Stat 928: Statistical Learning Theory **Lecture: 16**

Uniform and Empirical Covering Numbers

Instructor: Sham Kakade

1 Warmup

Assume that for every $\alpha > 0$ that we have a (finite) set $\hat{\mathcal{F}}_\alpha$ such that for all $f \in \mathcal{F}$ there exists an $\hat{f} \in \hat{\mathcal{F}}_\alpha$ such that $x \in \mathcal{X}, y \in \mathcal{Y}$:

$$|\phi(\hat{f}(x), y) - \phi(f(x), y)| \leq \alpha \quad .$$

Such an $\hat{\mathcal{F}}_\alpha$ is a α-cover of $\mathcal{F}$. Clearly, this implies that:

$$|\mathcal{L}(\hat{f}(x)) - \mathcal{L}(f(x))| \leq \alpha \quad .$$

Hence, we can view $\hat{\mathcal{F}}_\alpha$ as implicitly providing a cover for the loss class.

Intuitively, with respect to obtaining a uniform convergence rate, we could work directly with $\hat{\mathcal{F}}_\alpha$. More precisely,

Theorem 1.1. *Assume that for all $f \in \mathcal{F}$ our predictions are in $[-1, 1]$. With probability greater than $1 - \delta$*

$$\sup_{f \in \mathcal{F}} |\hat{\mathcal{L}}(f) - \mathcal{L}(f)| \leq \inf_\alpha 2\sqrt{\frac{\log|\hat{\mathcal{F}}_\alpha| + \log\frac{1}{\delta}}{2n}} + 2\alpha$$

Proof. Fix α. Using the union bound, we have:

$$\sup_{\hat{f} \in \hat{\mathcal{F}}_\alpha} |\hat{\mathcal{L}}(\hat{f}) - \mathcal{L}(\hat{f})| \leq 2\sqrt{\frac{\log|\hat{\mathcal{F}}_\alpha| + \log\frac{1}{\delta}}{2n}}$$

Let $c(f)$ be the function $\hat{\mathcal{F}}_\alpha$ which covers f. Following from the definition of $c(f)$ and $\hat{\mathcal{F}}_\alpha$, we have that for all $f \in \mathcal{F}$,:

$$|\mathcal{L}(f) - \mathcal{L}(c(f))| \leq \alpha$$
$$|\hat{\mathcal{L}}(f) - \hat{\mathcal{L}}(c(f))| \leq \alpha$$

It follows that:

$$\begin{aligned}
\sup_{f \in \mathcal{F}} |\hat{\mathcal{L}}(f) - \mathcal{L}(f)| &= \sup_{f \in \mathcal{F}} |\hat{\mathcal{L}}(f) - \hat{\mathcal{L}}(c(f)) - (\mathcal{L}(f) - \mathcal{L}(c(f))) + \hat{\mathcal{L}}(c(f)) - \mathcal{L}(c(f))| \\
&\leq 2\alpha + \sup_{f \in \mathcal{F}} |\hat{\mathcal{L}}(c(f)) - \mathcal{L}(c(f))| \\
&\leq 2\alpha + \sup_{\hat{f} \in \hat{\mathcal{F}}_\alpha} |\hat{\mathcal{L}}(\hat{f}) - \mathcal{L}(\hat{f})| \\
&\leq 2\alpha + \sqrt{\frac{\log|\hat{\mathcal{F}}_\alpha| + 2\log\frac{1}{\delta}}{2n}}
\end{aligned}$$

The proof is completed by noting that α is arbitrary, so we can take a $\inf$ over α. $\square$

2 General covering numbers

Consider function class $G = \{g_\theta(Z) : \theta \in \Theta\}$. Given any metric $d(g, g')$, an ϵ cover of G in metric d is a set $G_d(\epsilon) = \{g_1(Z), \ldots, g_N(Z)\}$ such that for all $g_\theta \in G$, there exists j: $d(g_\theta, g_j) \le \epsilon$.

