Machine Learning Theory (CS 6783)

Lecture 8: Covering Numbers

1 Covering Numbers

We already saw how to bound Rademacher Complexity in the cases where $\mathcal{F}$ is a finite set of mappings. We are often interested in infinite $\mathcal{F}$. To this end, we will use the notion of covering to bound Rademacher complexity. At a high level, the idea of covering is to approximate $\mathcal{F}$ by a finite family. Recall that the Sequential Rademacher complexity is defined as:

$$\mathcal{R}_n(\mathcal{F}) := \frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n \epsilon_t f(\epsilon_{1:t-1})\right]$$

To understand the notion of cover, let us first start with a simple example. Say we have a family of 2^{n-1} functions indexed by $\epsilon_{1:n-1} \in \{\pm 1\}^{n-1}$ as follows. $\mathcal{F} = \{f_{\epsilon_{1:n-1}} : \epsilon_{1:n-1} \in \{\pm1\}^{n-1}\}$ where $f_{\epsilon_{1:n-1}}(\epsilon_{1:t-1}) = 0$ for any $\epsilon_{1:t-1} \neq \epsilon_{1:n-1}$ and $f_{\epsilon_{1:n-1}}(\epsilon_{1:n-1}) = 1$. That is, $f_{\epsilon_{1:n-1}}$ evaluates to a 1 only on $\epsilon_{1:n-1}$ and 0 for any other input. Clearly $|\mathcal{F}| = 2^{n-1}$. But the claim is that for the purpose of Rademacher complexity, we can cover this class of mappings with just two functions, given by $\overline{\mathcal{F}} = \{f_1, f_2\}$ where f_1 is the constant 0 function and f_2 is a mapping such that for any $t < n-1$, $f_2(\epsilon_{1:t}) = 0$ and $f_2(\epsilon_{1:n-1}) = 1$. That is, f_2 is 0 for any input of length less than $n-1$ and is $+1$ on any input of length $n-1$. Now note that:

$$\mathcal{R}_n(\mathcal{F}) := \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{f\in\mathcal{F}}\sum_{t=1}^n \epsilon_t f(\epsilon_{1:t-1})\right] = \frac{1}{n}\mathbb{E}_\epsilon\left[\max_{f\in\overline{\mathcal{F}}}\sum_{t=1}^n \epsilon_t f(\epsilon_{1:t-1})\right] = \mathcal{R}_n(\overline{\mathcal{F}})$$

Clearly, using the finite bound on $\overline{\mathcal{F}}$ yields a way better bound.

Inspired by this observation let us define the notion of cover and covering numbers.

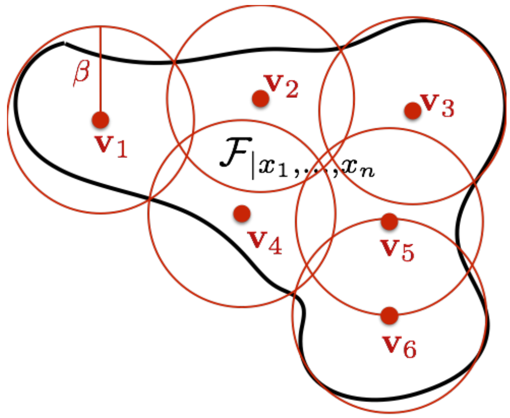
Definition 1. *$V \subset \mathbb{R}^{\bigcup_{t=1}^n \{\pm1\}^{t-1}}$ is an ℓ_p cover of $\mathcal{F} \subset \mathbb{R}^{\bigcup_{t=1}^n \{\pm1\}^{t-1}}$ at scale $\beta > 0$ if, for every $\epsilon \in \{\pm1\}^n$ and for all $f \in \mathcal{F}$, there exists $\mathbf{v}_{f,\epsilon} \in V$ such that*

$$\left(\frac{1}{n}\sum_{t=1}^n |f(\epsilon_{1:t-1}) - \mathbf{v}_{f,\epsilon}(\epsilon_{1:t-1})|^p\right)^{1/p} \leq \beta$$

Covering number is then defined as:

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

To give you a picture, consider the classic Rademacher complexity case. You can think of $V \subset \mathbb{R}^n$ as a finite discretization of $\mathcal{F} \subset \mathbb{R}^n$ 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 $p' \leq p$, $\mathcal{N}_{p'}(\mathcal{F},\beta) \leq \mathcal{N}_p(\mathcal{F},\beta)$.



