CS281B/Stat241B (Spring 2008) Statistical Learning Theory **Lecture: 19**

Covering Numbers

Lecturer: Peter Bartlett *Scribe: Ma'ayan Bresler*

1 Review: Covering Numbers and Discretization Theorem

In the last lecture we saw the following result:

Theorem 1.1. Discretization Theorem:

$$\hat{R}(f) \leq \inf_{\alpha} \left(\alpha + \sqrt{\frac{2 \log N(\alpha, F, L_2(P_n))}{n}} \right)$$

where the best α can be found by setting the two terms equal, and for last week's example $\alpha = \frac{1}{\sqrt{n}}$. The covering number can be useful when it grows polynomially (but sub-exponentially) for a rich class.

The covering numbers can be used to find an approximation to a rich class. For f constrained to $[-1, 1]$, the points formed by evaluating f at n locations is a subset of the n-cube:

$$\left\{ \begin{pmatrix} f(x_1) \\ \vdots \\ f(x_n) \end{pmatrix} : f \in F \right\} \subseteq [-1, 1]^n \tag{1}$$

The norm we'll use is the distance metric on the n-cube, $||v|| = \sqrt{\frac{1}{n} \sum v_i^2}$.

The covering number N is the smallest covering set such that $\ln(N)$ gives the number of bits needed to specify the function to within α.

If it were $\log_2(N)$ instead, we could think of it as a game of 20 questions, where the approximation to the function improves with the number of questions. In playing 20-questions, the most natural strategy is to start with broad questions and increase their precision, rather than using a constant α. We'll see later in today's lecture that we can get a higher rate doing this.

2 Examples of using the covering number

Example. Consider $F_{|x_1^n} = [-1, 1]^n$. Then $N(\alpha, F, L_2(P_n)) \sim (\frac{1}{\alpha})^n$, so

$$\sqrt{\frac{\log N}{n}} \sim \sqrt{\log(1/\alpha)},$$

which is no good for us, because then $\hat{R}(f) \nrightarrow 0$ as $n \to \infty$.

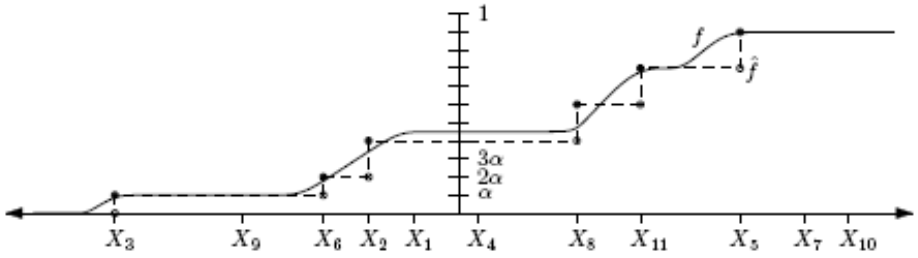

Figure 1: Fig copied from Benjamin Rubinstein's 2006 notes. A non-decreasing function from $\mathbb{R}$ to $[0, 1]$ can be approximated by discretizing the range to $1/\alpha$ levels of width α, and for each such level specifying one of the n points at which it increases above that level.

The covering number can be thought of as asking to what extent the class fills up the n-cube.

An example in which one can do quite a bit better than the discretization theorem is as follows:

Example. $F =$ the non-decreasing function from $\mathbb{R}$ to $[0, 1]$.

We can actually cover such a function uniformly. We only need to approximate it at n points, marked in the figure. If it is within α at each of these points then the L_2 distance will be no more than α. From the approximating points one can produce a non-decreasing function: for each of the α-levels (of which there will be $1/\alpha$), just specify one of the n points at which it increases above that level. From this we can (loosely, but to the right order of magnitude) upper bound the size of the class of estimate functions: $|\hat{F}| \leq n^{1/\alpha}$.

We see that we can cover F in L_2:

$$N(\alpha, F, L_2(P_n)) \leq Cn^{1/\alpha}.$$

As a result of the theorem, we can bound the empirical Rademacher averages by

$$\hat{R}_n(F) \leq \inf_{\alpha}(\alpha + C\sqrt{\frac{\log n}{\alpha n}}) \leq C(\frac{\log n}{n})^{1/3}$$

where the last step is obtained by optimizing over α (just equate the two terms). This bound is loose. We will see later that we can get an exponent of $1/2$ instead of $1/3$.

