CMPUT 654 Fa 23: Theoretical Foundations of Machine Learning **Fall 2023**

Lecture 9: Symmetrization Lemma

Lecturer: Csaba Szepesvári *Scribes: Shuai Liu*

Note: *LaTeX template courtesy of UC Berkeley EECS dept. (link to directory)*
Disclaimer: *These notes have* ***not*** *been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

9.1 Motivation and Overview

Consider the following example. We have $\mathcal{X} = \mathbb{R}^d$, $\mathcal{Y} = \{0, 1\}$ and our function class is $\mathcal{F} = \{f_w : w \in \mathbb{R}^d\}$ where $f_w = \mathbb{I}(w^\top x \geq 0)$. In this example, the lower bracketing cover is hard to find. Bracketing cover works but in order to get a better rate, we need some other tools: L_p-empirical covering number, uniform entropy (L_1, L_∞) and symmetrization. Back to the example, the set of our function $\mathcal{F}$ is an infinity set so the covering size has to go to infinity in order to cover $\mathbb{R}^d$ up to ε. However, if we take a closer look at the structure of $\mathcal{F}$, the behavior of this function class is restrictive, for example, the magnitude does not matter (we only care about the sign of the inner product, which is only relevant to the direction of w). What's more, note that with a dataset $(X_i, Y_i)_{i=1}^n$, the number of behaviors, i.e., all possible outcomes that $(Y_i)_{i=1}^n$ can take, could go up to 2^n. But with a function from $\mathcal{F}$, the total number of behaviors increases as n^d instead of 2^n (we will see that later). This structure could be taken use of to increase the learning efficiency.

9.2 Empirical Covering Number

We start with a pseudo-meric space $(\mathcal{X}, d)$ where $\mathcal{X}$ is a set of points and $d : \mathcal{X} \times \mathcal{X} \to [0, \infty)$ is a pseudo-meric that satisfies the following properties:

1. For all $x \in \mathcal{X}$, $d(x, x) = 0$.
2. (non-negativity) For all $x, y \in \mathcal{X}$, $d(x, y) \geq 0$.
3. (symmetry) For all $x, y \in \mathcal{X}$, $d(x, y) = d(y, x)$.
4. (triangle inequality) For all $x, y, z \in \mathcal{X}$, $d(x, z) \leq d(x, y) + d(y, z)$.

Note that it does not satisfy the property that $d(x, y) = 0$ only if $x = y$.

Definition 9.1 (ε-cover)**.** The ε-cover of $(\mathcal{X}, d)$ is a finite set $\{x_i\}_{i=1}^n$ such that for all $x \in \mathcal{X}$, there exists $i \in [n]$ satisfying $d(x, x_i) \leq \varepsilon$.

The empirical cover is w.r.t. the empirical metric, which we define below.

Definition 9.2 (Empirical L_p metric)**.** Let $\mathcal{G} \subseteq \mathbb{R}^{\mathcal{Z}}$ and $z_{1:n} \subset \mathcal{Z}^n$. For all $g \in \mathcal{G}$, the empirical L_p-norm $\|\cdot\|_{L_p(z_{1:n})}$ is defined to be

$$\|g\|_{L_p(z_{1:n})} = \left(\frac{1}{n}\sum_{i=1}^n |g(z_i)|^p\right)^{1/p} .$$

The empirical L_p-norm induces a metric d, i.e., for all $g_1, g_2 \in \mathcal{G}$, the empirical L_p metric is defined to be $d(g_1, g_2) = \|g_1 - g_2\|_{L_p(z_{1:n})}$.

Different from bracketing cover, the empirical covering has to be a subset of the whole set. As in bracketing number, we define the empirical L_p covering number

$$\mathcal{N}_p(\varepsilon, \mathcal{G}, z_{1:n}) = \min\{n \geq 1 : \exists g_1, ..., g_n \text{ that forms an } \varepsilon\text{-cover w.r.t. the empirical } L_p \text{ metric}\}.$$

We further define the uniform empirical L_p covering number as $\mathcal{N}(\varepsilon, \mathcal{G}, n) = \sup_{z_{1:n}} \mathcal{N}_p(\varepsilon, \mathcal{G}, z_{1:n})$.

