Machine Learning Theory (CS 6783)

Lecture 8 : Covering Numbers

1 Recap

1. For any statistical learning problem we have,

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

2. For any L-Lipchitz loss

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

Analogue of growth function and VC dimension?

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*

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

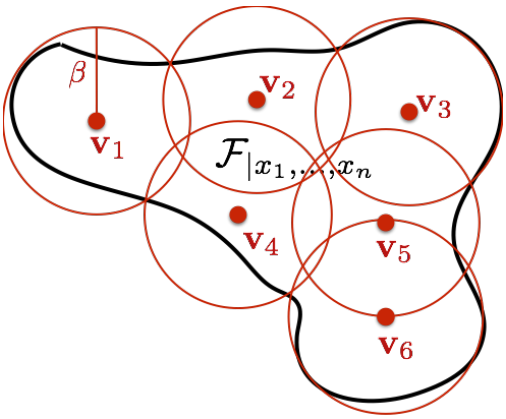
Empirical covering number

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

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 $\mathcal{F}_{|x_1,\ldots,x_n} \subset \mathbb{R}^n$ 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 $p' \leq p$, $\mathcal{N}_{p'}(\mathcal{F}, \beta; x_1, \ldots, x_n) \leq \mathcal{N}_p(\mathcal{F}, \beta; x_1, \ldots, x_n)$.



3 Pollard's bounds

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

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

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

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

Since above statement holds for any cover V, we have

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

Since above statement holds for all β we have,

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

□

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

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

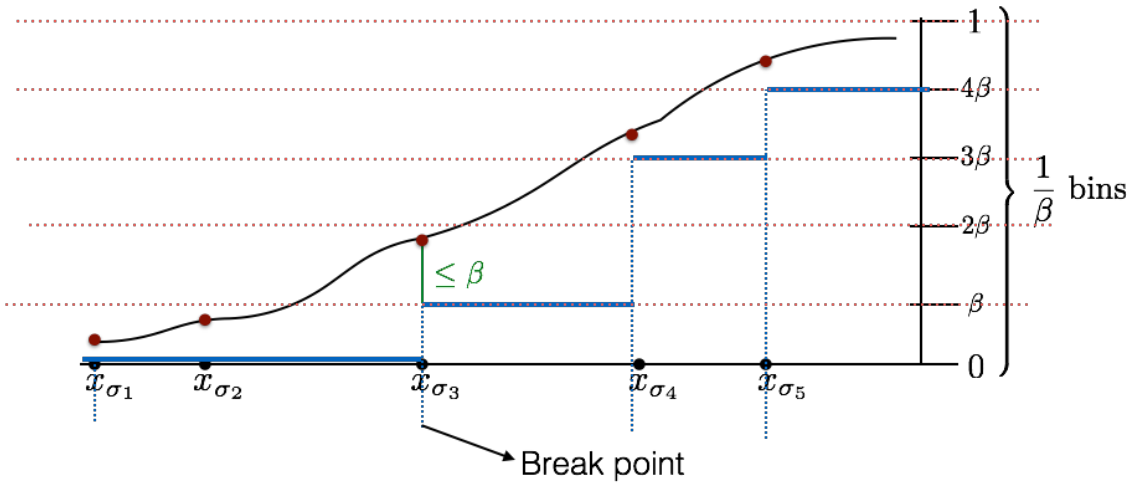
Example : Non-decreasing functions mapping from $\mathbb{R}$ to $\mathcal{Y} = [0,1]$
Discretize $\mathcal{Y} = [-1,1]$ to β granularity as bins $[0,\beta],[\beta,2\beta],\ldots,[1-\beta,1]$. There are $1/\beta$ bins. Now given n points, $x_1,\ldots,x_n$ sort them in ascending order. Any non-decreasing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

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

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

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



4 Dudley Chaining

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

$$\hat{\mathcal{R}}_S(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{6}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\log\left(\mathcal{N}_2(\mathcal{F}, \delta, n)\right)} d\delta \right\} =: \mathcal{D}_S(\mathcal{F})$$

Proof. Let V_j be an ℓ_2 cover of $\mathcal{F}$ on $x_1, \ldots, x_n$ at scale $\beta_j = 2^{-j}$.