On the last homework we looked at the relationship between bounded variation and decision stumps. Functions of bounded variation can be represented as the difference of non-decreasing functions, so the above covering can be extended to functions of bounded variation as well.

Now, rather than choosing the level of discretization α and then approximating the function, let's look at choosing increasingly small α, which is called Chaining.

3 Chaining and Dudley's Theorem

The idea here is to look at progressively smaller scales. Rather than choosing an α of a single fixed scale, we choose among increasing refinements.

Theorem 3.1. Dudley:

$$\hat{R}(F) \leq 12 \int_0^{\infty} \frac{logN(\epsilon, F, L_2(P_n))}{n} d\epsilon$$

If the covering numbers get big quickly as ϵ gets small then the right hand side will be infinite. In particular, if $\log N()$ grows faster than $1/\epsilon^2$ the integrand will not integrable. In this case, it turns out that the empirical process no longer obeys the central limit theorem. The Discretization Theorem avoids these problematically small ϵ, and so will give a tighter bound for *very* large classes. But for classes of interest, where $\log N()$ is not growing too rapidly (i.e. less complex function classes), Dudley's theorem will be useful.

PROOF. Fix $B = \sup_{f\in F} ||F||_{L_2(P_n)}$. Set $\alpha_0 = B$, $\alpha_i = 2^{-i}B$. Let T_i be an α_i cover of F in $L_2(P_n)$. Consider $f \in F$. Choose an $\hat{f}_i \in T_i$ such that $||f - \hat{f}_i|| \leq \alpha_i$. Fix $T_0 = \{0\} = \{\hat{f}_0\}$ (for every f).

The Chaining idea is to rewrite f as follows:

$$f = f + \sum_{j=1}^{N} (\hat{f}_j - \hat{f}_{j-1}) + \cancel{\hat{f}_0} - \hat{f}_N.$$

This holds for any N, and we use it in the definition of the rademacher averages to write

$$\begin{aligned}\hat{R}_n(F) = \mathbb{E}\left[\sup_{f\in F} \frac{1}{n}\sum_{i=1}^{N} \epsilon_i f(x_i)\right] &= \mathbb{E}\sup_{f\in F}\left[\frac{1}{n}\sum_{i=1}^{N}\epsilon_i\left(f(x_i) - f_N(x_i)\right) + \frac{1}{n}\sum_{i=1}^{N}\epsilon_i\sum_{j=1}^{N}\left(f_j(x_i) - f_{j-1}(x_i)\right)\right] \\ &\leq \mathbb{E}\sup_{f\in F}\left[\frac{1}{n}\sum_{i=1}^{N}\epsilon_i\left(f(x_i) - f_N(x_i)\right)\right] + \mathbb{E}\sup_{f\in F}\left[\frac{1}{n}\sum_{i=1}^{N}\epsilon_i\sum_{j=1}^{N}\left(f_j(x_i) - f_{j-1}(x_i)\right)\right]\end{aligned}$$

The first term of the right hand side is the inner product $\left\langle \epsilon, f - \hat{f}_N \right\rangle$. This term is small when $\hat{f}_N$ is a good approximation to f: $\left\langle \epsilon, f - \hat{f}_N \right\rangle \leq ||\epsilon||||f - \hat{f}_N|| \leq 1 \cdot \alpha_N$.

The second term on the right hand side is $\sum_{j=1}^{N}\left\langle \epsilon, \hat{f}_j - \hat{f}_{j-1} \right\rangle$, and we observe that $||\hat{f}_j - \hat{f}_{j-1}|| \leq ||\hat{f}_j - f|| + ||f - \hat{f}_{j-1}|| \leq \alpha_j + \alpha_{j-1} = 3\alpha_j$ (where the last step uses the fact that $a_i = 2^{-i}\alpha_0$).

Note that one could nest the T_i's, possibly improving the constant factor.

The Finite Set Lemma implies

$$\mathbb{E}\sup_{f\in F}\frac{1}{n}\sum_{i=1}^{N}\epsilon_i\left(f_j(x_i) - f_{j-1}(x_i)\right) \leq 3\alpha_j\sqrt{\frac{2\log(|T_j||T_{j-1}|)}{n}} \leq 6\alpha_j\sqrt{\frac{\log|T_j|}{n}}$$

where in the last step we've used $|T_{j-1}| \leq |T_j|$.

