m^T)$. 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 $f \colon \mathcal{X} \to \mathbb{R}$

we define $\|f\|_{L_p} := \left(\int |f(x)|^p d\mu(x)\right)^{1/p}$ if the integral exists and $\|f\|_{L_\infty} := \operatorname{ess\,sup}_{x\in\mathcal{X}} |f(x)|$. For $1 \leq p \leq \infty$, we let $L_p(\mathcal{X}) := \{f\colon \mathcal{X} \to \mathbb{R}\colon \|f\|_{L_p} < \infty\}$. 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 $U_E := \{x \in E\colon \|x\|_E \leq 1\}$ is bounded. The smallest such bound is called the *operator norm*, $\|T\| := \sup_{x\in U_E} \|Tx\|_F$. The nth *entropy number of a set* $M \subset E$, for $n \in \mathbb{N}$, is

$$\epsilon_n(M) := \inf\{\epsilon > 0\colon \exists \text{ an } \epsilon\text{-cover for } M \text{ in } E \text{ containing } n \text{ or fewer points}\}.$$

The *entropy numbers of an operator* $T \in \mathfrak{L}(E,F)$ are defined as $\epsilon_n(T) := \epsilon_n(T(U_E))$. 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 $\overline{T(U_E)}$ is compact. The *dyadic entropy numbers of an operator* are defined by $e_n(T) := \epsilon_{2^{n-1}}(T)$, $n \in \mathbb{N}$. A very nice introduction to entropy numbers of operators is [8]. 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. 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 $(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_m, y_m) \in \mathbb{R}^d \times \mathbb{R}$, respectively. We will map the input data into a feature space via a mapping Φ. We let $\tilde{\mathbf{x}} := \Phi(\mathbf{x})$.

2 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

$$\mathcal{N}^m(\epsilon, \mathcal{F}) := \sup_{\mathbf{x}_1,\ldots,\mathbf{x}_m \in \mathcal{X}} \mathcal{N}(\epsilon, \mathcal{F}, \ell_\infty^{\mathbf{X}^m}), \tag{1}$$

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].

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 [36]. 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 [36].

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 [24] and was studied in detail by Kolmogorov and others [19]. The use of metric entropy to say something about linear operators was developed independently by several people. Prosser [25] 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 [27], where he considered the effect of different convolution operators on the entropy of an ensemble. Prosser's paper [25] led to a handful of papers (see e.g. [26,15,3,21]) 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 [8] studied covering numbers [31] and later entropy numbers [23] 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. [22]). More recently Gurvits [14] 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 [36]; 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 [17,18] and Defant and Junge [12,16]. Note that the exact formulation of their results differs. Kolchinskiĭ was motivated by probabilistic problems not unlike ours.

3 Generalization Bounds via Uniform Convergence

The generalization performance of learning machines can be bounded via uniform convergence results as in [34]. 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 $P_m(f) := \frac{1}{m}\sum_{j=1}^{m} f(\mathbf{x}_j)$ 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 $m \geq \frac{2}{\epsilon^2}$,*

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

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 $\tilde{\mathbf{X}}^m = (\tilde{\mathbf{x}}_1^T, \ldots, \tilde{\mathbf{x}}_{2m}^T)$ also drawn i.i.d. from P.*

In order to use this lemma one can make use of the fact that $\mathbf{E}\left[\mathcal{N}(\epsilon, \mathcal{F}, \ell_\infty^{\tilde{\mathbf{X}}^m})\right] \leq \mathcal{N}^m(\epsilon, \mathcal{F})$. The above result can be used to give a generalization error result by applying it to the loss-function induced class using standard techniques. Furthermore, one can obtain bounds on the generalization error of classifiers in terms of the margin achieved on a training sample in terms of these covering numbers — see [28].

4 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 4.3 we will show how a more direct approach could be taken towards upper–bounding entropy numbers.

