Machine Learning Theory (CS 6783)

Lecture 9: Covering Numbers, Pollard and Dudley Bounds

1 Recap

1. For any statistical learning problem we have,

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

2. For any L-Lipchitz loss

$$\frac{1}{n}\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n \epsilon_t \ell(f(x_t),y_t)\right] \le \frac{L}{n}\mathbb{E}_S\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n \epsilon_t f(x_t)\right]$$
$$\mathbb{E}_S\left[\hat{\mathcal{R}}_S(\ell\circ\mathcal{F})\right] \le L\ \mathbb{E}_S\left[\hat{\mathcal{R}}_S(\mathcal{F})\right]$$

2 Covering Number

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

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

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 β.

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

$$\left(\frac{1}{n}\sum_{t=1}^n |f(x_t)-\mathbf{v}_f[t]|^p\right)^{1/p} \le \beta$$

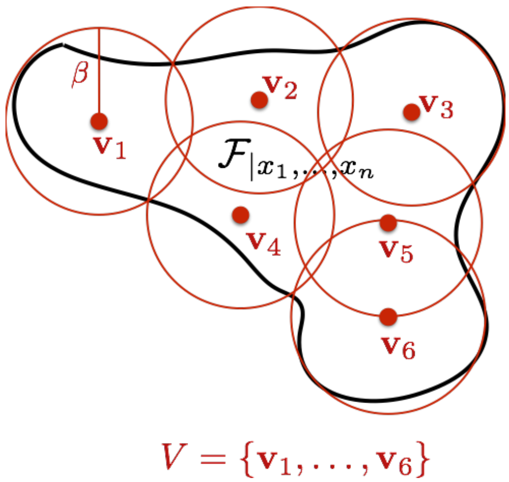
Empirical covering number

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

Covering number

$$\mathcal{N}_p(\mathcal{F},\beta,n) = \sup_{x_1,\ldots,x_n}\mathcal{N}_p(\mathcal{F},\beta;x_1,\ldots,x_n)$$

You can think of $V \subset \mathbb{R}^n$ as a finite discretization of $\mathcal{F}_{|x_1,\dots,x_n} \subset \mathbb{R}^n$ to scale β in the normalize ℓ_p distance as shown in Figure below. It can easily be verified that for any $p, p' \in [1, \infty)$ such that $p' \le p$, $\mathcal{N}_{p'}(\mathcal{F}, \beta; x_1, \dots, x_n) \le \mathcal{N}_p(\mathcal{F}, \beta; x_1, \dots, x_n)$.



3 Pollard's bounds

Lemma 1. *For any given sample $x_1, \dots, x_n$, we have*

$$\hat{\mathcal{R}}_S(\mathcal{F}) \le \inf_{\beta \ge 0} \left\{ \beta + \sqrt{\frac{2 \log \mathcal{N}_1(\mathcal{F}, \beta, x_1, \dots, x_n)}{n}} \right\}$$

Proof. Let V be any ℓ_1 cover of $\mathcal{F}$ on $x_1, \dots, x_n$ at scale β to be set later.

$$\begin{aligned}
\frac{1}{n}\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t) \right] &= \frac{1}{n}\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t \left(f(x_t) - \mathbf{v}_f[t] \right) + \epsilon_t \mathbf{v}_f[t] \right] \\
&\le \frac{1}{n}\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t \left(f(x_t) - \mathbf{v}_f[t] \right) \right] + \frac{1}{n}\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t \mathbf{v}_f[t] \right] \\
&\le \frac{1}{n}\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t \left(f(x_t) - \mathbf{v}_f[t] \right) \right] + \frac{1}{n}\mathbb{E}_\epsilon \left[\sup_{\mathbf{v} \in V} \sum_{t=1}^n \epsilon_t \mathbf{v}_f[t] \right] \\
&\le \frac{1}{n} \sup_{f \in \mathcal{F}} \sum_{t=1}^n |f(x_t) - \mathbf{v}_f[t]| + \frac{1}{n}\mathbb{E}_\epsilon \left[\sup_{\mathbf{v} \in V} \sum_{t=1}^n \epsilon_t \mathbf{v}_f[t] \right] \\
&\le \beta + \sqrt{\frac{2 \log V}{n}}
\end{aligned}$$