An example is least squares sub-Gaussian analysis, where $g_\theta(\xi) = \theta^\top P_X \xi$, and the covering is with respect to the Euclidean distance in the parameter space $\Theta = S^{d-1}$.

We are particularly interested in distances with respect to the true or empirical underlying distribution of Z. Let D be a distribution over Z, then we can define L_p distance between two functions $g(z)$ and $g'(z)$ as $d_D^p(g, g') = [E_D|g(z) - g'(z)|^p]^{1/p}$. We know that $d_P^p(g, g')$ increases as p increases (property of L_p distance).

Now $G_p(\epsilon) = \{g_1(Z), \ldots, g_N(Z)\}$ is an L_p cover of G with respect to D if for all $g_\theta \in G$, there exists j such that

$$[E_{Z\sim D}|g_j(Z) - g_\theta(Z)|^p]^{1/p} \le \epsilon.$$

Moreover, consider an empirical distribution $Z_1^n = \{Z_1, \ldots, Z_n\}$ over Z, then we may define empirical L_p cover of G as L_p cover of G with respect to the empirical p-norm:

$$[n^{-1}\sum_{i=1}^{n}|g(Z_i) - g'(Z_i)|^p]^{1/p}.$$

The smallest number of ϵ-cover, is called ϵ-covering number, and the log of covering number is called ϵ-entropy. Uniform (empirical) L_p entropy is the maximum L_p entropy of G under the worst case empirical distribution. Since L_p distance increases, therefore L_p entropy increases when p increases. However, the most interesting cases are $p \ge 2$, specially $p = 2$ and $p = \infty$.

Relation to bracketing cover: L_∞ cover is stronger than Bracketing cover. This is because if $\{g_j\}$ is an ϵ cover of g_θ, then $g_j^L = g_j - \epsilon$ is 2ϵ lower and g_j^U is 2ϵ upper bracketing cover. g_j^L and g_j^U is 2ϵ bracketing cover. The reverse is not necessarily true. For example, the classification example has finite bracketing cover but does not have finite L_∞ cover. Because of the relationship, the analysis of bracketing cover can be used with L_∞ cover. However, some times empirical L_∞ cover is useful and one does not necessarily have a bracketing cover counterpart.

3 p-norm Covering Numbers

The problem with the previous notion of a cover is that it *uniformly* demands a good approximation to each f by an element in $\hat{\mathcal{F}}_\alpha$. Intuitively, it seems more natural to have a cover such that for each $f \in \mathcal{F}$ there is an element in the cover which is only on average close f. We now formalize this.

Assume that all hypotheses in our class $\mathcal{F}$ make real valued predictions. Let $x_{1:n}$ be a set of n points. A set of vectors $V \subset \mathbb{R}^n$ is an α-*cover*, with respect to the p-norm, of $\mathcal{F}$ on $x_{1:n}$ if for all $f \in \mathcal{F}$ there exists a $v \in V$ such that:

$$\left(\frac{1}{n}\sum_{i=1}^{n}|v_i - f(x_i)|^p\right)^{\frac{1}{p}} \le \alpha$$

We define the *p-norm covering number* $\mathcal{N}_p(\alpha, \mathcal{F}, x_{1:n})$ as the size of the minimal such cover V, i.e.:

$$\mathcal{N}_p(\alpha, \mathcal{F}, x_{1:n}) = \min\{|V| : V \text{ is an } \alpha\text{-cover, under the p-norm, of F on } x_{1:n}\}$$

Also define:

$$\mathcal{N}_p(\alpha, \mathcal{F}, n) = \sup_{x_{1:n}} \mathcal{N}_p(\alpha, \mathcal{F}, x_{1:n}) \quad .$$

In other words, $\mathcal{N}_p(\alpha, \mathcal{F}, n)$ is the worst case covering number over $x_{1:n}$.

