Machine Learning Theory (CS 6783)

Lecture 9: Covering Numbers, Pollard and Dudley Bounds

1 Recap

1. For any statistical learning problem we have,

\[ E_S \left[ L_D(\hat{y}_{\text{erm}}) - \inf_{f \in \mathcal{F}} L_D(f) \right] \leq \frac{2}{n} E_S E_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t) \right] = 2 E_S \left[ \hat{R}_S(\ell \circ \mathcal{F}) \right] \]

2. For any $L$-Lipschitz loss

\[ \frac{1}{n} E_S E_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \ell(f(x_t), y_t) \right] \leq \frac{L}{n} E_S E_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \]

\[ E_S \left[ \hat{R}_S(\ell) \right] \leq L E_S \left[ \hat{R}_S(\mathcal{F}) \right] \]

2 Covering Number

Conditioned on $x_1, \ldots, x_n$, we are interested in bounding :

\[ \frac{1}{n} E_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] = \frac{1}{n} E_\epsilon \left[ \sup_{v \in \mathcal{F}_{|x_1, \ldots, x_n}} \sum_{t=1}^{n} \epsilon_t v[t] \right] \]

Recall the projection of $\mathcal{F}$ on sample :

\[ \mathcal{F}_{|x_1, \ldots, x_n} = \{(f(x_1), \ldots, f(x_n)) \in \mathbb{R}^d : f \in \mathcal{F}\} \]

For real valued functions of course $|\mathcal{F}_{|x_1, \ldots, x_n}|$ could very well be infinite. But now given the $n$ data points, we can ask how large a set do we need to discretize $\mathcal{F}_{|x_1, \ldots, x_n}$ to accuracy $\beta$.

**Definition 1.** $V \subset \mathbb{R}^n$ is an $\ell_p$ cover of $\mathcal{F}$ on $x_1, \ldots, x_n$ at scale $\beta > 0$ if for all $f \in \mathcal{F}$, there exists $v_f \in V$ such that

\[ \left( \frac{1}{n} \sum_{t=1}^{n} |f(x_t) - v_f[t]|^p \right)^{1/p} \leq \beta \]

**Empirical covering number**

\[ N_p(\mathcal{F}, \beta; x_1, \ldots, x_n) = \min \{|V| : V \text{ is an } \ell_p \text{ cover of } \mathcal{F} \text{ on } x_1, \ldots, x_n \text{ at scale } \beta\} \]

**Covering number**

\[ N_p(\mathcal{F}, \beta, n) = \sup_{x_1, \ldots, x_n} N_p(\mathcal{F}, \beta; x_1, \ldots, x_n) \]
You can think of \( V \subset \mathbb{R}^n \) as a finite discretization of \( F|_{x_1, \ldots, x_n} \subset \mathbb{R}^n \) to scale \( \beta \) in the normalize \( \ell_p \) distance as shown in Figure below. It can easily be verified that for any \( p, p' \in [1, \infty) \) such that \( p' \leq p \), \( N_{p'}(F, \beta; x_1, \ldots, x_n) \leq N_p(F, \beta; x_1, \ldots, x_n) \).

\[ V = \{v_1, \ldots, v_6\} \]

## 3 Pollard’s bounds

**Lemma 1.** For any given sample \( x_1, \ldots, x_n \), we have

\[
\hat{R}_S(F) \leq \inf_{\beta \geq 0} \left\{ \beta + \sqrt{\frac{2 \log N_1(F, \beta, x_1, \ldots, x_n)}{n}} \right\}
\]

**Proof.** Let \( V \) be any \( \ell_1 \) cover of \( F \) on \( x_1, \ldots, x_n \) at scale \( \beta \) to be set later.

\[
\frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in F} \sum_{t=1}^n \epsilon_t f(x_t) \right] = \frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in F} \sum_{t=1}^n \epsilon_t (f(x_t) - v_f[t]) + \epsilon_t v_f[t] \right]
\]

\[
\leq \frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in F} \sum_{t=1}^n \epsilon_t (f(x_t) - v_f[t]) \right] + \frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in F} \sum_{t=1}^n \epsilon_t v_f[t] \right]
\]

\[
\leq \frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in F} \sum_{t=1}^n \epsilon_t (f(x_t) - v_f[t]) \right] + \frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{v \in V} \sum_{t=1}^n \epsilon_t v_f[t] \right]
\]

\[
\leq \frac{1}{n} \sup_{f \in F} \sum_{t=1}^n |f(x_t) - v_f[t]| + \frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{v \in V} \sum_{t=1}^n \epsilon_t v_f[t] \right]
\]

\[
\leq \beta + \sqrt{\frac{2 \log V}{n}}
\]

Since above statement holds for any cover \( V \), we have

\[
\frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in F} \sum_{t=1}^n \epsilon_t f(x_t) \right] \leq \beta + \sqrt{\frac{2 \log N_1(F, \beta, x_1, \ldots, x_n)}{n}}
\]
Since above statement holds for all $\beta$ we have,

$$\frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t) \right] \leq \inf_{\beta \geq 0} \left\{ \beta + \sqrt{\frac{2 \log N_1(\mathcal{F}; \beta, x_1, \ldots, x_n)}{n}} \right\}$$