Since above statement holds for any cover V, we have

$$\frac{1}{n}\mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t) \right] \le \beta + \sqrt{\frac{2 \log \mathcal{N}_1(\mathcal{F}, \beta, x_1, \dots, x_n)}{n}}$$

Since above statement holds for all β we have,

$$\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(x_t)\right] \leq \inf_{\beta\geq 0}\left\{\beta + \sqrt{\frac{2\log\mathcal{N}_1(\mathcal{F},\beta,x_1,\ldots,x_n)}{n}}\right\}$$

□

Example : Binary function class $\mathcal{F}$
By VC/Sauer/Shelah lemma, for any $\alpha \in [0,1)$:

$$\mathcal{N}_\infty(\mathcal{F},\alpha,n) = \Pi(\mathcal{F},n) \leq \left(\frac{e\ n}{\mathrm{VC}(\mathcal{F})}\right)^{\mathrm{VC}(\mathcal{F})}$$

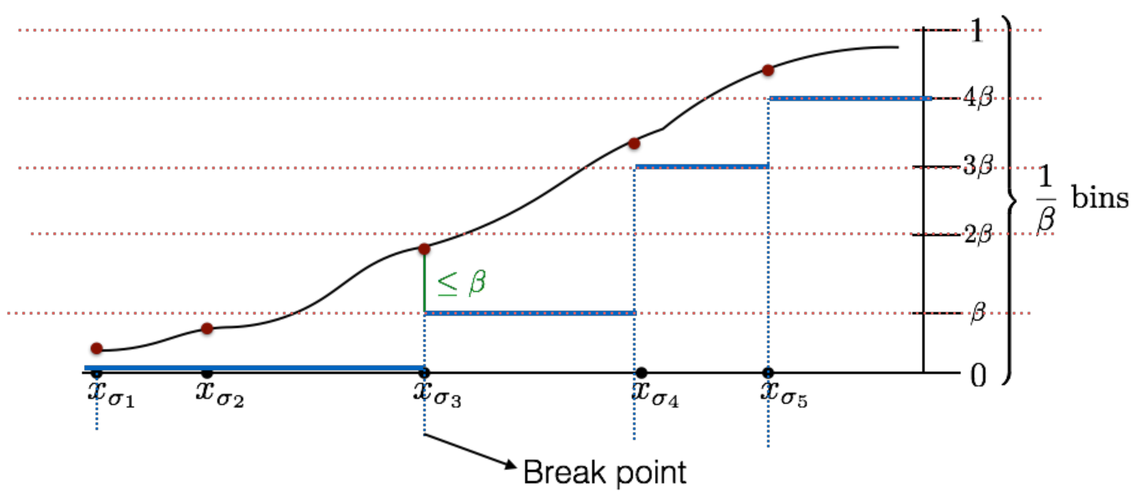
Example : Non-decreasing functions mapping from $\mathbb{R}$ to $\mathcal{Y} = [0,1]$
Discretize $\mathcal{Y} = [-1,1]$ to β granularity as bins $[0,\beta],[\beta,2\beta],\ldots,[1-\beta,1]$. There are $1/\beta$ bins. Now given n points, $x_1,\ldots,x_n$ sort them in ascending order. Any non-decreasing function can be approximated to accuracy β (in the ℓ_∞ 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,

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

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

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



4 Dudley Chaining

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

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

Proof. Let V^j be an ℓ_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 $|V^j| = \mathcal{N}_2(\mathcal{F}, \beta_j, x_1, \ldots, x_n)$. 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 $\mathbf{v}_f^j$ correspond to the element in V^j that is β_j close to f on the sample in the normalized ℓ_2 sense. Such element is guaranteed to exist by definition of the cover. Now note that by telescoping sum,

$$f(x_t) = f(x_t) - \mathbf{v}_f^0 = \left(f(x_t) - \mathbf{v}_f^N[t]\right) + \sum_{j=1}^N \left(\mathbf{v}_f^j[t] - \mathbf{v}_f^{j-1}[t]\right)$$