4.1 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 [20, 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 1 (Mercer). *Suppose $k \in L_\infty(\mathcal{X}^2)$ such that the integral operator $T_k : L_2(\mathcal{X}) \to L_2(\mathcal{X})$,*

$$T_k f(\cdot) := \int_{\mathcal{X}} k(\cdot, \mathbf{y}) f(\mathbf{y}) d\mu(\mathbf{y}) \tag{3}$$

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 $\|\psi_j\|_{L_2} = 1$ and let $\overline{\psi_j}$ denote its complex conjugate. Then

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. *$k(\mathbf{x}, \mathbf{y}) = \sum_{j \in \mathbb{N}} \lambda_j \overline{\psi_j(\mathbf{x})} \psi_j(\mathbf{y})$ 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

$$|\psi_j(\mathbf{x})| \leq C_k \text{ for all } j \in \mathbb{N} \text{ and } \mathbf{x} \in \mathcal{X}. \tag{4}$$

(Actually (4) 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

$$\begin{aligned} \Phi &: \mathcal{X} \to \ell_2 \\ \mathbf{x} &\mapsto (\phi_j(\mathbf{x}))_j := (\sqrt{\lambda_j} \psi_j(\mathbf{x}))_j \end{aligned} \tag{5}$$

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 1 (Mapping $\Phi(\mathbf{x})$ into ℓ_2). *Let S be the diagonal map*

$$\begin{aligned} S &: \mathbb{R}^{\mathbb{N}} \to \mathbb{R}^{\mathbb{N}} \\ S &: (x_j)_j \mapsto S(x_j)_j = (s_j x_j)_j. \end{aligned} \tag{6}$$

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

$$\|S\Phi(\mathbf{x})\|_{\ell_2}^2 = \sum_{j\in\mathbb{N}} s_j^2\lambda_j|\psi_j(\mathbf{x})|^2 \leq \sum_{j\in\mathbb{N}} s_j^2\lambda_j C_k^2 = R_S^2. \tag{7}$$

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

$$a_n(\mathbf{x}) := \sum_{j=1}^{n} s_j^2\lambda_j|\psi_j(\mathbf{x})|^2. \tag{8}$$

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

$$a_n(\mathbf{x}) \geq \frac{A_n}{\mu(\mathcal{X})} \quad \text{for all } \mathbf{x} \in \tilde{\mathcal{X}}. \tag{9}$$

Combining the left side of (7) with (8) we obtain $\|S\Phi(\mathbf{x})\|_{\ell_2}^2 \geq a_n(\mathbf{x})$ for all $n \in \mathbb{N}$ and almost all $\mathbf{x}$. Since $a_n(\mathbf{x})$ is unbounded for a set $\tilde{\mathcal{X}}$ 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.

$$\begin{tikzcd}
\mathcal{X} \arrow[r, "\Phi"] & \Phi(\mathcal{X}) \arrow[r, "A^{-1}"] \arrow[d, phantom, "\cap"] & U_{\ell_2} \arrow[dl, "A"] \\
& \mathcal{E} &
\end{tikzcd} \tag{10}$$

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

$$A(U_{\ell_2}) \subseteq \mathcal{E}. \tag{11}$$

We can ensure this by constructing A such that

$$A\colon (x_j)_j \mapsto (R_A a_j x_j)_j \tag{12}$$

with $R_A := C_k \|(\sqrt{\lambda_j}/a_j)_j\|_{\ell_2}$. From Proposition 1 it follows that all those operators A for which $R_A < \infty$ will satisfy (11). We call such scaling (inverse) operators *admissible.*

4.2 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 [13, p. 226] (stated in the present form in [8, p. 17]).

