A Note on Different Covering Numbers in Learning Theory

Massimiliano Pontil[a*]

[a]Department of Computer Sciences, University College London
Gower Street London WC1E, England
Email: m.pontil@cs.ucl.ac.uk

The covering number of a set $\mathcal{F}$ in the space of continuous functions on a compact set X plays an important role in learning theory. In this paper we study the relation between this covering number and its discrete version, obtained by replacing X with a finite subset. We formally show that when $\mathcal{F}$ is made of smooth functions, the discrete covering number is close to its continuous counterpart. In particular, we illustrate this result in the case that $\mathcal{F}$ is a ball in a reproducing kernel Hilbert space.

1. INTRODUCTION

Let $C(X)$ be the Banach space of continuous functions on a compact set $X \subset \mathbb{R}^n$ with the norm $\|f\| = \sup_{x \in X} |f(x)|$, and $\mathcal{H} \subset C(X)$ a Hilbert space with the norm $\| \cdot \|_{\mathcal{H}}$. We denote by B_R the ball of radius R in $\mathcal{H}$ and by $\mathcal{N}(B_R, \eta)$ the $\eta-$covering number of B_R using the norm of $C(X)$, i.e. the minimal $\ell \in \mathbb{N} \cup \{\infty\}$ such that there exist ℓ disks in B_R of radius η covering B_R. We assume that this number is finite for every $\eta > 0$ or, equivalently, that B_R is pre-compact in $C(X)$.

We study the dependency of $\mathcal{N}(B_R, \eta)$ on the space X. In particular, we consider the case where X is replaced by a finite subset. This problem is motivated by recent results in [3] where the covering numbers of compact sets of $C(X)$ are shown to play a fundamental role in the problem of bounding the deviation between expected and empirical error functionals studied in learning theory.

In the related statistical learning theory [9] the setting of the problem is similar but with the important difference that the covering number is computed by using a semi-norm in $C(X)$, namely the maximum norm of f with respect to (w.r.t) a finite set of points belonging to X. Let $\mathbf{x} = \{x_1, \ldots, x_m\} \subset X$ be such a set. We denote by $\mathcal{N}_{\mathbf{x}}(B_R, \eta)$ the $\eta-$covering number of B_R when the maximum norm over the set $\mathbf{x}$ is used, i.e. $\max_{i=1}^{m} |f(x_i)|$.

We show that, if $\mathcal{H}$ has some kind of Hölder continuous property, the covering number of B_R does not change much as a function of X. This is summarized by the following theorem.

---

*Most of this work was made while the author was visiting the City University of Hong Kong in February and March 2002. The work has been supported by the Research Grants Council of the Hong Kong SAR (project number: CityU 1002/99P).

Theorem 1 *Suppose that for any* $f \in \mathcal{H}$ *and* $x, t \in X$ *such that* $\|x - t\| \leq \delta$ *we can write*

$$|f(x) - f(t)| \leq \|f\|_{\mathcal{H}} \Delta(\|x - t\|)$$

with $\Delta(\cdot)$ *a positive continuous function which satisfies* $\Delta(0) = 0$. *Then, for every* $\eta > 0$, *we have*

$$\mathcal{N}(B_R, \eta + 2R\Delta(\nu(\mathbf{x}))) \leq \mathcal{N}_{\mathbf{x}}(B_R, \eta) \leq \mathcal{N}(B_R, \eta)$$

where we have defined $\nu(\mathbf{x}) = \inf\{a > 0 \mid X \subseteq \bigcup_{i=1}^{m} D(x_i, a)\}$.

The proof of Theorem 1 is given in Section 2 where we also discuss its implications in learning theory. In Section 3 we discuss Theorem 1 in the context of reproducing kernel Hilbert spaces.

2. RELATION BETWEEN THE COVERING NUMBER IN $C(X)$ AND ITS DISCRETE APPROXIMATION

The idea behind proving Theorem 1 is based on the simple observation that, under the Hölder property hypothesis, the norm in $C(X)$ can be bounded by a linear function of the semi-norm w.r.t to a finite set of points $\mathbf{x}$.