Hence we have that,

$$\begin{aligned}
\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t f(x_t)\right] &= \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t \left(f(x_t) - \mathbf{v}_f^N[t]\right) + \epsilon_t \sum_{j=1}^N \left(\mathbf{v}_f^j[t] - \mathbf{v}_f^{j-1}[t]\right)\right] \\
&\le \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t \left(f(x_t) - \mathbf{v}_f^N[t]\right)\right] + \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f \in \mathcal{F}} \sum_{j=1}^N \sum_{t=1}^n \epsilon_t \left(\mathbf{v}_f^j[t] - \mathbf{v}_f^{j-1}[t]\right)\right]
\end{aligned}$$

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

$$\begin{aligned}
&\le \frac{1}{n}\mathbb{E}_\epsilon\left[\sqrt{\sum_{t=1}^n \epsilon_t^2}\right] \sqrt{\sup_{f \in \mathcal{F}} \sum_{t=1}^n \left(f(x_t) - \mathbf{v}_f^N[t]\right)^2} + \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f \in \mathcal{F}} \sum_{j=1}^N \sum_{t=1}^n \epsilon_t \left(\mathbf{v}_f^j[t] - \mathbf{v}_f^{j-1}[t]\right)\right] \\
&= \sup_{f \in \mathcal{F}} \sqrt{\frac{1}{n} \sum_{t=1}^n \left(f(x_t) - \mathbf{v}_f^N[t]\right)^2} + \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f \in \mathcal{F}} \sum_{j=0}^N \sum_{t=1}^n \epsilon_t \left(\mathbf{v}_f^j[t] - \mathbf{v}_f^{j-1}[t]\right)\right] \\
&\le \beta_N + \frac{1}{n} \sum_{j=1}^N \mathbb{E}_\epsilon\left[\sup_{f \in \mathcal{F}} \sum_{t=1}^n \epsilon_t \left(\mathbf{v}_f^j[t] - \mathbf{v}_f^{j-1}[t]\right)\right]
\end{aligned}$$

where the last step we replaced the first term by β_N since $\mathbf{v}_f^N$ is the element that is β_N close to f in the normalized ℓ_2 sense. Now define set $W^j \subset \mathbb{R}^n$ as

$$W^j = \{\mathbf{w} = (\mathbf{v}_f^j[1] - \mathbf{v}_f^{j-1}[1], \ldots, \mathbf{v}_f^j[n] - \mathbf{v}_f^{j-1}[n]) : f \in \mathcal{F}\}$$

Note that for any $\mathbf{w} \in W^j$,

$$\begin{aligned}
\|\mathbf{w}\|_2 &\leq \sup_{f \in \mathcal{F}} \left\| \mathbf{v}_f^j - \mathbf{v}_f^{j-1} \right\|_2 \\
&\leq \sup_{f \in \mathcal{F}} \left\{ \left\| \mathbf{v}_f^j - (f(x_1), \ldots, f(x_n)) \right\|_2 + \left\| \mathbf{v}_f^{j-1} - (f(x_1), \ldots, f(x_n)) \right\|_2 \right\} \\
&\leq \sqrt{n} \, (\beta_j + \beta_{j-1})
\end{aligned}$$

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 $\mathcal{W}^j$ is the difference between one element in V^j and one from V^{j-1}. Therefore :

