The Entropy Regularization Information Criterion

A.J. Smola\textsuperscript{5,*}, J. Shawe-Taylor\textsuperscript{4}, B. Schölkopf\textsuperscript{5,*}, R.C. Williamson\textsuperscript{*}
\textsuperscript{5}GMD FIRST, Rudower Chaussee 5, 12489 Berlin, Germany
\textsuperscript{4}Department of Engineering, Australian National University, Canberra 0200, Australia
\textsuperscript{1}Royal Holloway College, University of London, Egham, UK
\{smola, bs\}@first.gmd.de, Bob.Williamson@anu.edu.au, john@dcs.rhbnc.ac.uk

Abstract

Effective methods of capacity control via uniform convergence bounds for function expansions have been largely limited to Support Vector machines, where good bounds are obtainable by the entropy number approach. We extend these methods to systems with expansions in terms of arbitrary (parametrized) basis functions and a wide range of regularization methods covering the whole range of general linear additive models. This is achieved by a data dependent analysis of the eigenvalues of the corresponding design matrix. Experimental evidence corroborates the new bounds.

1 INTRODUCTION

Model selection criteria based on the Vapnik-Chervonenkis (VC) dimension are known to be difficult to obtain, worst case, and often not very tight. Yet they have the theoretical appeal of providing bounds, with little or no assumptions made.

Recently new methods [10, 8, 7] have been developed which are able to provide a better characterization of the complexity of function classes than the VC dimension, and moreover, are easily obtainable and take advantage of the data at hand (i.e. they employ the concept of luckiness). These techniques, however, have been limited to linear functions or expansions of functions in terms of kernels as happens to be the case in Support Vector (SV) machines.

In this paper we show that the previously mentioned techniques can be extended to expansions in terms of arbitrary basis functions, covering a large range of practical algorithms such as general linear models, weight decay, sparsity regularization [4], and regularization networks [5].

2 SUPPORT VECTOR MACHINES

Support Vector machines carry out an effective means of capacity control by minimizing a weighted sum of the training error

\[ R_{\text{emp}}[f] \geq \frac{1}{m} \sum_{i=1}^{m} c(x_i, y_i, f(x_i)) \] (1)
and a regularization term \( Q[f] = \frac{1}{2}||w||^2 \); i.e. they minimize the regularized risk functional

\[
R_{\text{reg}}[f] := R_{\text{emp}}[f] + \lambda Q[f] = \frac{1}{m} \sum_{i=1}^{m} c(x_i, y_i, f(x_i)) + \frac{\lambda}{2}||w||^2.
\]

(2)

Here \( X := \{x_1, \ldots, x_m\} \subset \mathcal{X} \) denotes the training set, \( Y := \{y_1, \ldots, y_m\} \subset \mathcal{Y} \) the corresponding labels (target values), \( \mathcal{X}, \mathcal{Y} \) the corresponding domains, \( \lambda > 0 \) a regularization constant, \( c \) a cost function, and \( f : \mathcal{X} \rightarrow \mathcal{Y} \) is given by

\[
f(x) := \langle x, w \rangle, \text{ or in the nonlinear case } f(x) := \langle \Phi(x), w \rangle.
\]

(3)

Here \( \Phi : \mathcal{X} \rightarrow \mathcal{F} \) is a map into a feature space \( \mathcal{F} \). Finally, dot products in feature space can be written as \( \langle \Phi(x), \Phi(x') \rangle = k(x, x') \) where \( k \) is a so-called Mercer kernel.

For \( n \in \mathbb{N}, \mathbb{R}^n \) denotes the \( n \)-dimensional space of vectors \( x = (x_1, \ldots, x_n) \). We define spaces \( \ell_p^n \) as follows: as vector spaces, they are identical to \( \mathbb{R}^n \); in addition, they are endowed with \( p \)-norms: for \( 0 < p < \infty \), \( ||x||_p := (\sum_{j=1}^{n} |x_j|^p)^{1/p} \); for \( p = \infty \), \( ||x||_\infty := \max_{j=1,...,n} |x_j| \). We write \( \ell_p = \ell_p^\infty \). Let \( U_{\varepsilon} := \{ x : ||x||_p \leq 1 \} \) be the unit \( \ell_p \)-ball.