Proof of Theorem 1: The right hand side (r.h.s.) follows immediately from the inequality

$$\max_{i=1,\ldots,m} |f(x_i)| \leq \sup_{x \in X} |f(x)|.$$

To prove the left hand side inequality note that, since X is compact and by hypothesis $\bigcup_{i=1}^{m} D(x_i, \nu(\mathbf{x}))$ covers X, we can rewrite the norm in $C(X)$ as

$$\sup_{x \in X} |f(x)| = \max_{i=1,\ldots,m} \left\{ \sup_{x \in D(x_i, \nu(\mathbf{x}))} |f(x)| \right\}.$$

When $f \in \mathcal{H}$, we also have $|f(x) - f(x_i)| \leq \|f\|_{\mathcal{H}} \Delta(\|x - x_i\|)$ for every $x_i \in \mathbf{x}$, which combined with the last equation gives

$$\sup_{x \in X} |f(x)| \leq \max_{i=1,\ldots,m} |f(x_i)| + \|f\|_{\mathcal{H}} \Delta(\nu(\mathbf{x})).$$

Let $N = \mathcal{N}_{\mathbf{x}}(B_R, \eta)$ and $f_1, \ldots, f_N$ be the elements in $B_R(\mathcal{H})$ which realize the covering, i.e. for every $f \in B_R(\mathcal{H})$, $\max_{i=1,\ldots,m} |f(x_i) - f_n(x_i)| \leq \eta$ for some $n \in \{1, \ldots, N\}$. From last equation it follows that

$$\begin{aligned} \sup_{x \in X} |f(x) - f_n(x)| &\leq \max_{i=1,\ldots,m} |f(x_i) - f_n(x_i)| + \|f - f_n\|_{\mathcal{H}} \Delta(\nu(\mathbf{x})) \\ &\leq \eta + 2R\Delta(\nu(\mathbf{x})). \end{aligned}$$

Then, when using the norm of $C(X)$, $B_R(\mathcal{H})$ is covered by balls with centers f_n and radius $\eta + 2R\Delta(\nu(\mathbf{x}))$. QED.

Theorem 1 holds for every finite subset of X. In particular, since we assumed X to be compact, we can take $\mathbf{x}$ to be a minimal $\epsilon-$net of X of size m. In this case $\nu(\mathbf{x})$ is the $m-$entropy number of X, $\epsilon_m(X)$, which is defined as the minimal positive a such that there exist m closed balls in X with radius a covering X. This number can be bounded as a function of $n = dim(X)$. For example, in [3] it is shown that

$$\epsilon_m(X) \leq 8r(m+1)^{-\frac{1}{n}}$$

where r is the radius of the smallest sphere containing X. Combining this inequality with Theorem 1 we have the following corollary.

Corollary 1 *Under the same hypotheses of Theorem 1 there exists, for every $m > 0$, a set of m points in X, $\widehat{\mathbf{x}} = \{\widehat{x}_1, \ldots, \widehat{x}_m\}$, such that*

$$\mathcal{N}\left(B_R, \eta + 2R\Delta\left(8r(m+1)^{-\frac{1}{n}}\right)\right) \leq \mathcal{N}_{\widehat{\mathbf{x}}}(B_R, \eta).$$

Remark 1: Theorem 1 also applies to the case that $\mathbf{x}$ is replaced by every subset of X. Let $\mathcal{N}_0(B_R, \eta)$ be the covering number w.r.t $X_0 \subset X$. If X_0 is dense in X, $\mathcal{N}_0(B_R, \eta) = \mathcal{N}(B_R, \eta)$. Thus, assuming that $\mathbf{x}$ becomes dense in X when $m \to \infty$, we also have $\lim_{m\to\infty} \mathcal{N}_{\mathbf{x}}(B_R, \eta) = \mathcal{N}(B_R, \eta)$.

Remark 2: If B_R is replaced by a compact subspace $\mathcal{F}$ of $\mathcal{H}$, Theorem 1 still holds true if we let R be the radius of $\mathcal{F}$, $R = \inf_{f\in\mathcal{H}} \sup_{g\in\mathcal{F}} \|f - g\|_{\mathcal{H}}$.

2.1. Covering number and sample complexity

Learning theory studies the problem of computing a function from a finite random sample. We briefly explain the problem here. For a more detailed account see, e.g., [1,3,5,9] and references therein.