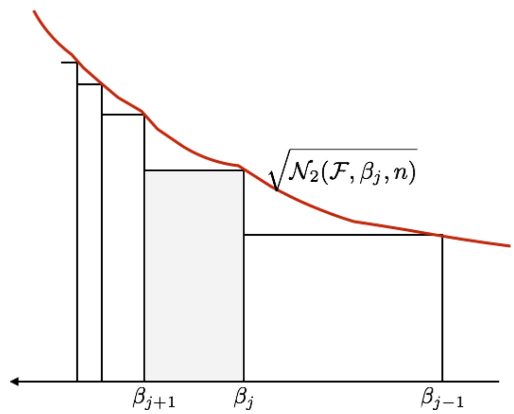
$$\frac{1}{n} \mathbb{E}_\epsilon \left[\sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(x_t) \right] \leq \beta_N + \frac{1}{n} \sum_{j=1}^{N} \mathbb{E}_\epsilon \left[\sup_{\mathbf{w} \in \mathcal{W}^j} \sum_{t=1}^{n} \epsilon_t \mathbf{w}[t] \right]$$

Using Masart's finite lemma, we have

$$\begin{aligned}
&\leq \beta_N + \frac{1}{n} \sum_{j=1}^{N} \sqrt{2 \left(\sup_{\mathbf{w} \in W^j} \|\mathbf{w}\|_2^2 \right) \log\left(|W^j|\right)} \\
&\leq \beta_N + \frac{1}{n} \sum_{j=1}^{N} \sqrt{18 n \beta_j^2 \log\left(|V^j| \times |V^{j-1}|\right)} \\
&= \beta_N + \frac{3}{n} \sum_{j=1}^{N} \beta_j \sqrt{2n \log\left(|V^j| \times |V^{j-1}|\right)} \\
&\leq \beta_N + \frac{3}{n} \sum_{j=1}^{N} \beta_j \sqrt{2n \log\left(|V^j| \times |V^j|\right)} \\
&\leq \beta_N + \frac{6}{n} \sum_{j=1}^{N} \beta_j \sqrt{n \log\left(|V^j|\right)}
\end{aligned}$$

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

$$\begin{aligned}
&\leq \beta_N + \frac{12}{n} \sum_{j=1}^{N} (\beta_j - \beta_{j+1}) \sqrt{n \log\left(|V^j|\right)} \\
&\leq \beta_N + \frac{12}{n} \sum_{j=1}^{N} (\beta_j - \beta_{j+1}) \sqrt{n \log\left(\mathcal{N}_2(\mathcal{F}, \beta_j, n)\right)} \\
&\leq \beta_N + \frac{12}{\sqrt{n}} \int_{\beta_{N+1}}^{\beta_0} \sqrt{\log\left(\mathcal{N}_2(\mathcal{F}, \delta, n)\right)} d\delta
\end{aligned}$$



Now for any α 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

$$\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(x_t)\right] \leq 4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^1 \sqrt{\log\left(\mathcal{N}_2(\mathcal{F},\delta,n)\right)}d\delta$$

Since choice of α is arbitrary we conclude the theorem taking infimum. □

Non-decreasing functions example : Lets go back to the non-decreasing functions example. In the case when $\mathcal{F} \subset [0,1]^{\mathbb{R}}$ corresponds to all non-decreasing functions on the real line, we saw that $\mathcal{N}_1(\mathcal{F},\beta,x_1,\ldots,x_n) \leq \mathcal{N}_2(\mathcal{F},\beta,x_1,\ldots,x_n) \leq \mathcal{N}_\infty(\mathcal{F},\beta,x_1,\ldots,x_n) \leq n^{1/\beta}$. Using the Pollard's bound we proved in previous class, we were only able to show that $\hat{\mathcal{R}}_S(\mathcal{F}) \leq O\left(\frac{\log n}{n}\right)^{1/3}$. Using the dudley integral bound we can improve this as follows :

$$\begin{aligned}\hat{\mathcal{R}}_S(\mathcal{F}) &\leq \inf_{\alpha\geq 0}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_\alpha^1\sqrt{\frac{\log n}{\delta}}d\delta\right\}\\ &\leq 12\sqrt{\frac{\log n}{n}}\int_0^1\sqrt{\frac{1}{\delta}}\,d\delta = 24\sqrt{\frac{\log n}{n}}\end{aligned}$$