CS281B/Stat241B (Spring 2006) Statistical Learning Theory **Lecture: 24**

Covering Numbers, Chaining & Dudley's Entropy Integral

Lecturer: Peter Bartlett *Scribe: Benjamin I. P. Rubinstein*

While continuing the discussion on covering numbers begun last time, we see in this lecture that the covering number bound on Rademacher averages can be quite loose. The main result of the lecture is a more refined chaining result that admits tighter rates on Rademacher averages (by computed covering numbers) for two natural examples.

1 Recap of Last Time

Last time we began studying covering numbers, the size of a minimal cover of a metric subspace.

Definition. Let (M,ρ) be a metric space. A subset $\hat{T} \subseteq M$ is called an ϵ-*cover* of $T \subseteq M$ if $T \subseteq \bigcup_{\hat{t}\in\hat{T}} \mathcal{B}_\epsilon(\hat{t})$, where $\mathcal{B}_\epsilon(c) = \{x \in M \mid \rho(x,c) < \epsilon\}$ denotes the ball of radius $\epsilon > 0$ centered at $c \in M$. An ϵ-cover is *proper* if $\hat{T} \subseteq T$. The ϵ-*covering number* of T is the cardinality of the smallest cover of T,

$$\mathcal{N}(\epsilon, T, \rho) = \min\{|\hat{T}| : \hat{T} \text{ an } \epsilon\text{-cover of T}\} \ .$$

Restricting ourselves to proper covers, we can also estimate ϵ-*proper covering numbers* $\mathcal{N}_{\text{proper}}(\epsilon, T, \rho)$.

Covering numbers arise naturally in empirical process theory, where in order to get useful bounds on empirical Rademacher averages it is important to show that the *metric entropy* $\log \mathcal{N}(\epsilon, \mathcal{F}, L_2(P_n))$ scales well with the sample size[1]. This consideration is demonstrated by the following result, which relates these two quantities.

Theorem 1.1. For $\mathcal{F} \subseteq [-1,1]^{\mathcal{X}}$ and $x_1,\ldots,x_n \in \mathcal{X}$,

$$\hat{R}_n(\mathcal{F}) \quad \leq \quad \inf_{\alpha>0} \sqrt{\frac{2\log\left(\mathcal{N}(\alpha,\mathcal{F},L_2(P_n))\right)}{n}} + \alpha \ .$$

We looked at the case of $\mathcal{F}$ in a d-dimensional linear subspace, where $\mathcal{N}$ grows like $\mathcal{O}\left(\left(\frac{1}{\alpha}\right)^d\right)$. The rate of the metric entropy at scale α has asymptotic *metric dimension* $\frac{\log \mathcal{N}(\alpha,f,l_2(p_n))}{\log \frac{1}{\alpha}}$, which in the linear subspace example agrees with the subspace's dimension of d.

2 An Example with Infinite Metric Dimension

Example. Consider the set of bounded non-decreasing functions[2] $\mathcal{F} = \{f \in [0,1]^{\mathbb{R}} \mid f \text{ non-decreasing}\}$, and how we might cover $(\mathcal{F}, L_2(P_n))$.

[1] P_n is the random empirical measure supported by the random sample, $\left([-1,1]^{\mathcal{X}}, L_2(P_n)\right)$ is the space of bounded functions with random metric (normalized euclidean distance) $\sqrt{\frac{1}{n}\sum_{i=1}^n \left(f(x_i) - g(x_i)\right)^2}$, and $\mathcal{F}$ is the subset we wish to cover.

[2] Note that the set of bounded-variation functions $\{f_1 - f_2 \mid f_1, f_2 \in \mathcal{F}\} = \text{co}(H \cup -H)$ where H is the set of linear threshold functions. So the covering numbers of this latter set is at most twice the covering numbers of $\mathcal{F}$.