We have two sets of variables $x \in X$ and $y \in Y \subseteq \mathbb{R}$ which are related by a probabilistic relationship $P(x, y)$ defined over the set $X \times Y$. Our desired function is the minimizer of the expected error

$$E(f) = \int (y - f(x))^2 P(x, y)\ dxdy.$$

Unfortunately this functional can not be computed because the probability distribution $P(x, y)$ is unknown. We are only provided with a training set of m pairs (x_i, y_i), $i = 1, \ldots, m$, sampled in $X \times Y$ according to $P(x, y)$. A natural approach is to replace the expected error with the empirical error

$$E_m(f) = \frac{1}{m}\sum_{i=1}^{m}(y_i - f(x_i))^2.$$

We then minimize E_m in a compact subset $\mathcal{F}$ of a Hilbert space $\mathcal{H}$. Let f_m be a minimizer. A main issue in the theory is to study conditions which guarantee that $E_m(f_m)$ is close to $E(f_m)$ in probability. Formally we require that

$$\text{Prob}\left\{|E(f_m) - E_m(f_m)| \leq \epsilon\right\} \geq 1 - \delta \qquad (1)$$

where the probability is w.r.t. the random draw of the training set and ϵ and δ are two small positive numbers. The answer to this question is related to the study of the covering number of $\mathcal{F}$. It is based on extending some classical probabilistic inequalities, such as Bernstein and Hoeffding's, to function spaces[2]. We assume that, for every $f \in \mathcal{F}$, $|y - f(x)| \leq M$ almost everywhere, and, without loss of generality we chose $M = 1$. For our purpose here it is sufficient to consider the results derived through Hoeffdings's Inequality [6]. A key result from Vapnik and Chervonenkis (see, e.g., Chapter 7 of [9] or [1]) establishes that

$$\delta = 12m \left[\sup_{|\mathbf{x}|=2m} \mathcal{N}_{\mathbf{x}} \left(\mathcal{F}, \frac{\epsilon}{6} \right) \right] e^{-\frac{\epsilon^2 m}{36}}. \tag{2}$$

A result with a similar flavor but with a much simpler proof, was recently derived by Cucker and Smale [3]. It says that[3]

$$\delta = 2\mathcal{N}(\mathcal{F}, \frac{\epsilon}{8}) e^{-\frac{\epsilon^2 m}{8}}. \tag{3}$$

Equations (2) and (3) can be inverted to obtain a lower bound on the number of samples m as a function of ϵ, δ and the covering number. For example, Equation (3) gives

$$m \geq \frac{8}{\epsilon^2} \left[\ln \mathcal{N}(\mathcal{F}, \frac{\epsilon}{8}) - \ln(\frac{\delta}{2}) \right].$$

This is also called a sample complexity bound: when m satisfies the bound, Inequality (1) holds true. Assuming that $\ln \mathcal{N}(\mathcal{F}, \eta)$ grows as η^{-q} [8], the sample complexity bound gives, for a fixed δ, $m = O(\epsilon^{-(2+q)})$. Now let us look at Equation (2). Corollary 1 implies that

$$\mathcal{N}(B_R, \eta') \leq \sup_{|\mathbf{x}|=m} \mathcal{N}_{\mathbf{x}}(B_R, \eta) \leq \mathcal{N}(B_R, \eta)$$

with $\eta' = \eta + 2R\Delta(8r(m+1)^{-\frac{1}{n}})$. We then see that $\mathcal{N}(B_R, \eta)$ is close to $\sup_{|\mathbf{x}|=m} \mathcal{N}_{\mathbf{x}}(B_R, \eta)$ if

$$2R\Delta \left(8r(m+1)^{-\frac{1}{n}} \right) \leq \eta.$$

Thus, assuming that $\Delta(\xi)$ goes to zero as ξ^s, $s > 0$, the last inequality implies

$$m \geq \frac{(2R)^{\frac{n}{s}} (8r)^n}{\eta^{\frac{n}{s}}} - 1 = O(\eta^{-\frac{n}{s}}).$$

We conclude that, under the assumption that $\ln \mathcal{N}(\mathcal{F}, \epsilon) = O(\epsilon^{-q})$, if $n \leq s(2+q)$, Equations (2) and (3) lead to the same sample complexity bound.

[2] For a nice introduction to this subject see Chapters 2 and 3 of [4].

[3] Note that the result in [3] in based on Bernstein's Inequality. However, the same argument in that paper remains true in the case of Hoeffdings's Inequality, leading to Equation (3).

