1 Covering Numbers

We already saw how to bound Rademacher Complexity in the cases where \( F \) is a finite set of mappings. We are often interested in infinite \( F \). 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 \( F \) by a finite family. Recall that the Sequential Rademacher complexity is defined as:

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

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. \( F = \{f_{\epsilon_{1:n-1}} : \epsilon_{1:n-1} \in \{\pm 1\}^{n-1}\} \) 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 \( |F| = 2^{n-1} \). But the claim is that for the purpose of Rademacher complexity, we can cover this class of mappings with just two functions, given by \( F = \{f_1, f_2\} \) 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:

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

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

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

**Definition 1.** \( V \subset \mathbb{R}^{\cup_{t=1}^{n}\{\pm 1\}}^{t-1} \) is an \( \ell_p \) cover of \( \mathcal{F} \subset \mathbb{R}^{\cup_{t=1}^{n}\{\pm 1\}}^{t-1} \) at scale \( \beta > 0 \) if, for every \( \epsilon \in \{\pm 1\}^n \) and for all \( f \in \mathcal{F} \), there exists \( v_{f,\epsilon} \in V \) such that

\[
\left( \frac{1}{n} \sum_{t=1}^{n} |f(\epsilon_{1:t-1}) - v_{f,\epsilon}(\epsilon_{1:t-1})|^p \right)^{1/p} \leq \beta
\]

Covering number is then defined as:

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

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 \( \beta \) 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 \( p' \leq p \), \( N_{p'}(\mathcal{F}, \beta) \leq N_p(\mathcal{F}, \beta) \).
Lemma 1. For any mapping $F \subset \mathbb{R}^n \cup \{\pm 1\}^{t-1}$,

$$\mathcal{R}_n(F) \leq \inf_{\beta \geq 0} \left\{ \beta + \sqrt{\frac{2 \log N_1(F, \beta)}{n}} \right\}$$

Proof. Let $V$ be any $\ell_1$ cover of $F$ at scale $\beta$ to be set later.

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

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

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

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

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

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

Since above statement holds for any cover $V$, we have

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

Since above statement holds for all $\beta$ we have,

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

$\square$
Example: Classical Rademacher complexity on Non-decreasing functions mapping to $\mathcal{Y} = [0, 1]$.
Discretize $\mathcal{Y} = [-1, 1]$ to $\beta$ 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 $\beta$ (even in the $\ell_\infty$ 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,

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

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

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

3 Dudley Chaining

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

$$\hat{R}_S(\mathcal{F}) \leq \inf_{\alpha \geq 0} \left\{ 4\alpha + \frac{10}{\sqrt{n}} \int_0^1 \sqrt{\log (\mathcal{N}_2(\mathcal{F}, \delta))} d\delta \right\} =: D_S(\mathcal{F})$$
Proof. Let \( V_j \) be an \( \ell_2 \) 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 \( v_f^j \) correspond to the element in \( V_j \) that is closest 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,

\[
f(\epsilon_{1:t-1}) - f(\epsilon_{1:t-1}) = f(\epsilon_{1:t-1}) - v_f^0(\epsilon_{1:t-1}) = (f(\epsilon_{1:t-1}) - v_f^N(\epsilon_{1:t-1})) + \sum_{j=1}^{N} (v_f^j(\epsilon_{1:t-1}) - v_f^{j-1}(\epsilon_{1:t-1}))
\]

Hence we have that,

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

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

\[
\frac{1}{n} \mathbb{E}_\epsilon \left[ \sum_{j=1}^{N} \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t \left( f(\epsilon_{1:t-1}) - v_f^N(\epsilon_{1:t-1}) \right) \right] + \frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{j=1}^{N} \sum_{t=1}^{n} \epsilon_t \left( v_f^j(\epsilon_{1:t-1}) - v_f^{j-1}(\epsilon_{1:t-1}) \right) \right]
= \sup_{f \in \mathcal{F}} \left[ \frac{1}{n} \sum_{t=1}^{n} \left( f(\epsilon_{1:t-1}) - v_f^N(\epsilon_{1:t-1}) \right)^2 + \frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{j=0}^{N} \sum_{t=1}^{n} \epsilon_t \left( v_f^j(\epsilon_{1:t-1}) - v_f^{j-1}(\epsilon_{1:t-1}) \right) \right] \right]
\leq \beta_N + \frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{j=1}^{N} \sum_{t=1}^{n} \epsilon_t \left( v_f^j(\epsilon_{1:t-1}) - v_f^{j-1}(\epsilon_{1:t-1}) \right) \right]
\]