Theorem 2. *Let $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_j \geq \cdots \geq 0$ be a non–increasing sequence of non–negative numbers and let*

$$D\mathbf{x} = (\sigma_1 x_1, \sigma_2 x_2, \ldots, \sigma_j x_j, \ldots) \tag{13}$$

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 $n \in \mathbb{N}$,

$$\sup_{j\in\mathbb{N}} n^{-\frac{1}{j}} (\sigma_1\sigma_2\cdots\sigma_j)^{\frac{1}{j}} \leq \epsilon_n(D) \leq 6 \sup_{j\in\mathbb{N}} n^{-\frac{1}{j}} (\sigma_1\sigma_2\cdots\sigma_j)^{\frac{1}{j}}. \tag{14}$$

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 1 (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 (12). Then*

$$\epsilon_n(A\colon \ell_2 \to \ell_2) \leq \inf_{(a_s)_s : \left(\sqrt{\lambda_s}/a_s\right)_s \in \ell_2} \sup_{j\in\mathbb{N}} 6C_k \left\| \left(\sqrt{\lambda_s}/a_s\right)_s \right\|_{\ell_2} n^{-\frac{1}{j}} (a_1 a_2 \cdots a_j)^{\frac{1}{j}}. \tag{15}$$

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 [7] and [11] to extend the Optimal Margin Hyperplane classifier to what is now known as the SV machine. (The "$+b$" term is readily dealt with; we omit such considerations here though.) 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 [33,32] 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

$$\begin{aligned} &S_{\tilde{\mathbf{X}}^m} : \ell_2 \to \ell_\infty^m \\ &S_{\tilde{\mathbf{X}}^m} : \mathbf{w} \mapsto (\langle \tilde{\mathbf{x}}_1, \mathbf{w} \rangle, \ldots, \langle \tilde{\mathbf{x}}_m, \mathbf{w} \rangle) \,. \end{aligned} \tag{16}$$

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 [36] for some extensions and historical remarks.

Theorem 3 (Carl and Stephani [8, 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 $1 \leq j \leq m$*

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

The restatement of Theorem 3 in terms of $\epsilon_{2^{n-1}} = e_n$ will be useful in the following. Under the assumptions above we have

$$\epsilon_n(S) \leq c\|S\| \left((\log n + 1)^{-1} \log\left(1 + \frac{m}{\log n + 1}\right)\right)^{1/2} . \tag{17}$$

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 2 (Carl and Stephani [8, p. 11]). *Let E, F, G be Banach spaces, $R \in \mathfrak{L}(F, G)$, and $S \in \mathfrak{L}(E, F)$. Then, for $n, t \in \mathbb{N}$,*

$$\epsilon_{nt}(RS) \leq \epsilon_n(R)\epsilon_t(S) \tag{18}$$

$$\epsilon_n(RS) \leq \epsilon_n(R)\|S\| \tag{19}$$

$$\epsilon_n(RS) \leq \epsilon_n(S)\|R\|. \tag{20}$$

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 4 (Bounds for SV classes). *Let k be a Mercer kernel, let Φ be induced via (5) and let $T := S_{\tilde{\mathbf{X}}^m} R_\mathbf{w}$ where $S_{\tilde{\mathbf{X}}^m}$ is given by (16) and $R_\mathbf{w} \in \mathbb{R}^+$. Let A be defined by (12) 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:*

$$\epsilon_n(T) \leq c\|A\|R_\mathbf{w} \log^{-1/2} n \log^{-1/2}\left(1 + \tfrac{m}{\log n}\right) \tag{21}$$

$$\epsilon_n(T) \leq R_\mathbf{w} \epsilon_n(A) \tag{22}$$

$$\epsilon_{nt}(T) \leq cR_\mathbf{w} \log^{-1/2} n \log^{-1/2}\left(1 + \tfrac{m}{\log n}\right) \epsilon_t(A)$$

where C_k and c are defined as in Corollary 1 and Lemma 3.

