Stat 928: Statistical Learning Theory **Lecture: 17**

Dudley's Theorem and Packing Numbers

Instructor: Sham Kakade

1 Chaining and Dudley's Theorem

Rather than choosing a fixed scale, one can integrate over all scales, as shown in the following theorem.

Theorem 1.1. *Assume that all $\mathcal{F}_{x_{1:n}} \subset \mathbb{R}^n$. Let $\hat{\mathfrak{R}}_n(\mathcal{F})$ be the empirical Rademacher number of $\mathcal{F}$ on $x_{1:n}$. We have:*

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left(4\alpha + 12 \int_{\alpha}^{\infty} \sqrt{\frac{\log N_2(\epsilon, \mathcal{F}, x_{1:n})}{n}} d\epsilon \right)$$

This theorem is subtly different from the discretization theorem. It is stated in terms of the 2-norm covering number rather than the 1-norm covering number. Also note that we do no not restrict the range of $f \in \mathcal{F}$.

Proof. Abusing notation we assume that $\mathcal{F} = \mathcal{F}_{x_{1:n}}$. Let

$$B = \sup_{f \in \mathcal{F}} \sqrt{\frac{1}{n} \sum_{i=1}^{n} f_i^2}$$

For $j \geq 0$, set $\alpha_j = 2^{-j} B_j$ and let T_j be a minimal α_j-cover of $\mathcal{F}$ with respect to the 2-norm. Let $c_j(f)$ be the element in T_j which covers f. We also use the notation $[c_j(f)]_i$ to denote the $i-th$ component. Assume that $T_0 = \{0\}$ (which is a B-cover).

The *chaining* method expresses f as $f = f - c_N(f) + \sum_{j=1}^{N} (c_j(f) - c_{j-1}(f))$, since $c_0(f) = 0$. Now we have:

$$\begin{aligned}
\hat{\mathfrak{R}}_n(\mathcal{F}) &= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i f_i \right) \right] \\
&= \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f_i - [c_N(f)]_i + \sum_{j=1}^{N} ([c_j(f)]_i - [c_{j-1}(f)]_i)) \right) \right] \\
&\leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f_i - [c_N(f)]_i) \right) \right] + \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N} \epsilon_i ([c_j(f)]_i - [c_{j-1}(f)]_i) \right) \right] \\
&\leq \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f_i - [c_N(f)]_i) \right) \right] + \sum_{j=1}^{N} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i ([c_j(f)]_i - [c_{j-1}(f)]_i) \right) \right] \\
&\leq \alpha_N + \sum_{j=1}^{N} \mathbb{E}\left[\sup_{f \in \mathcal{F}} \left(\frac{1}{n} \sum_{i=1}^{n} \epsilon_i ([c_j(f)]_i - [c_{j-1}(f)]_i) \right) \right]
\end{aligned}$$

where we have appealed to Cauchy-Schwarz in the last step.

Appealing to Massart's lemma:

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i([c_j(f)]_i-[c_{j-1}(f)]_i)\right)\right] \leq \frac{(\sup_{f\in\mathcal{F}}||c_j(f)-c_{j-1}(f)||_2)\sqrt{2\log|T_j||T_{j-1}|}}{n}$$

And by the triangle inequality:

$$\begin{aligned}\sup_{f\in\mathcal{F}}||c_j(f)-c_{j-1}(f)||_2 &\leq \sup_{f\in\mathcal{F}}||c_j(f)-f||_2+\sup_{f\in\mathcal{F}}||f-c_{j-1}(f)||_2\\ &\leq \sqrt{n}\alpha_j+\sqrt{n}\alpha_{j-1}\\ &= \sqrt{n}\alpha_j+2\sqrt{n}\alpha_j\\ &= 3\sqrt{n}\alpha_j\end{aligned}$$

where $\sqrt{n}$ factor comes from the fact that the 2-norm in the covering number definition is a normalized quantity. So we have

$$\begin{aligned}\mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^{n}\epsilon_i([c_j(f)]_i-[c_{j-1}(f)]_i)\right)\right] &\leq \frac{3\alpha_j\sqrt{2\log|T_j||T_{j-1}|}}{\sqrt{n}}\\ &\leq \frac{6\alpha_j\sqrt{\log|T_j|}}{\sqrt{n}}\end{aligned}$$

since $|T_j|\geq|T_{j-1}|$.

Continuing, and using that $2(\alpha_j-\alpha_{j+1})=\alpha_j$

$$\begin{aligned}\hat{\mathfrak{R}}_n(\mathcal{F}) &\leq \alpha_N+6\sum_{j=1}^{N}\alpha_j\sqrt{\frac{\log|T_j|}{n}}\\ &= \alpha_N+6\sum_{j=1}^{N}\alpha_j\sqrt{\frac{\log N_2(\alpha_j,\mathcal{F},x_{1:n})}{n}}\\ &\leq \alpha_N+12\sum_{j=1}^{N}(\alpha_j-\alpha_{j+1})\sqrt{\frac{\log N_2(\alpha_j,\mathcal{F},x_{1:n})}{n}}\\ &\leq \alpha_N+12\int_{\alpha_{N+1}}^{\alpha_0}\sqrt{\frac{\log N_2(\alpha,\mathcal{F},x_{1:n})}{n}}d\alpha\\ &\leq \alpha_N+12\int_{\alpha_{N+1}}^{\infty}\sqrt{\frac{\log N_2(\alpha,\mathcal{F},x_{1:n})}{n}}d\alpha\end{aligned}$$