Example : Binary function class $\mathcal{F}$

By VC/Sauer/Shelah lemma, for any $\alpha \in [0, 1)$:

$$N_\infty(\mathcal{F}, \alpha, n) = \Pi(\mathcal{F}, n) \leq \left( \frac{e \cdot n}{\text{VC}(\mathcal{F})} \right)^{\text{VC}(\mathcal{F})}$$

Example : Non-decreasing functions mapping from $\mathbb{R}$ to $\mathcal{Y} = [0, 1]$

Discretize $\mathcal{Y} = [-1, 1]$ to $\beta$ granularity as bins $[0, \beta], [\beta, 2\beta], \ldots, [1 - \beta, 1]$. There are $1/\beta$ bins.

Now given $n$ points, $x_1, \ldots, x_n$ sort them in ascending order. Any non-decreasing function can be approximated to accuracy $\beta$ (in the $\ell_\infty$ metric) by picking on these $x_i$’s the lower limit of the interval of the bin the function evaluation at that point belongs to. This is shown in the figure below.

What is the size of this cover?

One possible approach to bound the size of the cover could be to note that there are $n$ points and each can fall in one of $1/\beta$ bins. However this would be too loose and lead to covering number $1/\beta^n$ which does not yield any useful bounds. Alternatively, to describe any element of the cover, all we need to do is to specify for each grid/bin on the $y$ axis, the smallest index $i$ amongst the sorted $x_{\sigma_1}, \ldots, x_{\sigma_n}$ at which the function $f(x_{\sigma_i})$ is larger than the upper end of the bin. One can think of this smallest index as a break-point in the cover for the specific function. Now to bound the size of the cover, note that there are $1/\beta$ bins and each bin can have a break-point that is one of the $n$ indices. Thus the total size of the cover is $n^{1/\beta}$. This is illustrated in the figure below. Hence we have,

$$N_\infty(\mathcal{F}, \beta, n) \leq n^{1/\beta}$$

If we use this with the Pollard’s bounds we get:

$$\hat{R} \leq \inf_{\beta \geq 0} \left\{ \beta + \sqrt{\frac{2 \log n}{n\beta}} \right\} = 2 \left( \frac{2 \log n}{n} \right)^{1/3}$$
4 Dudley Chaining

**Lemma 2.** For any function class $\mathcal{F}$ bounded by 1,

$$\hat{R}_S(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_0^1 \sqrt{\log (\mathcal{N}_2(\mathcal{F}, \alpha, n))} d\alpha \right\} =: \mathcal{D}_S(\mathcal{F})$$

**Proof.** Let $V_j^j$ be an $\ell_2$ cover of $\mathcal{F}$ on $x_1, \ldots, x_n$ at scale $\beta_j = 2^{-j}$. We assume that $V_j$ is the minimal cover so that $|V_j^j| = \mathcal{N}_2(\mathcal{F}, \beta_j, x_1, \ldots, x_n)$. Note that since the function class is bounded by 1, the singleton set

$$V^0 = \{ x \mapsto 0 \}$$

is a cover at scale 1. Now further, for any $f \in \mathcal{F}$ let $v_j^j$ correspond to the element in $V_j^j$ that is $\beta_j$ close to $f$ on the sample in the normalized $\ell_2$ sense. Such element is guaranteed to exist by definition of the cover. Now note that by telescoping sum,

$$f(x_t) = f(x_t) - v_j^0 = (f(x_t) - v_j^N) + \sum_{j=1}^N (v_j^t - v_j^{j-1})$$

Hence we have that,

$$\frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t) \right] = \frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t (f(x_t) - v_j^N) + \epsilon_t \sum_{j=1}^N (v_j^t - v_j^{j-1}) \right]$$

$$\leq \frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t (f(x_t) - v_j^N) \right] + \frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{j=1}^N \sum_{t=1}^n \epsilon_t (v_j^t - v_j^{j-1}) \right]$$

Using Cauchy Shwartz inequality on the first of the two terms above,

$$\leq \frac{1}{n} \mathbb{E}_\epsilon \left[ \left( \sum_{t=1}^n \epsilon_t^2 \right)^{\frac{1}{2}} \sup_{f \in \mathcal{F}} \sum_{t=1}^n \left( f(x_t) - v_j^N \right)^2 \right] + \frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{j=1}^N \sum_{t=1}^n \epsilon_t (v_j^t - v_j^{j-1}) \right]$$

$$= \sup_{f \in \mathcal{F}} \left\{ \frac{1}{n} \sum_{t=1}^n \left( f(x_t) - v_j^N \right)^2 \right\} + \frac{1}{n} \mathbb{E}_\epsilon \left[ \sum_{j=1}^N \sum_{t=1}^n \epsilon_t (v_j^t - v_j^{j-1}) \right]$$