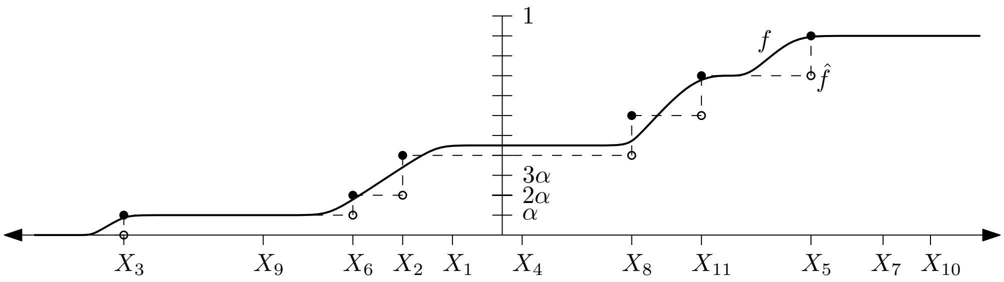

Figure 1: Approximating an $f \in \mathcal{F}$ on $S = (X_1, \ldots, X_n)$ by a step-function $\hat{f} : \mathbb{R} \to \{0, \alpha, 2\alpha, \ldots, 1\}$.

We can approximate any $f \in \mathcal{F}$ by discretizing $[0,1]$ to an α-grid of $\lceil \frac{1}{\alpha} \rceil + 1$ elements, and selecting an appropriate $\hat{f} : \mathbb{R} \to \{0, \alpha, 2\alpha, \ldots, 1\}$ that steps up appropriately to stay within α of f, at each point in $S = (X_1, \ldots, X_n)$, as depicted in Figure 1. This produces an $L_\infty(P_n)$-cover of $\mathcal{F}$: $f(X_i)$ is within α of $\hat{f}(X_i)$ for each i and so equivalently $\|\hat{f} - f\|_{L_\infty(P_n)} = \max_{1 \le i \le n} |\hat{f}(X_i) - f(X_i)| \le \alpha$.

We now bound the α-covering number with respect to the $L_\infty(P_n)$-norm. To specify any approximating $\hat{f}$ we need only specify, for each y in the α-grid, the smallest X_i for which $f(X_i) \ge y$. Thus there are at most $n^{\lceil \frac{1}{\alpha} \rceil + 1}$ functions in the constructed cover. and since $\mathcal{N}(\alpha, \mathcal{F}, L_2(P_n)) \le \mathcal{N}(\alpha, \mathcal{F}, L_\infty(P_n))$ we have

$$\begin{aligned} \mathcal{N}(\alpha, \mathcal{F}, L_2(P_n)) &\le \mathcal{N}(\alpha, \mathcal{F}, L_\infty(P_n)) \\ &\le n^{\lceil \frac{1}{\alpha} \rceil + 1} \\ \log \mathcal{N}(\alpha, \mathcal{F}, L_2(P_n)) &\le \frac{C \log n}{\alpha} \ , \end{aligned}$$

where C is a constant.

Note that this example has no finite metric dimension since $\frac{\log \mathcal{N}(\alpha, \mathcal{F}, L_2(P_n))}{\log \frac{1}{\alpha}}$ grows exponentially. Applying Theorem 1.1 with $\alpha = \left(\frac{\log n}{n}\right)^{1/3}$ leads to[3],

$$\begin{aligned} \hat{R}_n(\mathcal{F}) &\le \inf_{\alpha > 0} \sqrt{\frac{2C \log n}{n\alpha}} + \alpha \\ &\le C \left(\frac{\log n}{n}\right)^{1/3} \ . \end{aligned}$$

This analysis based on Theorem 1.1 is in fact quite loose – the rate is too slow. The main idea of the theorem has been to focus on a single scale α; the contribution of functions below α are bounded by α while those above are counted by the α-covering number. In the next section we turn to a clever improvement to this analysis where we look at all scales simultaneously.