3. SPACES WITH A REPRODUCING KERNEL

In this section we take the space $\mathcal{H}$ to be a reproducing kernel Hilbert space (RKHS) [2], which we farther denote by $\mathcal{H}_K$. We first recall few facts concerning the RKHS that we need in order to analyze Theorem 1 in this context. For a detailed overview on RKHS's consistent with our notation see [3].

Given a continuous, symmetric, and positive definite function $K : X \times X \to \mathbb{R}$, called *kernel*, the associated RKHS is defined as the completion of the span of the set $\{K_x = K(x, \cdot) \mid x \in X\}$ with the norm $\|\cdot\|_K$ induced by the inner product $(K_x, K_t)_K = K(x,t)$. Two important examples of kernels are the *polynomial kernel*, $K(x,t) = (x,t)^d$, with d a positive integer, and the *Gaussian kernel*, $K(x,t) = \exp\{-\beta\|x-t\|^2\}$, $\beta > 0$, where we denoted by $(\cdot,\cdot)$ be the scalar product in $\mathbb{R}^n$.

Let $\mathcal{L}^2_\mu(X)$ be the space of square integrable functions on X w.r.t the positive measure μ. We consider the integral operator associated to kernel K, $L_K : \mathcal{L}^2_\mu(X) \to C(X)$ defined as

$$(L_K g)(x) = \int_X K(x,t) g(t) d\mu(t)$$

and let $\{\phi_i(x), \lambda_i\}_{i=1}^\infty$ be a system of eigenvectors and eigenvalues of L_K.

Theorem 2 *If K is continuous, B_R is compact in $C(X)$. In addition the following inequalities hold for every $f \in \mathcal{H}_K$ and $x, t \in X$:*

$$|f(x)| \leq \|f\|_K \sqrt{K(x,x)} \tag{4}$$

$$|f(x) - f(t)| \leq \|f\|_K \sqrt{K(x,x) + K(t,t) - 2K(x,t)}. \tag{5}$$

Proof: We first notice that $\mathcal{H}_K$ can be seen as the image of an injective operator $L_{\sqrt{K}} : \mathcal{L}^2_\mu(X) \to C(X)$ defined by $L_{\sqrt{K}}\phi_i = \sqrt{\lambda_i}\phi_i$. Then, for every $f \in \mathcal{H}_K$ we can write $f = L_{\sqrt{K}} g$, with $g = \sum_{i=1}^\infty a_i \phi_i$. We have

$$|f(x)| = |(L_{\sqrt{K}} g)(x)| = \sum_{i=1}^\infty a_i \sqrt{\lambda_i} \phi_i(x) = (a, \Phi(x))_{\ell^2}$$

where we have defined the map $\Phi : X \to \ell^2$ by $\Phi_i(x) = \sqrt{\lambda_i}\phi_i(x)$. As shown in Theorem 3, Chapter 3 of [3], this map is well defined, continuous and satisfies $(\Phi(x), \Phi(t))_{\ell^2} = K(x,t)$. Applying the Cauchy-Schwartz inequality to the r.h.s. of inequality above, we obtain

$$|f(x)| \leq \|a\|_{\ell^2} \sqrt{\sum_{i=1}^\infty \lambda_i \phi_i^2(x)} = \|f\|_K \sqrt{K(x,x)}.$$

This proves Inequality (4). Inequality (5) is proved similarly, by observing that

$$|f(x) - f(t)| \leq \|f\|_K \|\Phi(x) - \Phi(t)\|_{\ell^2}$$

and using $(\Phi(x), \Phi(t))_{\ell^2} = K(x,t)$.

Finally, we show that $L_{\sqrt{K}}$ is compact. This implies that B_R is compact in C(X). First notice that, since K is continuous, Inequality (4) implies that $L_{\sqrt{K}}$ is bounded and $\|L_{\sqrt{K}}\| \leq \sup_{x\in X}\sqrt{K(x,x)}$. To see that $L_{\sqrt{K}}$ is compact, consider a bounded sequence $\{f_n\}_{n=1}^{\infty}$ in $\mathcal{L}^2_{\mu}(X)$. By Inequality (4), $(L_{\sqrt{K}}f_n)$ is uniformly bounded and by Inequality (5) it is equicontinuous. Therefore by Arzelà's Theorem (see, e.g., Chapter 11.4 of [7]) $L_{\sqrt{K}}$ is compact. QED.