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

$$\begin{tikzcd}
U_{\ell_2} \arrow[r, "T"] \arrow[d, "R_{\mathbf{w}}"'] & \ell_\infty^m \\
R_{\mathbf{w}} U_{\ell_2} \arrow[r, "A"] \arrow[ur, "S_{\tilde{\mathbf{X}}^m}"] & R_{\mathbf{w}}\mathcal{E} \arrow[u, "S_{(A^{-1}\tilde{\mathbf{X}}^m)}"']
\end{tikzcd} \tag{23}$$

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

$$\langle \mathbf{w}, \tilde{\mathbf{x}} \rangle = \langle \mathbf{w}, AA^{-1}\tilde{\mathbf{x}} \rangle = \langle A\mathbf{w}, A^{-1}\tilde{\mathbf{x}} \rangle. \tag{24}$$

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 $\Phi(\mathcal{X})A^{-1} \in U_{\ell_2}$ filling a larger proportion of it than just $\frac{1}{\rho}\Phi(\mathcal{X})$.
By construction of A and the Cauchy-Schwarz inequality we have $\|S_{A^{-1}\tilde{\mathbf{X}}^m}\| = 1$. Thus applying lemma 2 to the factorization of T and using Theorem 3 proves the theorem. ■

One can give (see below) asymptotic rates of decay for $\epsilon_n(A)$. (In fact we can determine 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 3 (Rate bounds on ϵ_n). *Let k be a Mercer kernel and suppose A is the scaling operator associated with it as defined by (12).*

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

This Lemma (the proof of which is omitted; see [35]) shows that in the first case, Maurey's result (theorem 3) allows an improvement in the exponent of the entropy number of T, whereas in the second, it affords none (since the entropy numbers decay so fast anyway). The Maurey result may still help in that case though for nonasymptotic n. 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 $k(x, y) = \exp(-(x - y)^2)$.

4.3 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 (10) which performs the mapping described above. We merely flag this here — the full development of these ideas requires considerable further work and will be deferred to a subsequent paper. There are some remarks in the full version of this paper [35]. Furthermore the statistical argument needed to exploit such techniques (bounding generalization error in terms of *empirical* covering numbers has now been developed — see [29].

5 Eigenvalue Decay Rates

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. A commonly used kernel is $k(x, y) = e^{-(x-y)^2}$ which has noncompact support. The induced integral operator $(T_k f)(x) = \int_{-\infty}^{\infty} k(x, y) f(y) dy$ then has a continuous spectrum 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? This problem can be readily resolved by analysing the v-periodic extension of the kernel in question $k_v(x) := \sum_{j=-\infty}^{\infty} k(x - jv)$. A simple argument gives

Lemma 4. *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* $k(\cdot)$ *and* k_v *denote the* v*–periodic 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* $k_v(x - y) = \sum\limits_{j=-\infty}^{\infty} \frac{\sqrt{2\pi}}{v} K(j\omega_0) e^{ij\omega_0 x}$ *Moreover* $\lambda_j = \sqrt{2\pi} K(j\omega_0)$ *for* $j \in \mathbb{Z}$ *and* $C_k = \sqrt{\frac{2}{v}}$.

This lemma tells one how to compute the discrete eigenvalue sequence for kernels with infinite support; for more details see [35].

The above results show the overall covering numbers of a SV machine are controlled by the entropy numbers of the admissible scaling operator A: $\epsilon_n(A\colon \ell_2 \to \ell_2)$. One can work this out (with constants), although it is somewhat intricate to do so. Here we simply state how $\epsilon_n(A)$ depends asymptotically on the eigenvalues of T_k for a certain class of kernels.

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

An example of such a kernel (for $p = 2$) is $k(x) = e^{-x^2}$. It can also be shown that the rate in the above proposition is asymptotically tight. For a proof, and related results, see [35].

6 The Missing Pieces and Some Conclusions

In this short version we have omitted many details and extensions such as

Discretization How should one choose v in periodizing a non-compact kernel?

Higher Dimensions The results need to be extended to multi-dimensional kernels to be practically useful. Several additional technical complications arise in doing so.