For model selection purposes one wants to obtain bounds on the richness of the map \( S_X \)

\[
S_X : w \mapsto (f(x_1), \ldots, f(x_m)) = (\langle \Phi(x_1), w \rangle, \ldots, \langle \Phi(x_m), w \rangle).
\]

(4)

where \( w \) is restricted to an \( \ell_2 \) unit ball of some radius \( \Lambda \) (this is equivalent to choosing an appropriate value of \( \lambda \)). By the “richness” of \( S_X \) specifically we mean the \( \ell_n^m \) \( \varepsilon \)-covering numbers of the set \( S_X(U_{\varepsilon}) \). (In the standard COLT notation, we mean \( N(\varepsilon, S_X(U_{\varepsilon}), \ell_n^m) \); see [10] for a fuller explanation.)

When carrying out model selection in this case, advanced methods [7] exploit the distribution of \( X \) mapped into feature space \( \mathcal{F} \), and thus of the spectral properties of the operator \( S_X \) by analyzing the eigenspectrum of the Gram matrix \( G = [g_{ij}]_{i,j} \), where \( g_{ij} := k(x_i, x_j) \).

All this is possible since \( k(x_i, x_j) \) can be seen as a dot product of \( x_i, x_j \) mapped into some feature space \( \mathcal{F} \), i.e. \( k(x_i, x_j) = \langle \Phi(x_i), \Phi(x_j) \rangle \). This property, whilst true for SV machines with Mercer kernels, does not hold in general case where \( f \) is expanded in terms of more or less arbitrary basis functions.

3 THE BASIC PROBLEMS

One basic problem is that when expanding \( f \) into

\[
f(x) = \sum_i \alpha_i f_i(x) \text{ where } \alpha_i \in \mathbb{R}
\]

(5)

with \( f_i(x) \) being arbitrary functions, it is not immediately obvious how to regard \( f \) as a dot product in some feature space. One can show that the VC dimension of a set of \( n \) linear independent functions is \( n \). Hence one would intuitively try to restrict the class of admissible models by controlling the number of basis functions \( n \) in terms of which \( f \) can be expanded.

Now consider an extreme case. In addition to the \( n \) basis functions \( f_i \) defined previously, we are given \( n \) further basis functions \( f'_i \), linearly independent of the previous ones, which differ from \( f_i \) only on a small domain \( \mathcal{X}' \), i.e. \( f'_i|_{\mathcal{X}\setminus\mathcal{X}'} = f_i|_{\mathcal{X}\setminus\mathcal{X}'} \). Since this new set of functions is linearly independent, the VC dimension of the joint set is given by \( 2n \). On the other hand, if hardly any data occurs on the domain \( \mathcal{X}' \), one would not notice the difference.
between \( f_i \) and \( f_i' \). In other words, the joint system of functions would behave as if we only had the initial system of \( n \) basis functions.

An analogous situation occurs if \( f_i' = f_i + \epsilon g_i \) where \( \epsilon \) is a small constant and \( g_i \) was bounded, say, within \([0, 1] \). Again, in this case, the additional effect of the set of functions \( f_i' \) would be hardly noticeable, but still, the joint set of functions would count as one with VC dimension \( 2n \). This already indicates, that simply counting the number of basis functions may not be a good idea after all.

Finally we have the practical problem that capacity control, which in SV machines was carried out by minimizing the “weight vector” \( w \), cannot be done in an analogous way either. There are several ways to do this. Below we consider three that have appeared in the literature and for which there exists effective algorithms.