Remark 3: Theorem 2 improves Proposition 1 in [3], where it is shown that L_K is compact. Our result indeed shows that L_{K^t} is compact if $t \geq 1/2$.

Equation (5) is not yet in the form required by the hypotheses of Theorem 1. In the case, common in practice, that the kernel K is smooth, we can explicitly characterize the form of the function Δ. We assume in particular that K belongs to $C^2(X \times X)$. Let $K^{[1,0]}(s,t)$ be the gradient of $K(s,t)$ w.r.t. to s, $K^{[2,0]}(s,t)$ the $n\times n$ matrix formed by the second order partial derivatives of $K(s,t)$ w.r.t to s, and $K^{[1,1]}(s,t)$ the $n \times n$ matrix formed by the second order partial derivatives of $K(s,t)$ w.r.t. to one component of s and one of t. Likewise, we define $K^{[0,1]}(s,t)$ and $K^{[0,2]}(s,t)$, and note that, since K is symmetric, $K^{[0,1]}(s,t) = K^{[1,0]}(t,s)$ and $K^{[0,2]}(s,t) = K^{[2,0]}(t,s)$. With this notation at hand, the expansion of $K(s,t)$ is power series reads:

$$\begin{aligned} K(s,t) &= K(x,x) + (K^{[1,0]}(x,x), s-x) + (K^{[1,0]}(x,x), t-x) + \\ &+\frac{1}{2}(s-x, K^{[2,0]}(x,x)(s-x)) + \frac{1}{2}(t-x, K^{[2,0]}(x,x)(t-x)) \\ &+(s-x, K^{[1,1]}(x,x)(t-x)) + O(\max[(t-x)^3, (s-x)^3]). \end{aligned}$$

Applying this formula to the r.h.s. of Equation (5), we obtain

$$|f(x)-f(t)|^2 \leq (x-t, K^{[1,1]}(x,x)(x-t)) \leq \|K^{[1,1]}(x,x)\|\|x-t\|^2.$$

Therefore, $\mathcal{H}_K$ satisfies the hypotheses of Theorem 1 with

$$\Delta(\nu) = \sup_{x\in X} \|K^{(1,1)}(x,x)\|^{\frac{1}{2}}\nu \tag{6}$$

where $\|K^{(1,1)}(x,x)\|$ is the operator norm. Note that for the Gaussian kernel we can directly compute the r.h.s of Equation (5), obtaining $\Delta^2(\nu) = 2(1-e^{-\beta\nu^2})$ which implies that $\Delta(\nu) \simeq \sqrt{2\beta}\nu$.

Going back to the discussion at the end of Section 2.1, we see that Equations (2) and (3) lead to the same sample complexity bound if $n \leq 2+q$. However, it should be possible to improve this result when K has higher order derivatives. This is left as a future problem.

Acknowledgments: The author would like to thank Steven Smale for valuable suggestions on an earlier version of the paper, and Tomaso Poggio, Enrico Zoli and the referee for useful comments.

REFERENCES

1. N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler, Scale-sensitive dimensions, uniform convergence, and learnability, *Journal of the ACM*, **44**, 4 (1997) 615-631.
2. N. Aronszajn, Theory of reproducing kernels, *Trans. Amer. Math. Soc.*, **686** (1950) 337-404.
3. F. Cucker, and S. Smale, On the mathematical foundations of learning, *Bulletin of the Amer. Math. Soc.*, **39**, 1 (2002) 1-49.
4. L. Devroye and G. Lugosi, "Combinatorial Methods in Density Estimation", Springer Verlag, 2001.
5. T. Evgeniou, M. Pontil, and T. Poggio, Regularization networks and support vector machines, *Adv. in Comp. Math.*, **13** (2000) 1-50.
6. W. Hoeffding, Probability inequality for sums of bounded random variables, *Journal of the Amer. Statist. Assoc.*, **58** (1963) 13–30.
7. A. Kolmogorov, and S. Fomin, "Introductory real analysis", Dover Publications Inc, 1975.
8. S. Smale, and D.X. Zhou, Estimating the approximation error in learning theory, *Anal. Appl.*, **1** (2003) 1-25.
9. V.N. Vapnik, "Estimation of dependencies based on empirical data", Springer Verlag, 1982.