There is a proposition in real analysis about the relationship between L_p and L_q spaces which we state below.

Proposition 9.3. *Let (Ω, Σ, μ) be a finite measure space and $1 \leq p \leq q \leq \infty$. Then $\|\cdot\|_{L_p} \leq C\|\cdot\|_{L_q}$, where $C = \mu(\Omega)^{1/p-1/q}$. In particular, if $\mu(\Omega) = 1$, then $L_p \leq L_q$.*

The measure corresponding to empirical L_p-norm is the mixture of diracs on $z_1, ..., z_n$, i.e., $\mu(z) = \frac{1}{n}$ if $z \in \{z_i\}_{i=1}^n$ and $\mu(z) = 0$ otherwise. Hence we have the following corollary.

Corollary 9.4. *For $1 \leq p \leq q \leq \infty$ and $z_{1:n} \in \mathcal{Z}^n$, it follows that $L_p(z_{1:n}) \leq L_q(z_{1:n})$.*

9.3 Symmetrization

The symmetrization lemma that we are going to display here is counter-intuitive so let's focus on the result itself and we will see why we need it later. We need to first introduce some notations. Let $P \in \mathcal{M}_1(\mathcal{Z})$ be a probability measure and $Z_{1:n}, Z'_{1:n} \sim P$ be i.i.d. samples from P where $Z'_{1:n}$ are called shadow samples. As before, we define the empirical measure on $Z_{1:n}, Z'_{1:n}$:

$$P_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z_i}, P'_n = \frac{1}{n}\sum_{i=1}^n \delta_{Z'_i}.$$

Take $s \in \{\pm 1\}^d$ and create signed empirical measures

$$P_{s,n} = \frac{1}{n}\sum_{i=1}^n s_i\delta_{Z_i}, P'_{s,n} = \frac{1}{n}\sum_{i=1}^n s_i\delta_{Z'_i}.$$

Theorem 9.5 (Symmetrization Lemma)**.** *Let $\sigma \sim \mathrm{Rad}(n)$ be a sample of Rademachar distribution (the discrete uniform distribution over $\{\pm 1\}^n$) that is independent of $Z_{1:n}, Z'_{1:n}$. For $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{Z}}$, functions $\psi : \mathcal{F} \times \mathcal{Z}^n \to \mathbb{R}$, $\tilde{\psi} : \mathcal{F} \times \mathcal{Z}^{2n} \to \mathbb{R}$ and $\varepsilon > 0$, $0 < \delta < 1$, assume the following holds:*

(U) w.p. $1 - \delta$, for all $f \in \mathcal{F}$, it follows that

$$P_{\sigma,n} f \leq \psi(f, Z_{1:n}) + \varepsilon.$$

(NU) For all $f \in \mathcal{F}$, $z_{1:2n} \in \mathcal{Z}^{2n}$, it follows that

$$\psi(f, z_{1:n}) + \psi(f, z_{n+1:2n}) \leq \tilde{\psi}(f, z_{1:2n}).$$

(S) For all $f \in \mathcal{F}$, $z_{1:2n} \in \mathcal{Z}^{2n}$ and $\pi \in \mathrm{Perm}([2n])$, it follows that

$$\tilde{\psi}(f, z_{1:2n}) = \tilde{\psi}(f, z_{\pi(1:2n)})$$

where $\mathrm{Perm}(A) = \{f : A \to A | f \text{ is a bijection}\}$ is the set of all the permutations on a finite set A and $z_{\pi(1:2n)} = z_{\pi(1):\pi(2n)}$.

Then w.p. $1 - 2\delta$, it holds that for all $f \in \mathcal{F}$,

$$P'_n f \leq P_n f + \tilde{\psi}(f, Z_{1:n} Z'_{1:n}) + 2\varepsilon.$$

For a fixed $s \in \{\pm 1\}^n$, let

$$\mathcal{E}_s := \{\forall f \in \mathcal{F}, P'_{s,n}f - P_{s,n}f \leq \tilde{\psi}(f, Z_{1:n}, Z'_{1:n}) + 2\varepsilon\}$$

and let $\hat{\mathcal{E}}$ be defined as

$$\hat{\mathcal{E}} := P'_{\sigma,n}f - P_{\sigma,n}f \leq \tilde{\psi}(f, Z_{1:n}, Z'_{1:n}) + 2\varepsilon. \tag{9.1}$$

We now state an intuitive lemma and delay the proof to the end.

