1 Warmup

Assume that for every $\alpha > 0$ that we have a (finite) set $\hat{\mathcal{F}}_\alpha$ such that for all $f \in \mathcal{F}$ there exists an $\hat{f} \in \hat{\mathcal{F}}_\alpha$ such that

$$|\phi(\hat{f}(x), y) - \phi(f(x), y)| \leq \alpha.$$ 

Such an $\hat{\mathcal{F}}_\alpha$ is a $\alpha$-cover of $\mathcal{F}$. Clearly, this implies that:

$$|\mathcal{L}(\hat{f}(x)) - \mathcal{L}(f(x))| \leq \alpha.$$

Hence, we can view $\hat{\mathcal{F}}_\alpha$ as implicitly providing a cover for the loss class.

Intuitively, with respect to obtaining a uniform convergence rate, we could work directly with $\hat{\mathcal{F}}_\alpha$. More precisely,

**Theorem 1.1.** Assume that for all $f \in \mathcal{F}$ our predictions are in $[-1, 1]$. With probability greater than $1 - \delta$

$$\sup_{f \in \mathcal{F}} |\hat{\mathcal{L}}(f) - \mathcal{L}(f)| \leq \inf_{\alpha} 2\sqrt{\log |\hat{\mathcal{F}}_\alpha| + \log \frac{1}{\delta}} + 2\alpha$$

**Proof.** Fix $\alpha$. Using the union bound, we have:

$$\sup_{f \in \mathcal{F}} |\hat{\mathcal{L}}(\hat{f}) - \mathcal{L}(\hat{f})| \leq 2\sqrt{\log |\hat{\mathcal{F}}_\alpha| + \log \frac{1}{\delta}}$$

Let $c(f)$ be the function $\hat{\mathcal{F}}_\alpha$ which covers $f$. Following from the definition of $c(f)$ and $\hat{\mathcal{F}}_\alpha$, we have that for all $f \in \mathcal{F}$,

$$|\mathcal{L}(f) - \mathcal{L}(c(f))| \leq \alpha$$

$$|\hat{\mathcal{L}}(f) - \hat{\mathcal{L}}(c(f))| \leq \alpha$$

It follows that:

$$\sup_{f \in \mathcal{F}} |\hat{\mathcal{L}}(f) - \mathcal{L}(f)| = \sup_{f \in \mathcal{F}} |\hat{\mathcal{L}}(f) - \hat{\mathcal{L}}(c(f)) - (\mathcal{L}(f) - \mathcal{L}(c(f))) + \hat{\mathcal{L}}(c(f)) - \mathcal{L}(c(f))|$$

$$\leq 2\alpha + \sup_{f \in \mathcal{F}} |\hat{\mathcal{L}}(c(f)) - \mathcal{L}(c(f))|$$

$$\leq 2\alpha + \sup_{f \in \hat{\mathcal{F}}_\alpha} |\hat{\mathcal{L}}(\hat{f}) - \mathcal{L}(\hat{f})|$$

$$\leq 2\alpha + \sqrt{\log |\hat{\mathcal{F}}_\alpha| + 2\log \frac{1}{\delta}}$$

The proof is completed by noting that $\alpha$ is arbitrary, so we can take a inf over $\alpha$. \qed
2 p-norm Covering Numbers

The problem with the previous notion of a cover is that it uniformly demands a good approximation to each \( f \) by an element in \( \hat{F}_\alpha \). Intuitively, it seems more natural to have a cover such that for each \( f \in F \) there is an element in the cover which is only on average close \( f \). We now formalize this.