where the last step we replaced the first term by \( \beta_N \) since \( v_f^N \) is the element that is closest to \( f \) in the normalized \( \ell_2 \) sense. Now define set \( W_j \subset \mathbb{R}^{\left\{ \pm 1 \right\}^{t-1}} \) as

\[
W_j = \left\{ w_{f,\epsilon}(\epsilon_{1:t-1}) = v_f^j(\epsilon_{1:t-1}) - v_f^{j-1}(\epsilon_{1:t-1}), \text{ and } 0 \text{ otherwise}: f \in \mathcal{F}, \epsilon \in \left\{ \pm 1 \right\}^n \right\}
\]

That is each \( w_{f,\epsilon} \) evaluates to \( v_f^j - v_f^{j-1} \) when input is a subsequence of \( \epsilon \) and is 0 otherwise. 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:

$$\frac{1}{n} E_{e} \left[ \sup_{f \in F} \sum_{t=1}^{n} \epsilon_{t} f(\epsilon_{1:t-1}) \right]$$

$$\leq \beta_{N} + \frac{1}{n} \sum_{j=1}^{N} E_{e} \left[ \sup_{w \in W_{j}} \sum_{t=1}^{n} \epsilon_{t} w(\epsilon_{1:t-1}) \right]$$

Using Masmart's finite lemma, we have

$$\leq \beta_{N} + \frac{1}{n} \sum_{j=1}^{N} \sqrt{2 \left( \max_{f \in F, \epsilon \in \{\pm 1\}^{n}} \sum_{t=1}^{n} w(\epsilon_{1:t-1})^{2} \right) \log (|W_{j}|)}$$

$$\leq \beta_{N} + \frac{1}{n} \sum_{j=1}^{N} \sqrt{2 \left( \max_{f \in F, \epsilon \in \{\pm 1\}^{n}} \sum_{t=1}^{n} \left( v_{j}^{f}(\epsilon_{1:t-1}) - v_{j-1}^{f}(\epsilon_{1:t-1}) \right)^{2} \right) \log ((|V_{j}| \times |V_{j-1}|))}$$

$$= \beta_{N} + \frac{1}{n} \sum_{j=1}^{N} \sqrt{4 \left( \max_{f \in F, \epsilon \in \{\pm 1\}^{n}} \sum_{t=1}^{n} \left( v_{j}^{f}(\epsilon_{1:t-1}) - f(\epsilon_{1:t-1}) + f(\epsilon_{1:t-1}) - v_{j-1}^{f}(\epsilon_{1:t-1}) \right)^{2} \right) \log ((|V_{j}| \times |V_{j-1}|))}$$

$$\leq \beta_{N} + \frac{1}{n} \sum_{j=1}^{N} \sqrt{4 \left( n \beta_{j}^{2} + n \beta_{j-1}^{2} \right) \log ((|V_{j}| \times |V_{j-1}|))}$$

$$= \beta_{N} + \sqrt{n} \sum_{j=1}^{N} \sqrt{12 \beta_{j}^{2} \log ((|V_{j}| \times |V_{j-1}|))}$$

$$\leq \beta_{N} + \sqrt{n} \sum_{j=1}^{N} \sqrt{12 \log (|V_{j}| \times |V_{j}|)}$$

$$\leq \beta_{N} + \sqrt{\frac{24}{n} \sum_{j=1}^{N} \sqrt{\log (|V_{j}|)}}$$

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

$$\leq \beta_{N} + 2 \sqrt{\frac{24}{n} \sum_{j=1}^{N} (\beta_{j} - \beta_{j+1}) \sqrt{\log (|V_{j}|)}}$$

$$\leq \beta_{N} + 2 \sqrt{\frac{24}{n} \sum_{j=1}^{N} (\beta_{j} - \beta_{j+1}) \sqrt{n \log (N_{2}(F, \beta_{j}))}}$$

$$\leq \beta_{N} + \frac{10}{\sqrt{n}} \int_{\beta_{N+1}}^{\beta_{0}} \sqrt{\log (N_{2}(F, \delta))} d\delta$$
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

$$\frac{1}{n} \mathbb{E}_\epsilon \left[ \sup_{f \in \mathcal{F}} \sum_{t=1}^{n} \epsilon_t f(\epsilon_{1:t-1}) \right] \leq 4\alpha + \frac{10}{\sqrt{n}} \int_{\alpha}^{1} \sqrt{\log \left( N_2(\mathcal{F}, \delta) \right)} d\delta$$

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