Observe that:

$$\mathcal{N}_p(\alpha, \mathcal{F}, \infty) \leq \mathcal{N}_q(\alpha, \mathcal{F}, \infty)$$

for $p \leq q$. This is consequence of using the (normalized) p-norm in the definition of the covering number.

Note that:

$$\mathcal{N}_\infty(\alpha, \mathcal{F}, \infty) \leq |\hat{\mathcal{F}}_\alpha|$$

which follows directly from the definition of $\hat{\mathcal{F}}_\alpha$.

4 Rademacher Bounds

Theorem 4.1. *(Discretization) Assume that all $f \in \mathcal{F}$ make predictions in $[-1,1]$. Let $\hat{\mathfrak{R}}_n(\mathcal{F})$ be the empirical Rademacher number of $\mathcal{F}$ on $x_{1:n}$. We have:*

$$\hat{\mathfrak{R}}_n(\mathcal{F}) \leq \inf_\alpha \sqrt{\frac{2\log N_1(\alpha, \mathcal{F}, x_{1:n})}{n}} + \alpha$$

Proof. Fix α and fix a minimal cover V. Define $B_\alpha(v)$ to be the hypothesis in $\mathcal{F}$ that are α-covered by v. Using that $\cup_{v\in V} B_\alpha(v) = \mathcal{F}$,

$$\begin{aligned}
\hat{\mathfrak{R}}_n(\mathcal{F}) &= \mathbb{E}\left[\sup_{f\in\mathcal{F}}\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right)\right] \\
&= \mathbb{E}\left[\sup_{v\in V}\sup_{f\in B_\alpha(v)}\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)\right)\right] \\
&= \mathbb{E}\left[\sup_{v\in V}\sup_{f\in B_\alpha(v)}\left(\frac{1}{n}\sum_{i=1}^n \epsilon_i v_i + \frac{1}{n}\sum_{i=1}^n \epsilon_i (f(x_i) - v_i)\right)\right] \\
&\leq \mathbb{E}\left[\sup_{v\in V}\frac{1}{n}\sum_{i=1}^n \epsilon_i v_i\right] + \mathbb{E}\left[\sup_{v\in V}\sup_{f\in B_\alpha(v)}\frac{1}{n}\sum_{i=1}^n \epsilon_i (f(x_i) - v_i)\right]
\end{aligned}$$

Using Holder's inequality for the second term,

$$\begin{aligned}
\mathbb{E}\left[\sup_{v\in V}\sup_{f\in B_\alpha(v)}\frac{1}{n}\sum_{i=1}^n \epsilon_i (f(x_i) - v_i)\right] &\leq \mathbb{E}\left[\sup_{v\in V}\sup_{f\in B_\alpha(v)}\frac{1}{n}\sum_{i=1}^n |f(x_i) - v_i|\right] \\
&\leq \alpha
\end{aligned}$$

Using Massart's finite lemma for the first term:

$$\begin{aligned}
\mathbb{E}\left[\sup_{v\in V}\frac{1}{n}\sum_{i=1}^n \epsilon_i v_i\right] &\leq \frac{\sup_{v\in V}||v||_2\sqrt{2\log|V|}}{n} \\
&\leq \sqrt{\frac{2\log|V|}{n}} \\
&= \sqrt{\frac{2\log N_1(\alpha, \mathcal{F}, x_{1:n})}{n}}
\end{aligned}$$

The proof is completed by combining these last two bounds and noting that α was arbitrary (so we can take an inf over all $\alpha > 0$). □

The following is immediate:

Corollary 4.2. *Assume that all $f \in \mathcal{F}$ make predictions in $[-1, 1]$. We have:*

$$\mathfrak{R}_n(\mathcal{F}) \leq \inf_{\alpha} \sqrt{\frac{2 \log N_1(\alpha, \mathcal{F}, n)}{n}} + \alpha$$