3 Chaining and Dudley's Entropy Integral

Theorem 3.1 (Dudley). For a function class $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{X}}$ and points $x_1, \ldots, x_n \in \mathcal{X}$,

$$\hat{R}_n(\mathcal{F}) \le 12 \int_0^\infty \sqrt{\frac{\log \mathcal{N}(\epsilon, \mathcal{F}, L_2(P_n))}{n}} d\epsilon$$

[3] Throughout these notes we absorb constants into C – i.e. the value of C changes from one line to the next.

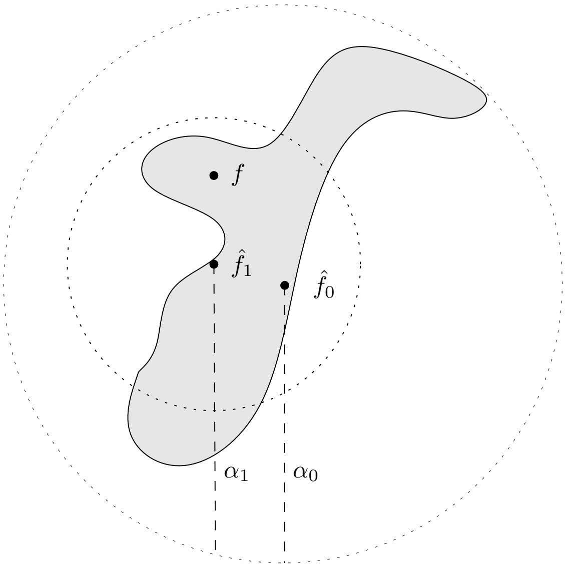

Figure 2: From the chaining proof of Theorem 3.1. For a diminishing sequence of scales α_i, $\hat{f}_i$ α_i-approximates f. If $\alpha_0 \geq 1$ then the choice of $\hat{f}_0 = 0$ satisfies this arrangement. Note that we claim each T_i is a proper cover; of course if 0 is not in $\mathcal{F}$ then $T_0 = \{\hat{f}_0\}$ as stated is not proper. However as we have seen, proper/improper covering numbers are essentially the same up to constants.

Before proceeding to prove the theorem, we use the result on our example to see what improvement it affords us over Theorem 1.1.

Example. Consider the class $\mathcal{F}$ of non-decreasing functions in $[0,1]^{\mathbb{R}}$. Applying Theorem 3.1,

$$\begin{aligned}\hat{R}_n(\mathcal{F}) &\leq C\int_0^\infty \sqrt{\frac{\log n}{\epsilon n}} d\epsilon \\ &= C\sqrt{\frac{\log n}{n}} \left[2\sqrt{\epsilon}\right]_0^1 \\ &= C\sqrt{\frac{\log n}{n}} .\end{aligned}$$

This rate is much closer to parametric and is a considerable improvement.

Proof of Theorem 3.1. Set $B = \sup_{f\in\mathcal{F}} \|f\|_{L_2(P_n)}$ and assume that this quantity is finite as otherwise the result claims a bound of ∞ and is vacuous. For $i \geq 0$ choose $\alpha_i = B2^{-i}$ and let T_i be a (proper) α_i-cover of $\mathcal{F}$ with respect to the $L_2(P_n)$-norm. For each $f \in \mathcal{F}$ and i pick an α_i-approximation $\hat{f}_i \in T_i$, e.g. such that $\|f - \hat{f}_i\|_{L_2(P_n)} \leq \alpha_i$ (see Figure 3).