**Example 1 (Weight Decay)** Define \( Q[f] := \frac{1}{2} \sum_i \alpha_i^2 \); i.e. the coefficients \( \alpha_i \) of the function expansion are constrained to an \( \ell_2 \) ball. In this case we can consider the following operator \( S_X^{(1)} : \ell_2^n \to \ell_\infty^n \), where

\[
S_X^{(1)}: \alpha \mapsto (f(x_1), \ldots, f(x_m)) = (\langle f(x_1), \alpha \rangle, \ldots, \langle f(x_m), \alpha \rangle)
\]  

Here \( f(x) := (f_1(x), \ldots, f_n(x)) \), \( \alpha := (\alpha_1, \ldots, \alpha_n) \) and \( \alpha \in \Lambda U_{\ell_2^n} \) for some \( \Lambda > 0 \).

**Example 2 (Sparsity Regularization)** In this case \( Q[f] := \sum_i |\alpha_i| \), i.e. the coefficients \( \alpha_i \) of the function expansion are constrained to an \( \ell_1 \) ball to enforce sparseness [4]. Thus \( S_X^{(2)} : \ell_1^n \to \ell_\infty^n \) with \( S_X^{(2)} \) mapping \( \alpha \) as in (6) except \( \alpha \in \Lambda U_{\ell_1^n} \). This is similar to expansions encountered in boosting or in linear programming machines.

**Example 3 (Regularization Networks)** Finally one could set \( Q[f] := \frac{1}{2} \alpha^\top Q \alpha \) for some positive definite matrix \( Q \). For instance, \( Q_{ij} \) could be obtained from \( (P f_1, P f_j) \) where \( P \) is a regularization operator penalizing unsmooth functions [5]. In this case \( \alpha \) lives inside some \( n \)-dimensional ellipsoid. By substituting \( \beta := Q^{1/2} \alpha \) one can reduce this setting to the case of example 1 and consider an evaluation operator \( S_X^{(3)} : \ell_2^n \to \ell_\infty^n \) given by

\[
S_X^{(3)}: \beta \mapsto (f(x_1), \ldots, f(x_m)) = (\langle Q^{-1/2} f(x_1), \beta \rangle, \ldots, \langle Q^{-1/2} f(x_m), \beta \rangle)
\]  

where \( \beta \in \Lambda U_{\ell_2^n} \) for some \( \Lambda > 0 \).

### 4 ENTROPY NUMBERS

Covering numbers characterize the difficulty of learning elements of a function class. Entropy numbers of operators can be used to compute covering numbers more easily and more tightly than the traditional techniques based on VC-like dimensions such as the fat shattering dimension [1]. Roughly speaking, knowing \( e_1(S_X) = \epsilon \) (see below for the definition) tells one that \( \log \mathcal{N}(\epsilon, F, \ell_\infty^n) \leq l \), where \( F \) is the effective class of functions used by the regularized learning machines under consideration. In this section we summarize a few basic definitions and results as presented in [10] and [3]. The \( l \)th entropy number \( e_l(F) \) of a set \( F \) with a corresponding metric \( d \) is the precision up to which \( F \) can be approximated by \( l \) elements of \( F \); i.e. for all \( f \in F \) there exists some \( f_i \in \{ f_1, \ldots, f_l \} \) such that \( d(f, f_i) \leq \epsilon_l \). Hence \( e_l(F) \) is the functional inverse of the covering number of \( F \). The entropy number of an bounded linear operator \( T : A \to B \) between normed linear spaces \( A \) and \( B \) is defined as \( e_l(T) := e_l(T(U_A)) \) with the metric \( d \) being induced by \( \| \cdot \|_B \). The dyadic entropy numbers \( e_l \) are defined by \( e_l := e_{2^l} \).