Glueing it all Together We have given the ingredients but not baked the cake. Since the approach we have taken is new, and since there are a wide range of different uniform convergence results one may use we have refrained from putting it all together into "master generalization error theorem." It should be clear that it is *possible* to do so.

Combining all these pieces together does give an (albeit complicated) answer to the question "what is the effect of the kernel?" Different kernels, or even different widths of the same kernel, give rise to different covering numbers and hence different generalization performance. We hope eventually to be able to give simple rules of thumb concerning the overall effect. The mere fact that entropy number techniques provide a handle on the question is interesting in itself though.

In summary, 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 method used to obtain the results (reasoning via entropy numbers of operators) would seem to be a nice new viewpoint and valuable for other problems.

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. An operator theoretical 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. *Journal 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. Robert Ash. *Information Theory.* Interscience Publishers, New York, 1965.
7. 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.
8. B. Carl and I. Stephani. *Entropy, compactness, and the approximation of operators.* Cambridge University Press, Cambridge, UK, 1990.
9. Bernd Carl. Entropy numbers of diagonal operators with an application to eigenvalue problems. *Journal of Approximation Theory*, 32:135–150, 1981.
10. Bernd 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. C. Cortes and V. Vapnik. Support vector networks. *Machine Learning*, 20:273 – 297, 1995.
12. Martin Defant and Marius Junge. Characterization of weak type by the entropy distribution of r-nuclear operators. *Studia Mathematica*, 107(1):1–14, 1993.
13. 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.
14. Leonid Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in Banach spaces. Technical report, NEC Research Institute, 1997. To appear in ALT97 Proceedings.
15. D. Jagerman. ε-entropy and approximation of bandlimited functions. *SIAM Journal on Applied Mathematics*, 17(2):362–377, 1969.
16. Marius Junge and Martin Defant. Some estimates of entropy numbers. *Israel Journal of Mathematics*, 84:417–433, 1993.
17. V.I. Kolchinskiĭ. Operators of type p and metric entropy. *Teoriya Veroyatnosteĭi Matematicheskaya Statistika*, 38:69–76, 135, 1988. (In Russian. MR 89j:60007).
18. 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.
19. 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.
20. H. König. *Eigenvalue Distribution of Compact Operators.* Birkhäuser Verlag, Basel, 1986.
21. 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.
22. Alain Pajor. *Sous-espaces ℓ_n^1 des espaces de Banach.* Hermann, Paris, 1985.
23. Albrecht Pietsch. *Operator ideals.* North-Holland, Amsterdam, 1980.
24. L.S. Pontriagin and L.G. Schnirelmann. Sur une propriété métrique de la dimension. *Annals of Mathematics*, 33:156–162, 1932.
25. R.T. Prosser. The ε–Entropy and ε–Capacity of Certain Time–Varying Channels. *Journal of Mathematical Analysis and Applications*, 16:553–573, 1966.

26. 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.
27. C.E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.
28. J. Shawe-Taylor, P.L. Bartlett, R.C. Williamson, and M. Anthony. Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, 44(5):1926–1940, 1998.
29. J. Shawe-Taylor and Robert C. Williamson. Generalization performance of classifiers in terms of observed covering numbers. 4th European Conference on Computational Learning Theory.
30. A. J. Smola, B. Schölkopf, and K.-R. Müller. The connection between regularization operators and support vector kernels. *Neural Networks*, 1998. in press.
31. H. Triebel. Interpolationseigenschaften von Entropie- und Durchmesseridealen kompackter Operatoren. *Studia Mathematica*, 34:89–107, 1970.
32. V. Vapnik. *The Nature of Statistical Learning Theory.* Springer Verlag, New York, 1995.
33. 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).
34. 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.
35. R.C. Williamson, A. Smola, and B. Schölkopf. Generalization performance of regularization networks and support vector machines entropy numbers of compact operators. Technical report, Neurocolt Technical Report 1998-019, 1998. ftp://www.neurocolt.com/pub/neurocolt/tech_reports/1998/98019.ps.Z.
36. Robert C. Williamson, Bernhard Schölkopf, and Alex Smola. A Maximum Margin Miscellany. Typescript, March 1998.