We put this into our earlier inequality, obtaining

$$\hat{R}_n(F) \leq \alpha_N + 6\sum_{j=1}^{N}\alpha_j\sqrt{\frac{\log|T_j|}{n}} \tag{2}$$

$$\leq \alpha_N + 12\sum_{j=1}^{N}(\alpha_j - \alpha_{j-1})\sqrt{\frac{\log|T_j|}{n}} \tag{3}$$

$$\leq 12\int_0^\infty \sqrt{\frac{\log N(\epsilon, F, L_2(P_n))}{n}}d\epsilon \tag{4}$$

where the second to last step uses the fact that $\alpha_j = 2(\alpha - \alpha_{j-1})$, and the final step is justified in Figure 2. □

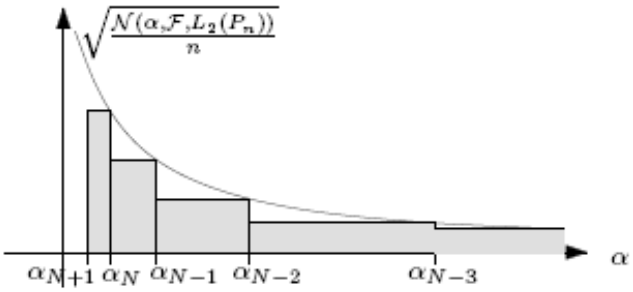

Figure 2: Fig copied from Benjamin Rubinstein's 2006 notes. The justification for the final step in the proof of Dudley's theorem

4 Examples

Let's look at some examples where the bound provided by Dudley's Theorem is better than that offered by the Discretization Theorem, which corresponded to minimization of the curve divided by the identity $y = \alpha$.

Example. F is a subset of a d-dimensional linear space. Then $\log N(\epsilon, F, L_2(P_n)) \sim d \log \frac{1}{\epsilon}$

1. The Discretization Theorem gives
$$\hat{R}_n(F) \leq c\sqrt{\frac{d \log n}{n}}$$
2. The Chaining Theorem gives
$$\hat{R}_n(F) \leq 12\sqrt{\frac{d}{n}} \int_0^1 \sqrt{\log \frac{1}{n}} d\epsilon = 12\sqrt{\frac{\pi}{2}}\sqrt{\frac{d}{n}}$$
where the last step is justified by Fig. 3.

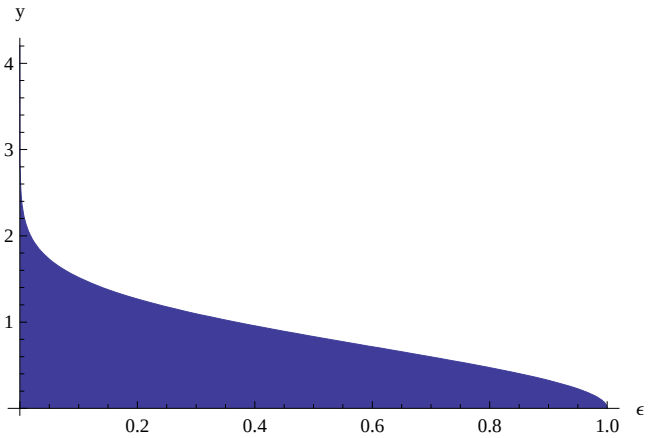

Figure 3: The area beneath $y = \sqrt{\log(1/\epsilon)}$ on $\epsilon \in [0, 1]$ is the same as the area encompassed by the curves $\epsilon = e^{-y^2}$, $y = 0$, and $x = 0$. The latter is simply half a gaussian, so we know its area to be $\sqrt{\pi/2}$.

So the Chaining Theorem allowed the elimination of the log factor to get the correct rate of $\sqrt{1/n}$.