We make use of the following three results on entropy numbers of the identity mapping from \( \ell_p^n \) into \( \ell_p^n \), diagonal operators, and products of operators. Let

\[
\text{id}^{\ell_2^n}_{p_1, p_2} : \ell_2^n \to \ell_2^n \quad ; \quad \text{id}^{\ell_2^n}_{p_1, p_2} : x \mapsto x
\]
The following result is due to Schütz; the constants 9.94 and 1.86 were obtained in [9].

**Proposition 1 (Entropy numbers for identity operators)** Be \( m \in \mathbb{N} \). Then

\[
e_1(\text{id}_{1,2}^m) \leq 9.94 \left( \frac{1}{l} \log \left( 1 + \frac{n}{l} \right) \right)^{\frac{1}{2}} \quad \text{and} \quad e_1(\text{id}_{2,\infty}^m) \leq 1.86 \left( \frac{1}{l} \log \left( 1 + \frac{n}{l} \right) \right)^{\frac{1}{2}}
\]  

(8)

**Proposition 2 (Carl and Stephani [3, p. 11])** Let \( E, F, G \) be Banach spaces, \( R : F \to G \), and \( S : E \to F \). Then, for \( n, t \in \mathbb{N} \),

\[
e_{nt} (RS) \leq e_n(R) e_t(S), \quad e_n(RS) \leq e_n(R) \|S\| \quad \text{and} \quad e_n(RS) \leq e_n(S) \|R\|.
\]  

(9)

Note that the latter two inequalities follow directly from the fact that \( e_1(R) = \|R\| \) for all \( R : F \to G \) by definition of the operator norm \( \|R\| \).

**Proposition 3** Let \( \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_j \geq \cdots \geq 0, 1 \leq p \leq \infty \) and

\[
Dx = (\sigma_1 x_1, \sigma_2 x_2, \ldots, \sigma_j x_j, \ldots)
\]  

(10)

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

\[
\sup_{j \in \mathbb{N}} n^{-\frac{1}{p}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{p}} \leq e_n(D) \leq 6 \sup_{j \in \mathbb{N}} n^{-\frac{1}{p}} (\sigma_1 \sigma_2 \cdots \sigma_j)^{\frac{1}{p}}.
\]  

(11)

**5 THE MAIN RESULT**

We can now state the main theorem which gives bounds on the entropy numbers of \( S_X^{(i)} \) for the three examples of model selection described above.

**Proposition 4** Let \( f \) be expanded in a linear combination of basis functions as \( f := \sum_{i=1}^n \alpha_i f_i \) and the coefficients \( \alpha \) restricted to one of the convex sets as described in the examples 1 to 3. Moreover denote by \( F_{ij} := f_j(x_i) \) the design matrix on a particular sample \( X \), and by \( Q \) the regularization matrix in the case of example 3. Then the following bound on \( S_X^{(i)} \) holds.

1. **In the case of weight decay (ex. 1) (with \( l_1 + l_2 \geq l + 1 \)**

\[
e_1(S_X^{(1)}) \leq 1.96 \left( l_1^{-1} \log (1 + m / l_1) \right)^{\frac{1}{2}} e_{l_1}(\Sigma).
\]  

(12)

2. **In the case of weight sparsity regularization (ex. 2) (with \( l_1 + l_2 + l_3 \geq l + 2 \)**

\[
e_1(S_X^{(2)}) \leq 18.48 \left( l_1^{-1} \log (1 + m / l_1) \right)^{\frac{1}{2}} e_{l_1}(\Sigma) \left( l_2^{-1} \log (1 + m / l_2) \right)^{\frac{1}{2}} e_{l_2}(\Sigma).
\]  

(13)

3. **Finally, in the case of regularization networks (ex. 3) (with \( l_1 + l_2 \geq l + 1 \)**

\[
e_1(S_X^{(3)}) \leq 1.96 \left( l_1^{-1} \log (1 + m / l_1) \right)^{\frac{1}{2}} e_{l_1}(\Sigma).
\]  

(14)