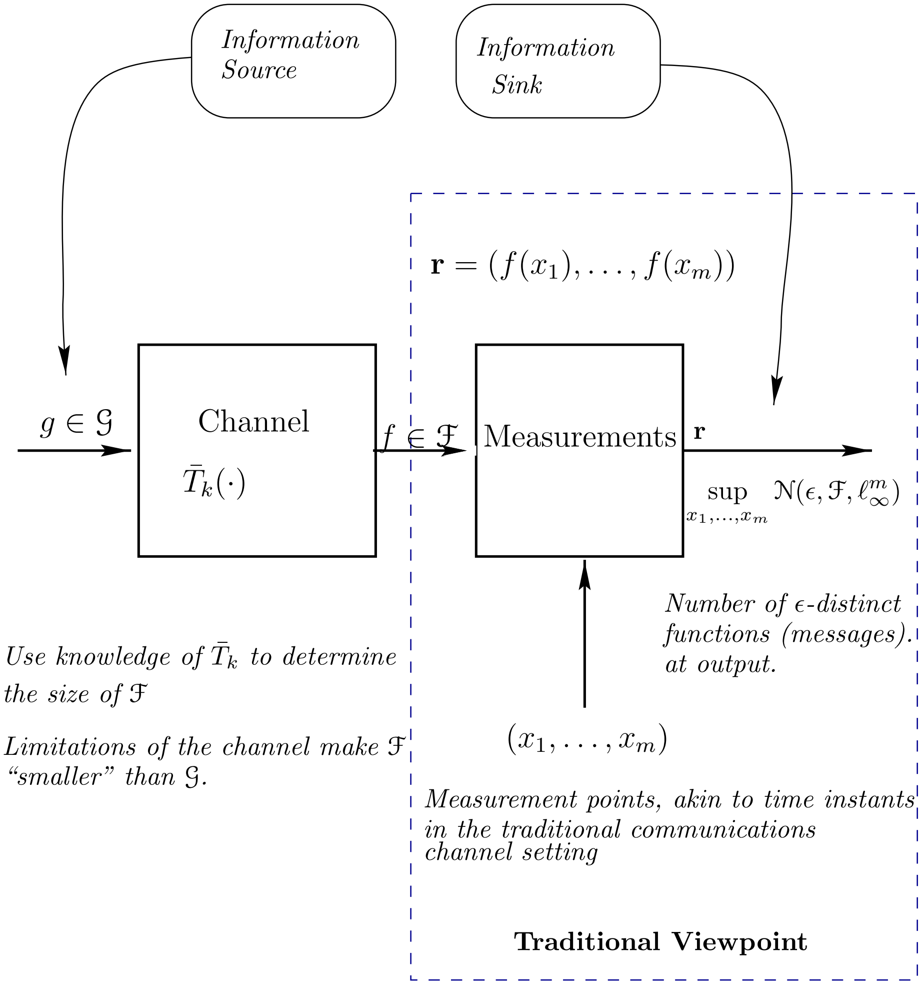
Information Source
Information Sink
$\mathbf{r} = (f(x_1), \ldots, f(x_m))$
$g \in \mathcal{G}$
Channel
$\bar{T}_k(\cdot)$
$f \in \mathcal{F}$
Measurements
$\mathbf{r}$
$\sup_{x_1,\ldots,x_m} \mathcal{N}(\epsilon, \mathcal{F}, \ell_\infty^m)$
Number of ϵ-distinct functions (messages). at output.
Use 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 Viewpoint

Fig. 1. Schematic picture of the new viewpoint.