$V = \{\mathbf{v}_1, \ldots, \mathbf{v}_6\}$

2 Pollard's bounds

Lemma 1. *For any mapping* $\mathcal{F} \subset \mathbb{R}^{\bigcup_{t=1}^n \{\pm 1\}^{t-1}}$,

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

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

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

Since above statement holds for any cover V, we have

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

Since above statement holds for all β we have,

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

$\square$

Example : Classical Rademacher complexity on Non-decreasing functions mapping 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 $f_1, \ldots, f_n$ are in ascending order. Any non-decreasing function can be approximated to accuracy β (even in the ℓ_∞ metric) as is shown in the figure below.

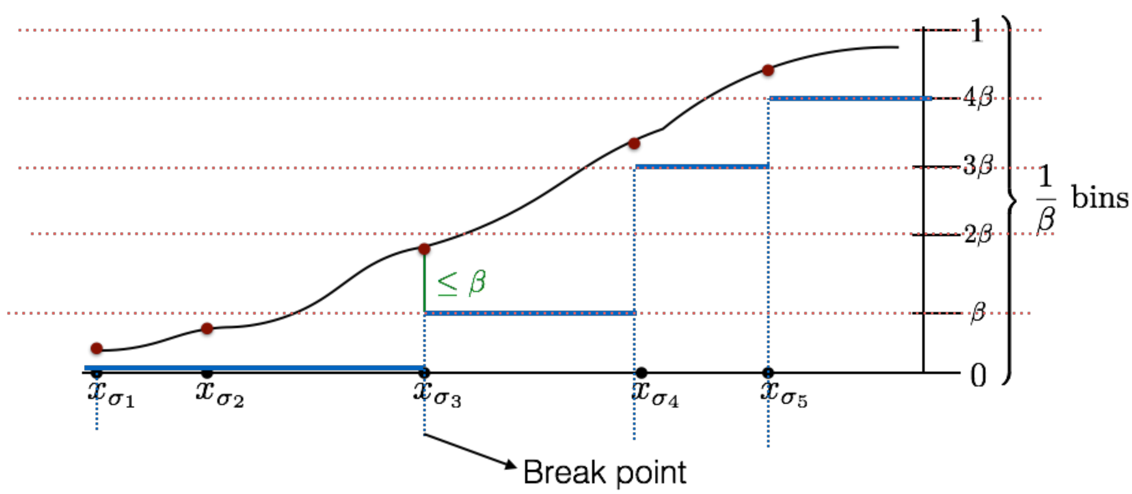
What is the size of this cover?

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 at which the f_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,

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

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

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



3 Dudley Chaining

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

$$\hat{\mathcal{R}}_S(\mathcal{F}) \le \inf_{\alpha \ge 0} \left\{ 4\alpha + \frac{10}{\sqrt{n}} \int_\alpha^1 \sqrt{\log\left(\mathcal{N}_2(\mathcal{F}, \delta)\right)} d\delta \right\} =: \mathcal{D}_S(\mathcal{F})$$

Proof. Let V^j be an ℓ_2 cover of $\mathcal{F}$ at scale $\beta_j = 2^{-j}$. We assume that V_j is the minimal cover so that $|V^j| = \mathcal{N}_2(\mathcal{F}, \beta_j)$. Note that since the function class is bounded by 1, the singleton set

$$V^0 = \left\{ \bigcup_{t=1}^{n} \{\pm 1\}^{t-1} \mapsto 0 \right\}$$

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,

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

Hence we have that,

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

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

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

where the last step we replaced the first term by β_N since $\mathbf{v}_{f,\epsilon}^N$ is the element that is β_N close to f in the normalized ℓ_2 sense. Now define set $W^j \subset \mathbb{R}^{\bigcup_{t=1}^{n}\{\pm1\}^{t-1}}$ as

$$W^j = \left\{ \mathbf{w}_{f,\epsilon}(\epsilon_{1:t-1}) = \mathbf{v}_{f,\epsilon}^j(\epsilon_{1:t-1}) - \mathbf{v}_{f,\epsilon}^{j-1}(\epsilon_{1:t-1}), \text{ and 0 otherwise } : f \in \mathcal{F}, \epsilon \in \{\pm1\}^n \right\}$$

That is each $\mathbf{w}_{f,\epsilon}$ evaluates to $\mathbf{v}_{f,\epsilon}^j - \mathbf{v}_{f,\epsilon}^{j-1}$ when input is a subsequence of ϵ and is 0 otherwise.Note that $|W^j| \le |V^j| \times |V^{j-1}|$, since each element in $\mathcal{W}^j$ is the difference between one element in V^j

and one from V^{j-1}. Therefore :