Lemma 9.6. *Under the conditions of Theorem 9.5, for all $s \in \{\pm 1\}^n$, $\mathbb{P}(\mathcal{E}_s) = \mathbb{P}(\mathcal{E}_1)$.*

proof of Theorem 9.5. Note that we only need to prove that $\mathbb{P}(\mathcal{E}_1) \geq 1 - 2\delta$ by definition. Let $\mathcal{E} = \{\forall f \in \mathcal{F} : -P_{\sigma,n}f \leq \psi(f, Z_{1:n}) + \varepsilon\}$ and $\mathcal{E}' = \{\forall f \in \mathcal{F} : P'_{\sigma,n}f \leq \psi(f, Z'_{1:n}) + \varepsilon\}$. Then from the assumptions specified in Theorem 9.5, we can obtain $\mathbb{P}(\mathcal{E}) = \mathbb{P}(\mathcal{E}')$ because $-P_{\sigma,n}f = P_{-\sigma,n}f$ by definition and $\sigma \overset{D}{=} -\sigma$, $Z_{1:n} \overset{D}{=} Z'_{1:n}$ where $\overset{D}{=}$ denotes equality in distribution. Then by union bound, $\tilde{\mathcal{E}} = \mathcal{E} \cap \mathcal{E}'$ holds w.p. $1 - 2\delta$. By definition of $\tilde{\mathcal{E}}$ and $\hat{\mathcal{E}}$, we have that $\tilde{\mathcal{E}} \subseteq \hat{\mathcal{E}}$ hence $\mathbb{P}(\hat{\mathcal{E}}) \geq \mathbb{P}(\tilde{\mathcal{E}}) \geq 1 - 2\delta$. Now it suffices to show that $\mathbb{P}(\hat{\mathcal{E}}) = \mathbb{P}(\mathcal{E}_1)$. Since $\hat{\mathcal{E}} = \cup_{s \in \{\pm 1\}^n} \{\sigma = s\} \cap \hat{\mathcal{E}}$,

$$\begin{aligned}
\mathbb{P}(\hat{\mathcal{E}}) &= \mathbb{P}\left(\bigcup_{s \in \{\pm 1\}^n} \{\sigma = s\} \cap \hat{\mathcal{E}}\right) \\
&= \sum_{s \in \{\pm 1\}^n} \mathbb{P}\left(\{\sigma = s\} \cap \hat{\mathcal{E}}\right) && (\{\sigma = s\} \text{ are disjoint sets}) \\
&= \sum_{s \in \{\pm 1\}^n} \mathbb{P}\left(\{\sigma = s\} \cap \mathcal{E}_s\right) \\
&= \sum_{s \in \{\pm 1\}^n} \mathbb{P}\left(\{\sigma = s\}\right) \mathbb{P}\left(\mathcal{E}_s\right) && (\text{independence between } \sigma \text{ and } Z_{1:n}, Z'_{1:n}) \\
&= \frac{1}{2^n} \sum_{s \in \{\pm 1\}^n} \mathbb{P}(\mathcal{E}_s) \\
&= \mathbb{P}(\mathcal{E}_1). && (\text{Lemma 9.6})
\end{aligned}$$