Assume that all hypotheses in our class \( F \) make real valued predictions. Let \( x_{1:n} \) be a set of \( n \) points. A set of vectors \( V \subset \mathbb{R}^n \) is an \( \alpha \)-cover, with respect to the \( p \)-norm, of \( F \) on \( x_{1:n} \) if for all \( f \in F \) there exists a \( v \in V \) such that:

\[
\left( \frac{1}{n} \sum_{i=1}^{n} |v_i - f(x_i)|^p \right)^{\frac{1}{p}} \leq \alpha
\]

We define the \( p \)-norm covering number \( N_p(\alpha, F, x_{1:n}) \) as the size of the minimal such cover \( V \), i.e.:

\[
N_p(\alpha, F, x_{1:n}) = \min \{ |V| : V \text{ is an } \alpha \text{-cover, under the } p \text{-norm, of } F \text{ on } x_{1:n} \}
\]

Also define:

\[
N_p(\alpha, F, n) = \sup x_{1:n} N_p(\alpha, F, x_{1:n})
\]

In other words, \( N_p(\alpha, F, n) \) is the worst case covering number over \( x_{1:n} \).

Observe that:

\[
N_p(\alpha, F, \infty) \leq N_q(\alpha, F, \infty)
\]

for \( p \leq q \). This is consequence of using the (normalized) \( p \)-norm in the definition of the covering number.

Note that:

\[
N_\infty(\alpha, F, \infty) \leq |\hat{F}_\alpha|
\]

which follows directly from the definition of \( \hat{F}_\alpha \).

3 Rademacher Bounds

**Theorem 3.1.** (Discretization) Assume that all \( f \in F \) make predictions in \([-1, 1]\). Let \( \hat{R}_n(F) \) be the empirical Rademacher number of \( F \) on \( x_{1:n} \). We have:

\[
\hat{R}_n(F) \leq \inf_{\alpha} \sqrt{\frac{2\log N_1(\alpha, F, x_{1:n})}{n}} + \alpha
\]

**Proof.** Fix \( \alpha \) and fix a minimal cover \( V \). Define \( B_\alpha(v) \) to be the hypothesis in \( F \) that are \( \alpha \)-covered by \( v \). Using that \( \cup_{v \in V} B_\alpha(v) = F \),

\[
\hat{R}_n(F) = \mathbb{E} \left[ \sup_{f \in F} \left( \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right) \right]
\]

\[
= \mathbb{E} \left[ \sup_{v \in V} \sup_{f \in B_\alpha(v)} \left( \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f(x_i) \right) \right]
\]

\[
= \mathbb{E} \left[ \sup_{v \in V} \sup_{f \in B_\alpha(v)} \left( \frac{1}{n} \sum_{i=1}^{n} \epsilon_i v_i + \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f(x_i) - v_i) \right) \right]
\]

\[
\leq \mathbb{E} \left[ \sup_{v \in V} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i v_i \right] + \mathbb{E} \left[ \sup_{v \in V} \sup_{f \in B_\alpha(v)} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f(x_i) - v_i) \right]
\]
Using Holder’s inequality for the second term,

\[
E \left[ \sup_{v \in V} \sup_{f \in B_\alpha(v)} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i (f(x_i) - v_i) \right] \leq E \left[ \sup_{v \in V} \sup_{f \in B_\alpha(v)} \frac{1}{n} \sum_{i=1}^{n} |f(x_i) - v_i| \right] \leq \alpha
\]

Using Massart’s finite lemma for the first term:

\[
E \left[ \sup_{v \in V} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i v_i \right] \leq \sup_{v \in V} \|v\|_2 \sqrt{\frac{2 \log |V|}{n}} \leq \sqrt{\frac{2 \log |V|}{n}} = \sqrt{\frac{2 \log N_1(\alpha, \mathcal{F}, x_{1:n})}{n}}
\]

The proof is completed by combining these last two bounds and noting that \( \alpha \) was arbitrary (so we can take an inf over all \( \alpha > 0 \)).

The following is immediate:

**Corollary 3.2.** Assume that all \( f \in \mathcal{F} \) make predictions in \([-1, 1]\). We have:

\[
\mathcal{R}_n(\mathcal{F}) \leq \inf_{\alpha} \sqrt{\frac{2 \log N_1(\alpha, \mathcal{F}, n)}{n}} + \alpha
\]