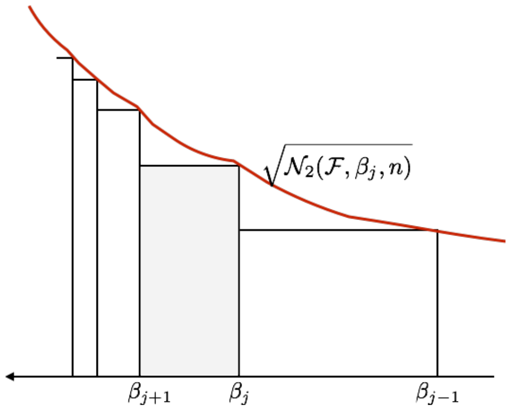
$$
\begin{aligned}
&\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^n \epsilon_t f(\epsilon_{1:t-1})\right] \\
&\quad \leq \beta_N + \frac{1}{n}\sum_{j=1}^N \mathbb{E}_\epsilon\left[\sup_{\mathbf{w}\in\mathcal{W}^j}\sum_{t=1}^n \epsilon_t \mathbf{w}(\epsilon_{1:t-1})\right]
\end{aligned}
$$

Using Masart's finite lemma, we have

$$
\begin{aligned}
&\leq \beta_N + \frac{1}{n}\sum_{j=1}^N \sqrt{2\left(\max_{\mathbf{w}\in W^j,\epsilon\in\{\pm1\}^n}\sum_{t=1}^n \mathbf{w}(\epsilon_{1:t-1})^2\right)\log\left(|W^j|\right)} \\
&\leq \beta_N + \frac{1}{n}\sum_{j=1}^N \sqrt{2\left(\max_{f\in\mathcal{F},\epsilon\in\{\pm1\}^n}\sum_{t=1}^n (\mathbf{v}^j_{f,\epsilon}(\epsilon_{1:t-1}) - \mathbf{v}^{j-1}_{f,\epsilon}(\epsilon_{1:t-1}))^2\right)\log\left((|V^j|\times|V^{j-1}|\right)} \\
&= \beta_N + \frac{1}{n}\sum_{j=1}^N \sqrt{2\left(\max_{f\in\mathcal{F},\epsilon\in\{\pm1\}^n}\sum_{t=1}^n (\mathbf{v}^j_{f,\epsilon}(\epsilon_{1:t-1}) - f(\epsilon_{1:t-1}) + f(\epsilon_{1:t-1}) - \mathbf{v}^{j-1}_{f,\epsilon}(\epsilon_{1:t-1}))^2\right)\log\left((|V^j|\times|V^{j-1}|\right)} \\
&\leq \beta_N + \frac{1}{n}\sum_{j=1}^N \sqrt{4\left(\max_{f\in\mathcal{F},\epsilon\in\{\pm1\}^n}\sum_{t=1}^n (\mathbf{v}^j_{f,\epsilon}(\epsilon_{1:t-1}) - f(\epsilon_{1:t-1}))^2 + (f(\epsilon_{1:t-1}) - \mathbf{v}^{j-1}_{f,\epsilon}(\epsilon_{1:t-1}))^2\right)\log\left((|V^j|\times|V^{j-1}|\right)} \\
&\leq \beta_N + \frac{1}{n}\sum_{j=1}^N \sqrt{4\left(n\beta_j^2 + n\beta_{j-1}^2\right)\log\left((|V^j|\times|V^{j-1}|\right)} \\
&= \beta_N + \frac{1}{\sqrt{n}}\sum_{j=1}^N \sqrt{12\beta_j^2\log\left((|V^j|\times|V^{j-1}|\right)} \\
&\leq \beta_N + \frac{1}{\sqrt{n}}\sum_{j=1}^N \beta_j\sqrt{12\log\left(|V^j|\times|V^j|\right)} \\
&\leq \beta_N + \sqrt{\frac{24}{n}}\sum_{j=1}^N \beta_j\sqrt{\log\left(|V^j|\right)}
\end{aligned}
$$

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

$$
\begin{aligned}
&\leq \beta_N + 2\sqrt{\frac{24}{n}}\sum_{j=1}^N (\beta_j - \beta_{j+1})\sqrt{\log\left(|V^j|\right)} \\
&\leq \beta_N + 2\sqrt{\frac{24}{n}}\sum_{j=1}^N (\beta_j - \beta_{j+1})\sqrt{n\log\left(\mathcal{N}_2(\mathcal{F},\beta_j)\right)} \\
&\leq \beta_N + \frac{10}{\sqrt{n}}\int_{\beta_{N+1}}^{\beta_0}\sqrt{\log\left(\mathcal{N}_2(\mathcal{F},\delta)\right)}d\delta
\end{aligned}
$$



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

$$\frac{1}{n}\mathbb{E}_\epsilon\left[\sup_{f\in\mathcal{F}}\sum_{t=1}^{n}\epsilon_t f(\epsilon_{1:t-1})\right] \leq 4\alpha + \frac{10}{\sqrt{n}}\int_\alpha^1 \sqrt{\log\left(\mathcal{N}_2(\mathcal{F},\delta)\right)}d\delta$$

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