$$\leq \beta_N + \frac{1}{n} \sum_{j=1}^N \sum_{t=1}^n \epsilon_t (v_j^t - v_j^{j-1})$$

where the last step we replaced the first term by $\beta_N$ since $v_j^N$ is the element that is $\beta_N$ close to $f$ in the normalized $\ell_2$ sense. Now define set $W_j \subset \mathbb{R}^n$ as

$$W_j = \{ w = (v_j^1[1] - v_{j-1}^1[1], \ldots, v_j^n[n] - v_{j-1}^n[n]) : f \in \mathcal{F} \}$$
Note that for any $\mathbf{w} \in W^j$,

$$\|\mathbf{w}\|_2 \leq \sup_{f \in F} \left\| \mathbf{v}_j^f - \mathbf{v}_j^{f-1} \right\|_2$$

$$\leq \sup_{f \in F} \left\{ \left\| \mathbf{v}_j^f - (f(x_1), \ldots, f(x_n)) \right\|_2 + \left\| \mathbf{v}_j^{f-1} - (f(x_1), \ldots, f(x_n)) \right\|_2 \right\}$$

$$\leq \sqrt{n} (\beta_j + \beta_{j-1})$$

But $\beta_{j-1} = 2\beta_j$. Hence $\|\mathbf{w}\|_2 \leq 3\sqrt{n} \beta_j$. Also note that $|W^j| \leq |V^j| \times |V^{j-1}|$, since each element in $W^j$ is the difference between one element in $V^j$ and one from $V^{j-1}$. Therefore:

$$\frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in F} \sum_{t=1}^n \epsilon_t f(x_t) \right] \leq \beta_N + \frac{1}{n} \sum_{j=1}^N \mathbb{E}_\epsilon \left[ \sup_{\mathbf{w} \in W^j} \sum_{t=1}^n \epsilon_t \mathbf{w}[t] \right]$$

Using Masart’s finite lemma, we have

$$\leq \beta_N + \frac{1}{n} \sum_{j=1}^N \sqrt{2 \left( \sup_{\mathbf{w} \in W^j} \|\mathbf{w}\|_2^2 \right) \log (|W^j|)}$$

$$\leq \beta_N + \frac{1}{n} \sum_{j=1}^N \sqrt{18n\beta_j^2 \log (|V^j| \times |V^{j-1}|)}$$

$$= \beta_N + \frac{3}{n} \sum_{j=1}^N \beta_j \sqrt{2n \log (|V^j| \times |V^{j-1}|)}$$

$$\leq \beta_N + \frac{6}{n} \sum_{j=1}^N \beta_j \sqrt{n \log (|V^j|)}$$

But $\beta_j = 2(\beta_j - \beta_{j+1})$ and so

$$\leq \beta_N + \frac{12}{n} \sum_{j=1}^N (\beta_j - \beta_{j+1}) \sqrt{n \log (|V^j|)}$$

$$\leq \beta_N + \frac{12}{n} \sum_{j=1}^N (\beta_j - \beta_{j+1}) \sqrt{n \log \left( \mathcal{N}_2(F, \beta_j, n) \right)}$$

$$\leq \beta_N + \frac{12}{\sqrt{n}} \int_{\beta_{N+1}}^{\beta_0} \sqrt{\log \left( \mathcal{N}_2(F, \delta, n) \right)} d\delta$$
Now for any $\alpha$ let $N = \max\{j : \beta_j = 2^j \geq 2\alpha\}$. Hence, for this choice of $N$ we have that $\beta_{N+1} \leq 2\alpha$ and so $\beta_N \leq 4\alpha$ also note that $\beta_{N+1} \geq \frac{\beta_N}{2} \geq \alpha$. Hence

$$\frac{1}{n} \mathbb{E}_t \left[ \sup_{f \in F} \sum_{t=1}^n \epsilon_t f(x_t) \right] \leq 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\log (N_2(F, \delta, n))} d\delta$$

Since choice of $\alpha$ is arbitrary we conclude the theorem taking infimum.

**Non-decreasing functions example:** Let’s go back to the non-decreasing functions example. In the case when $F \subset [0,1]^\mathbb{R}$ corresponds to all non-decreasing functions on the real line, we saw that $N_1(F, \beta, x_1, \ldots, x_n) \leq N_2(F, \beta, x_1, \ldots, x_n) \leq N_\infty(F, \beta, x_1, \ldots, x_n) \leq n^{1/\beta}$. Using the Pollard’s bound we proved in previous class, we were only able to show that $\hat{R}_S(F) \leq O\left(\frac{\log n}{n}\right)^{1/3}$. Using the dudley integral bound we can improve this as follows:

$$\hat{R}_S(F) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{12}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\frac{\log n}{\delta}} d\delta \right\}
\leq 12 \sqrt{\frac{\log n}{n}} \int_{0}^{1} \sqrt{\frac{1}{\delta}} d\delta = 24 \sqrt{\frac{\log n}{n}}$$