Pick any $\alpha\geq 0$. Let N be the largest integer j such that $\alpha_j\geq 2\alpha$. Hence, $\alpha_N\geq 2\alpha$ and $\alpha_{N+1}=\alpha_N/2\geq\alpha$, so the integral from α_{N+1} to ∞ is upper bounded by the integral from α to ∞. Also, since $\alpha_{N+1}\leq 2\alpha$, we have that $\alpha_N\leq 4\alpha$, which completes the proof. □

2 Examples

2.1 Vector Spaces

Assume that $F|_{x_{1:n}}$ is finite dimensional vector space of dimension k. Then one can show $N_2(\alpha, \mathcal{F}, x_{1:n}) \leq (1/\alpha)^k$. Then using the discretization bound, with $\alpha = \sqrt{1/n}$, we have:

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq 2\sqrt{\frac{k \log n}{n}}$$

From Dudley's Thm, we get the sharper bound that:

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq 12\sqrt{\frac{k}{n}} \int_0^1 \sqrt{\log \frac{1}{\alpha}} d\alpha \leq 12\sqrt{\frac{\pi}{2}}\sqrt{\frac{k}{n}}$$

Here, note that the covering number is 1 when $\alpha > 1$. Also, the integral can be evaluated using the change of variables $y = e^{-y^2}$ (which ends up looking like a Gaussian).

2.2 Increasing Functions

Now assume that $F|_{x_{1:n}}$ is the set of non-decreasing functions. Considering discretizing $\mathcal{Y} = [-1, 1]$ into bins of size $1/\alpha$. We can approximate a function in $f \in \mathcal{F}$ by: for each of the bins (there are $1/\alpha$ bins), just specify one of the n points at which the function increases above that bin. This means the covering number is at most $n^{1/\alpha}$.

Hence, from the discretization theorem, we get;

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq \text{constant} \left(\frac{\log n}{n}\right)^{1/3}$$

(by optimizing α). From Dudley's theorem, we get the sharper result:

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq \text{constant}\sqrt{\frac{\log n}{n}}$$

The improved exponent is actually quite important.

3 Combinatorial Dimensions for Real Valued Function Classes

We will now define some combinatorial dimensions for real valued function classes. We will then see how to get upper bounds for covering numbers in terms of these dimensions. The first dimension is called Pollard's pseudodimension.

Definition 3.1. *Let $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$. We say that $\mathcal{F}$ P-shatters a set $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ if there exist $s_1, \ldots, s_n$ such that, for all $E \subseteq X$, there exists $f_E \in \mathcal{F}$ such that,*

$$\begin{aligned} &\forall x_i \in E, && f_E(x_i) \geq s_i \\ &\forall x_i \in X - E, && f_E(x_i) < s_i \end{aligned}$$

The pseudodimension of $\mathcal{F}$, denoted by $\mathrm{Pdim}(\mathcal{F})$, is the size of a largest P-shattered set.

The pseudodimension does not take any scale into account. The fat shattering dimensions measures the complexity of the class $\mathcal{F}$ at some scale α.

Definition 3.2. *Let $\mathcal{F} \subseteq [0,1]^{\mathcal{X}}$. We say that $\mathcal{F}$ P_α-shatters a set $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ if there exist $s_1, \ldots, s_n$ such that, for all $E \subseteq X$, there exists $f_E \in \mathcal{F}$ such that,*

$$\begin{aligned} &\forall x_i \in E, && f_E(x_i) \geq s_i + \alpha \\ &\forall x_i \in X - E, && f_E(x_i) \leq s_i - \alpha \end{aligned}$$

The fat shattering dimension of $\mathcal{F}$ at scale α, denoted by $\mathrm{fat}_\alpha(\mathcal{F})$, is the size of a largest P_α-shattered set.

4 Covering and Packing Numbers

We have already seen the definition of covering numbers before. Packing numbers are closely related to covering numbers. Depending on the situation, we might prefer to work with one or the other. Note that both these quantities can be defined in a general (pseudo)metric space.

Let $(\mathcal{X}, d)$ be a pseudometric space. Let $A \subseteq \mathcal{X}$ and $\alpha > 0$. We say that $B \subseteq A$ is an α-cover of A iff $\forall a \in A$, $\exists b \in B$ such that $d(a,b) < \alpha$. Now we define the covering number as

$$\mathcal{N}_d(\alpha, A) := \min \{|B| \,|\, B \text{ is an } \alpha\text{-cover of } A\} \ .$$

Let $A \subseteq \mathcal{X}$ and $\alpha > 0$. We say that A is α-separated if $\forall a, b \in A, a \neq b$, $d(a,b) \geq \alpha$. Define the packing number as,

$$\mathcal{M}_d(\alpha, A) := \max \{|A'| \,|\, A' \text{ is } \alpha\text{-separated}, \ A' \subseteq A\} \ .$$

As we said, these two numbers are closely related.

Lemma 4.1. *For any pseudometric space $(\mathcal{X}, d)$ and any $A \subseteq \mathcal{X}$ and $\alpha > 0$,*

$$\mathcal{M}_d(2\alpha, A) \leq \mathcal{N}_d(\alpha, A) \leq \mathcal{M}_d(\alpha, A) \ .$$

Proof. If M is a 2α-separated subset of A and N is an α-cover of A then N must select a point within α distance of each of the points in M. These points will necessarily be distinct since points in M are at least 2α apart. Thus $|M| \leq |N|$. This proves the first half of the lemma.

If M is a *maximal* α-separated subset of A then M has to be an α-cover. Because if it is not, then there is a point $x \in A$ such there is no point of M within a distance of α from x. In that case, x can be added to M while still keeping it α-separated. This violates the maximality of M. Thus, $\mathcal{N}_d(\alpha, A) \leq |M| = \mathcal{M}_d(\alpha, A)$. This proves the second half of the lemma. □