Here \( \Sigma \) is a diagonal scaling operator (matrix) with \((i, i)\) entries \( \sqrt{\sigma_i} \) and \((\sqrt{\sigma_i})_i\) are the eigenvalues (sorted in decreasing order) of the matrix \( FF^T \) in the case of examples 1 and 2, and \( FQ^{-1}F^T \) in the case of example 3.
The entropy number of $\Sigma$ is readily bounded in terms of $(\sigma_i)_i$ by using (3). One can see that the first setting (weight decay) is a special case of the third one, namely when $Q = I$, i.e. when $Q$ is just the identity matrix.

**Proof** The proof relies on a factorization of $S_X^{(i)} (i = 1, 2, 3)$ in the following way. First we consider the equivalent operator $S_X$ mapping from $\ell^2_2$ to $\ell^m_2$ and perform a singular value decomposition [6] of the latter into $S_X = V\Sigma W$ where $V, W$ are operators of norm 1, and $\Sigma$ contains the singular values of $S_X^{(i)}$, i.e. the singular values of $F$ and $FQ^{-\frac{1}{2}}$ respectively. The latter, however, are identical to the square root of the eigenvalues of $FF^\top$ or $FQ^{-1}F^\top$. Consequently we can factorize $S_X^{(i)}$ as in the diagram

$$
\begin{align*}
\ell_2^1 \xrightarrow{id} \ell_2^2 & \xrightarrow{S_X^{(i)}} \ell_2^m \\
\ell_2^2 & \xrightarrow{W} \ell_2^3 & \xrightarrow{V} \ell_2^2
\end{align*}
$$

Finally, in order to compute the entropy number of the overall operator one has to use the factorization of $S_X$ into $S_X^{(i)} = \text{id}_{2,\infty}V\Sigma W$ for $i \in \{1, 3\}$, and into $S_X^{(2)} = \text{id}_{1,\infty}V\Sigma W\text{id}_{1,2}$ for example 2, and apply Proposition 2 several times. We also exploit the fact that for singular value decompositions $||V||, ||W|| \leq 1$.

The present theorem allows us to compute the entropy numbers (and thus the complexity) of a class of functions on the current sample $X$. Going back to the examples of section 3, which led to large bounds on the VC dimension one can see that the new result is much less susceptible to such modifications: the addition of $f_1, \ldots, f_n$ to $f_1, \ldots, f_m$ does not change the eigenspectrum $\Sigma$ of the design matrix significantly (possibly only doubling the nominal value of the singular values), if the functions $f_i$ differ from $f_j$ only slightly. Consequently the bounds will not change significantly even though the number of basis functions just doubled.

Also note that the current error bounds reduce to the results of [7] in the SV case: here $Q_{ij} = F_{ij} = k(x_i, x_j)$ (both the design matrix $F$ and the regularization matrix $Q$ are determined by kernels) and therefore $FQ^{-1}F = Q$. Thus the analysis of the singular values of $FQ^{-1}F$ leads to an analysis of the eigenvalues of the kernel matrix, which is exactly what is done when dealing with SV machines.

6  **ERROR BOUNDS**

To use the above result we need a bound on the expected error of a hypothesis $f$ in terms of the empirical error (training error) and the observed entropy numbers $\epsilon_n(\mathcal{F})$. We use [7, Theorem 4.1] with a little modification.

**Theorem 1** Let $\mathcal{F}$ be a set of linear functions as described in the previous examples with $\epsilon_n(S_X)$ as the corresponding bound on the observed entropy numbers of $\mathcal{F}$ on the dataset $X$. Moreover suppose that for a fixed threshold $b \in \mathbb{R}$ for some $f \in \mathcal{F}$, $\text{sgn}(f - b)$ correctly classifies the set $X$ with a margin $\gamma := \min_{1 \leq i \leq m} |f(x_i) - b|$.