Example. F is the set of non-decreasing functions, $\mathbb{R} \to [0, 1]$. Recall that we can take differences of these functions to obtain the convex hull of decision stumps.

$$N(\epsilon, F, L_2(P_n)) \leq n^{1/\alpha}$$

1. The Discretization Theorem gives
$$\hat{R}_n(F) \leq c\left(\frac{\log n}{n}\right)^{1/3}$$
2. The Chaining Theorem gives
$$\hat{R}_n(F) \leq 12\int_0^1 \sqrt{\frac{\log n}{\alpha n}}d\alpha = 12\sqrt{\frac{\log n}{n}}\int_0^1 \sqrt{\frac{1}{\alpha}}d\alpha = 24\sqrt{\frac{\log n}{n}}$$

Observe that the exponent given by the Chaining Theorem is 1/2 instead of 1/3, which is a dramatic improvement.[1]

5 A Partial Converse to Dudley's Theorem

We can also produce a lower bound for $\hat{R}_n(F)$:

Theorem 5.1. Sudakov's Theorem

$$\hat{R}_n(F) \geq \frac{c}{\log n}\sup_{\alpha>0}\alpha\sqrt{\frac{logN(\epsilon, F, L_2(P_n))}{n}}$$

By ignoring the $\log n$ we can interpret this bound as the biggest rectangle we can fit under the curve $\sqrt{\frac{\log N(\alpha)}{n}}$. The integral of the curve is clearly an upper bound for that area.

We will not prove this theorem (see, for example, Weak Convergence and Empirical Processes: With Applications to Statistics. Aad van der Vaart and Jon Wellner. Springer. 1996).

The gap between the lower bound and upper bound implies that the covering number was not exactly the right approach to take. If interested in uniform convergence then the Rademacher averages are exactly the right approach.

Let's look at the consequence of these results for VC classes.

Example. $F \subset \{\pm 1\}^{\mathcal{X}}$, $d_{VC}(F) = d$.

$$N(\epsilon, F, L_2(P_n)) \leq |F_{|x_1^N}| \leq \Pi_F(n).$$

So

$$\sqrt{\frac{logN(\epsilon, F)}{n}} \leq \sqrt{\frac{d\log(\epsilon n/d)}{n}} \qquad n \geq d$$

[1] Student asks "how much does this matter?" Answer: This is the rate at which the maximal deviation between risk and empirical risk approaches zero. Note that these are not asymptotic results, but true for all n! However, it turns out that the constant is larger, so the better rate is not an improvement until n is sufficiently large.

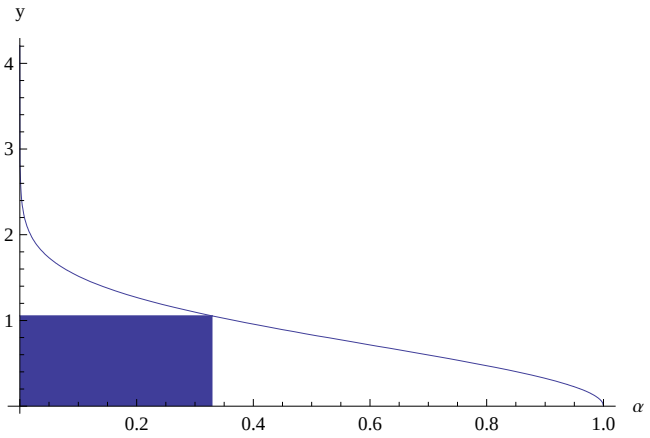

Figure 4: The product $\sup_{\alpha>0} \alpha\sqrt{\frac{logN(\epsilon,F,L_2(P_n))}{n}}$ can be interpreted as the largest rectangle you can fit under the curve $\sqrt{\frac{\log N(\alpha)}{n}}$. This is clearly less than the total area under that curve (which is the upper bound provided by Dudley's theorem).

Theorem 5.2. Haussler

For all $\mathcal{P}$ on $\mathcal{X}$,

$$N(\epsilon, F, L_2(P_n)) \leq \left(\frac{c}{\epsilon}\right)^{2d}$$

Hence

$$\hat{R}_n(F) \leq C\sqrt{\frac{d}{n}}.$$

Note that the log factor is not there.

So, we can do better than just using the growth function. The proof is quite pretty and uses richer properties of VC-classes than we've looked at. See the web site for notes on the proof.

Summary: Covering numbers are convenient for getting bounds.

- If we think about combining classes to get a new class, then it is very easy to get the new covering number, if you know the previous ones.
- There is not a tight relationship with the Radmeacher bounds. The covering number is not quite the right thing to work with in cases of uniform convergence.