Express f by *chaining* $f = f - \hat{f}_N + \sum_{j=1}^N \left(\hat{f}_j - \hat{f}_{j-1}\right)$, where $\hat{f}_0 = 0$. Note that taking the initial $\alpha_0 = B$

sufficiently large (e.g. at least 1) guarantees that $\hat{f}_0 = 0$ is an α_0-approximation of f. Now for any N,

$$\begin{aligned}
\hat{R}_n(\mathcal{F}) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \left(f(X_i) - \hat{f}_N(X_i) + \sum_{j=1}^N \left(\hat{f}_j(X_i) - \hat{f}_{j-1}(X_i)\right)\right)\right] \\
&\le \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \left(f(X_i) - \hat{f}_N(X_i)\right)\right] + \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{j=1}^N\sum_{i=1}^n \epsilon_i \left(\hat{f}_j(X_i) - \hat{f}_{j-1}(X_i)\right)\right] \\
&\le \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \left(f(X_i) - \hat{f}_N(X_i)\right)\right] + \sum_{j=1}^N \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \left(\hat{f}_j(X_i) - \hat{f}_{j-1}(X_i)\right)\right] \quad . \qquad (1)
\end{aligned}$$

For the first term of (1) we can appeal to the Cauchy-Schwartz inequality and the identity $\|\epsilon\|_{L_2(P_n)} = 1$,

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \left(f(X_i) - \hat{f}_N(X_i)\right)\right] &= \mathbb{E}\left[\sup_{f\in\mathcal{F}} \langle \epsilon, f - \hat{f}_N\rangle_{L_2(P_n)}\right] \\
&\le \mathbb{E}\left[\sup_{f\in\mathcal{F}} \|\epsilon\|_{L_2(P_n)} \cdot \left\|f - \hat{f}_N\right\|_{L_2(P_n)}\right] \\
&\le \mathbb{E}\left[\sup_{f\in\mathcal{F}} 1 \cdot \alpha_N\right] \\
&= \alpha_N \quad . \qquad (2)
\end{aligned}$$

And for the second term of (1), noting that

$$\begin{aligned}
\left\|\hat{f}_j - \hat{f}_{j-1}\right\|_{L_2(P_n)}^2 &= \left\|\hat{f}_j - f + f - \hat{f}_{j-1}\right\|^2 \\
&\le \left(\left\|\hat{f}_j - f\right\| + \left\|f - \hat{f}_{j-1}\right\|\right)^2 \\
&\le \left(\alpha_j + \alpha_{j-1}\right)^2 \\
&= \left(\alpha_j + 2\alpha_j\right)^2 \\
&= 9\alpha_j^2 \ ,
\end{aligned}$$

we can apply Massart's Finite Lemma[4]

$$\begin{aligned}
\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i \left(\hat{f}_j(X_i) - \hat{f}_{j-1}(X_i)\right)\right] &\le 3\alpha_j \sqrt{\frac{2\log(|T_j||T_{j-1}|)}{n}} \\
&\le 6\alpha_j \sqrt{\frac{\log|T_j|}{n}} \quad . \qquad (3)
\end{aligned}$$

[4]That states: if $|\mathcal{F}|$ is finite and $\sup_{f\in\mathcal{F}} \|f\|_{L_2(P_n)}^2 \le R^2$ then $\mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n \epsilon_i f(X_i)\right] \le \sqrt{\frac{2R^2\log|\mathcal{F}|}{n}}$.

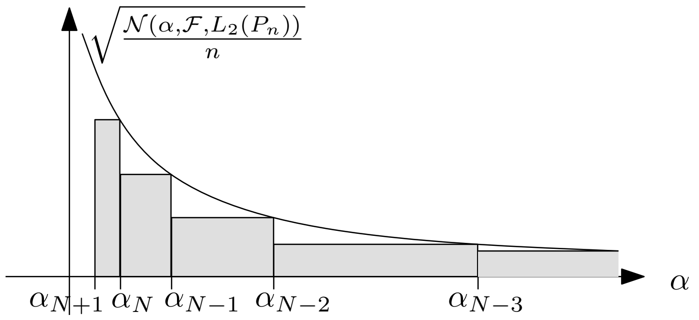

Figure 3: From the chaining proof of Theorem 3.1. Noting that $\mathcal{N}(\alpha, \mathcal{F}, L_2(P_n))$ is a non-increasing function of α and that α_i is a decreasing sequence, it is easy to see that $\sum_{j=1}^{N} (\alpha_j - \alpha_{j+1}) \sqrt{\frac{\log \mathcal{N}(\alpha_j, \mathcal{F}, L_2(P_n))}{n}} \le \int_{\alpha_{N+1}}^{\alpha_0} \sqrt{\frac{\log \mathcal{N}(\alpha, \mathcal{F}, L_2(P_n))}{n}}$.