Finally let $U := \min \{ n \in \mathbb{N} \mid \epsilon_n(S_X) \leq \gamma/8.001 \}$ and $\alpha(U, \delta) := 3.08(1 + \frac{1}{\ln \frac{1}{\delta}})$. Then with confidence $1 - \delta$ over $X$ (drawn randomly from $P^n$ where $P$ is some probability distribution) the expected error of $\text{sgn}(f - b)$ is bounded from above by

$$
\epsilon(m, U, \delta) = \frac{2}{m} \left( \frac{1 + \alpha(U, \delta)}{2} \log \left( \frac{5 n}{m} \right) \log(17m) \right) + \log \left( \frac{4 m}{\delta} \right).
$$

(16)
Figure 1: Task: separate digits 0 through 4 from 5 through 9 (top row), and separate even from odd digits (bottom row). Shown are the test error, the new bound (more precisely, the "effective VC-Dimension $U$), and the log of the old $R^2/\gamma^2$ bound (cf. text), all as functions of $\log(\sigma)$. In all experiments, we divided the training set into 23 random subsets of size 317 ($23 \cdot 317$ being the prime factorization of the training set size). Error bars in figures 1 denote standard deviations over the 23 trials. At the beginning of the experiment, the whole USPS set (training plus test set) was permuted, to ensure that the distribution of training and test data is the same.

The proof is essentially identical to that of [7, Theorem 4.1] and is omitted. [7] also shows how to compute $e_n(S_X)$ efficiently including an explicit formula for evaluating $e_l(\Sigma)$.

7 EXPERIMENTS

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

For convenience and in order to be able to compare our results with other tight VC–type bounds (using the number of basis functions as a comparison would have led to trivial bounds since it is identical to the number of samples) we tested the new bounds on a Support Vector classifier. SVMs usually come with two free parameters: the regularization constant $C$ and the kernel parameter. To make our bounds applicable, we chose the regularization such that the SVM attains zero training error over all experiments, thus focusing our model selection efforts on the kernel parameter, which in our case was the width $\sigma$ of a Gaussian kernel $k(x, y) = \exp(-||x - y||^2/\sigma)$.

For a range of values of $\sigma$, we computed the SVM hyperplane and evaluated the bound (16). Due to the constant offset term $b$ in the SV expansion we computed the kernel Gram matrix $g_{ij} = k(x_i, x_j)$ for centered data. In other words we performed a rank 1 update of $g_{ij}$ to achieve zero mean in rows and columns. Thus we computed the eigenspectrum of $(1 - 1_m)g(1 - 1_m)$, where $(1_m)_{ij} := 1/m$ and $1_{ij} = \delta_{ij}$ for $1 \leq i, j \leq m$. Incidentally this is precisely the matrix used in kernel PCA.

¹The final version will include an analysis of further learning models.
The results are given in figure 1. For both classification problems, the minimum of the test error occurs at the value of $\sigma$ which minimizes the new bound. Moreover, the new bound even resembles the shape of the test error curve closely.

The previously known bound [2], involving the fat-shattering dimension $R^2/\gamma^2$, led to worse predictions of the optimal $\sigma$. Essentially, in our situation this bound states that we should select the $\sigma$ which leads to the maximum margin: the value of $R$ was estimated as 1 since in the case of RBF-kernels, $k(x, x) = 1$ for all $x$.

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

8 DISCUSSION

We showed how good bounds could be obtained on the entropy numbers of a wide class of popular statistical estimators ranging from weight decay to sparsity regularization (with SV machines being a special case thereof). The results are given in a way that is directly usable for practitioners without any tedious calculations of the VC dimension or similar combinatorial quantities. In particular, our method ignores (nearly) linear dependent basis functions automatically. Finally, it takes advantage of favourable distributions of data by using the observed entropy numbers as a base for stating bounds on the true entropy numbers with respect to the function class under consideration.

Acknowledgements: This work was supported by the Australian Research Council.

References