□

proof of Lemma 9.6. Fix $(s_1, ..., s_n) = s \in \{\pm 1\}^n$ and let $s_{i,-} = (s_1, ..., -s_i, ..., s_n)$ be the sign vector that flips s in the i-th position. Then it suffices to show that $\mathbb{P}(\mathcal{E}_s) = \mathbb{P}(\mathcal{E}_{s_{i,-}})$ for all $i \in [n]$ because for all $s, s' \in \{\pm 1\}^n$, we can transform $\mathcal{E}_s$ to $\mathcal{E}_{s'}$ by flipping signs for at most n times without changing the probability. Then for $f \in \mathcal{F}$, we introduce the abbreviated notation

$$\begin{aligned}
R_{-i} &= (Z_1, \ldots, Z_{i-1}, Z_{i+1}, \ldots, Z_n, Z'_1, \ldots, Z'_{i-1}, Z'_{i+1}, \ldots, Z'_n), \\
U(Z_i, Z'_i, R_{-i}, f) &= \frac{s_i(f(Z_i) - f(Z'_i))}{n} + \frac{1}{n}\left(\sum_{j \neq i} s_j(f(Z'_j) - f(Z_j))\right) \\
V(Z_i, Z'_i, R_{-i}, f) &= \tilde{\psi}(f, Z_{1:n}, Z'_{1:n}) + 2\varepsilon
\end{aligned}$$

we write out $\mathcal{E}_s$ and $\mathcal{E}_{s_{i,-}}$:

$$\begin{aligned}
\mathcal{E}_s &= \{\forall f \in \mathcal{F} : U(Z_i, Z'_i, R_{-i}, f) \leq V(Z_i, Z'_i, R_{-i}, f)\} \\
\mathcal{E}_{s_{i,-}} &= \{\forall f \in \mathcal{F} : U(Z'_i, Z_i, R_{-i}, f) \leq V(Z_i, Z'_i, R_{-i}, f)\}
\end{aligned}$$

Then by tower rule,

$$\mathbb{P}(\forall f \in \mathcal{F} : U(Z_i, Z_i', R_{-i}, f) \leq V(Z_i, Z_i', R_{-i}, f)) = \mathbb{E}[\mathbb{P}(\forall f \in \mathcal{F} : U(Z_i, Z_i', R_{-i}, f) \leq V(Z_i, Z_i', R_{-i}, f) | R_{-i})].$$

It suffices to prove that

$$\mathbb{P}(\forall f \in \mathcal{F} : U(Z_i, Z_i', R_{-i}, f) \leq V(Z_i, Z_i', R_{-i}, f) | R_{-i}) = \mathbb{P}(\forall f \in \mathcal{F} : U(Z_i', Z_i, R_{-i}, f) \leq V(Z_i, Z_i', R_{-i}, f) | R_{-i}).$$

Assume the existence of the regular conditional distribution $P_{Z_i, Z_i' | R}(dz_i, dz_i' | R)$, the LHS can be written as

$$\begin{aligned}
&\int_{\mathcal{Z}^2} P_{Z_i, Z_i' | R}(dz_i, dz_i' | R) \mathbb{I}(\forall f \in \mathcal{F}, U(z_i, z_i', R_{-i}, f) \leq V(z_i, z_i', R_{-i}, f)) \\
= &\int_{\mathcal{Z}^2} P_{Z_i, Z_i'}(dz_i, dz_i') \mathbb{I}(\forall f \in \mathcal{F}, U(z_i, z_i', R_{-i}, f) \leq V(z_i, z_i', R_{-i}, f)) && \text{(Independence)} \\
= &\int_{\mathcal{Z}^2} P_{Z_i', Z_i}(dz_i', dz_i) \mathbb{I}(\forall f \in \mathcal{F}, U(z_i', z_i, R_{-i}, f) \leq V(z_i', z_i, R_{-i}, f)) && ((Z_i, Z_i') \overset{D}{=} (Z_i', Z_i)) \\
= &\int_{\mathcal{Z}^2} P_{Z_i', Z_i}(dz_i', dz_i) \mathbb{I}(\forall f \in \mathcal{F}, U(z_i', z_i, R_{-i}, f) \leq V(z_i, z_i', R_{-i}, f)) && \text{(Assumption (S))} \\
= &\int_{\mathcal{Z}^2} P_{Z_i', Z_i | R}(dz_i', dz_i | R) \mathbb{I}(\forall f \in \mathcal{F}, U(z_i', z_i, R_{-i}, f) \leq V(z_i, z_i', R_{-i}, f)), && \text{(Independence)}
\end{aligned}$$

which is RHS by definition. □