And so by (1)–(3), choosing minimal α_j-covers of $\mathcal{F}$ for each T_j,

$$
\begin{aligned}
\hat{R}_n(\mathcal{F}) &\le \alpha_N + 6 \sum_{j=1}^{N} \alpha_j \sqrt{\frac{\log \mathcal{N}(\alpha_j, \mathcal{F}, L_2(P_n))}{n}} \\
&= \alpha_N + 12 \sum_{j=1}^{N} (\alpha_j - \alpha_{j+1}) \sqrt{\frac{\log \mathcal{N}(\alpha_j, \mathcal{F}, L_2(P_n))}{n}} \\
&\le \alpha_N + 12 \int_{\alpha_{N+1}}^{\alpha_0} \sqrt{\frac{\log \mathcal{N}(\alpha, \mathcal{F}, L_2(P_n))}{n}} d\alpha \\
&\le \alpha_N + 12 \int_{\alpha_{N+1}}^{\infty} \sqrt{\frac{\log \mathcal{N}(\alpha, \mathcal{F}, L_2(P_n))}{n}} d\alpha \ , \qquad (4)
\end{aligned}
$$

where the second inequality replaces the sum by an integral via the argument summarized in Figure 3. Taking $N \to \infty$, (4) implies $\hat{R}_n(\mathcal{F}) \le 12 \int_0^\infty \sqrt{\frac{\log \mathcal{N}(\alpha, \mathcal{F}, L_2(P_n))}{n}} d\alpha$. □

We have already seen that this refined result provides a faster rate for the class of non-decreasing functions. Another example demonstrating an improvement is one we have seen earlier.

Example. Suppose that $\mathcal{F} \subseteq [-1, 1]^{\mathcal{X}}$ is a finite-dimensional class (or that it lies in one). Then

$$
\begin{aligned}
\log \mathcal{N}(\alpha, \mathcal{F}) &\le d \log \frac{1}{\epsilon} \\
\hat{R}_n(\mathcal{F}) &\le 12 \sqrt{\frac{d}{n}} \int_0^1 \sqrt{\log \frac{1}{\epsilon}} \ d\epsilon \\
&= 12 \sqrt{\frac{d}{n}} \int_{-\infty}^{0} e^{-y^2} \ dy \\
&= 12 \sqrt{\frac{\pi}{2}} \sqrt{\frac{d}{n}} \ ,
\end{aligned}
$$

as was claimed several lectures ago. This improves on the $C\sqrt{\frac{d \log n}{n}}$ achieved with Theorem 1.1.

3.1 A Partial Converse to Dudley's Theorem

Theorem 3.2 (Sudakov). Under the same conditions as Theorem 3.1,

$$\hat{R}_n(\mathcal{F}) \quad \geq \quad \frac{C}{\log n} \sup_{\alpha > 0} \alpha \sqrt{\frac{\log \mathcal{N}(\alpha, \mathcal{F}, L_2(P_n))}{n}} \ .$$

This result is analogous to finding the largest rectangle (by area) under the $\left(\sqrt{\frac{\mathcal{N}(\alpha, \mathcal{F}, L_2(P_n))}{n}}\right)$-curve in Figure 3. It turns out that $\hat{R}_n$ and $\mathcal{N}$ are not the same quantities and that this is the best lower-bound in general; the gap between it and Dudley's upper-bound is necessary.

4 Next Time

We shall see that a tight VC-dimension bound on covering numbers provides a useful bound on Rademacher averages. For $\mathbb{R}$-valued functions there exist related combinatorial dimensions with relations to covering numbers.