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

Define set $W^j \subset \mathbb{R}^n$ as

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

Clearly $|W^j| \leq |V^j| \times |V^{j-1}|$. Also note that for any $\mathbf{w} \in W^j$,

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

But $\beta_{j-1} = 2\beta_j$. Hence

$$\|\mathbf{w}\|_2 \leq 3\, \sqrt{n}\, \beta_j$$

Using above we have :

$$\begin{aligned}\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(x_t)\right] &\leq \beta_N + \frac{1}{n}\sum_{j=0}^{N}\mathbb{E}_\epsilon\left[\sup_{\mathbf{w}\in\mathcal{W}^j}\sum_{t=1}^{n}\epsilon_t \mathbf{w}_f^j[t]\right] \\ &\leq \beta_N + \frac{3}{n}\sum_{j=0}^{N}\beta_j\sqrt{2n\log\left(|V^j|\times|V^{j-1}|\right)} \\ &\leq \beta_N + \frac{3}{n}\sum_{j=0}^{N}\beta_j\sqrt{2n\log\left(|V^j|\times|V^{j}|\right)} \\ &\leq \beta_N + \frac{6}{n}\sum_{j=0}^{N}\beta_j\sqrt{n\log\left(|V^j|\right)}\end{aligned}$$

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

$$\begin{aligned}&\leq \beta_N + \frac{12}{n}\sum_{j=0}^{N}(\beta_j-\beta_{j+1})\sqrt{n\log\left(|V^j|\right)} \\ &\leq \beta_N + \frac{12}{n}\sum_{j=0}^{N}(\beta_j-\beta_{j+1})\sqrt{n\log\left(\mathcal{N}_2(\mathcal{F},\beta_j,n)\right)} \\ &\leq \beta_N + \frac{12}{\sqrt{n}}\int_{\beta_{N+1}}^{\beta_0}\sqrt{\log\left(\mathcal{N}_2(\mathcal{F},\delta,n)\right)}d\delta\end{aligned}$$

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

$$\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(x_t)\right] \leq 4\alpha + \frac{13}{\sqrt{n}}\int_{\alpha}^{1}\sqrt{\log\left(\mathcal{N}_2(\mathcal{F},\delta,n)\right)}d\delta$$

Finally since choice of α is arbitrary we conclude that :

$$\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(x_t)\right] \leq \inf_{\alpha\geq 0}\left\{4\alpha + \frac{12}{\sqrt{n}}\int_{\alpha}^{1}\sqrt{\log\left(\mathcal{N}_2(\mathcal{F},\delta,n)\right)}d\delta\right\}$$

□

Lets go back to the non-decreasing functions example :

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

5 Sudakov's Theorem and Partial Converse

Theorem 3. *There is a universal constant $c > 0$ such that*

$$\hat{\mathcal{R}}_S(\mathcal{F}) \geq \frac{c}{\log n} \sup_{\alpha > 0} \alpha \sqrt{\frac{\log \mathcal{N}_2(\mathcal{F}, \alpha, x_1, \ldots, x_n)}{n}}$$

The above theorem (paraphrased) is due to Sudakov. We shall not go over its proof. However we will see what this implies.

Theorem 4.

$$\frac{c}{12 \log^2 n} \left(\mathcal{D}_S(\mathcal{F}) - \frac{4}{n} \right) \leq \hat{\mathcal{R}}_S(\mathcal{F}) \leq \mathcal{D}_S(\mathcal{F})$$

Proof. We already showed that $\hat{\mathcal{R}}_S(\mathcal{F}) \leq \mathcal{D}_S(\mathcal{F})$. Now on the other hand, we have

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

However by Sudakov's theorem we have that for any $\delta > 0$, we have

$$\sqrt{\frac{\log \mathcal{N}_2(\mathcal{F}, \delta, x_1, \ldots, x_n)}{n}} \leq \frac{\log n \hat{\mathcal{R}}_S(\mathcal{F})}{c \ \delta}$$

Using this,

$$\begin{aligned} \mathcal{D}_S(\mathcal{F}) &\leq \inf_{\alpha > 0} \left\{ 4\alpha + \frac{12}{c} \log n \ \hat{\mathcal{R}}_S(\mathcal{F}) \int_\alpha^1 \frac{1}{\delta} d\delta \right\} \\ &= \inf_{\alpha > 0} \left\{ 4\alpha + \frac{12}{c} \log n \ \log(1/\alpha) \ \hat{\mathcal{R}}_S(\mathcal{F}) \right\} \end{aligned}$$

Picking $\alpha = \frac{1}{n}$ we conclude that

$$\mathcal{D}_S(\mathcal{F}) \leq \frac{4}{n} + \frac{12}{c} \log^2 n \ \hat{\mathcal{R}}_S(\mathcal{F})$$

□