36-710: Advanced Statistical Theory **Fall 2018**

Lecture 23: November 19

Lecturer: Alessandro Rinaldo *Scribe: Ron Yurko*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

23.1 Last Time

Want to bound supremum of empirical process,

$$||P_n - P||_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} (f(X_i) - \mathbb{E}[f(X_i)])|$$

where $(X_1, \dots, X_n) \overset{iid}{\sim} P$ on probability space $(\mathcal{X}, \mathcal{B})$. $\mathcal{F}$ is a class of real value functions on $\mathcal{X}$, and uniformly bounded:

$$\sup_{x \in \mathcal{X}} |f(x)| \leq B, \ \forall f \in \mathcal{F}.$$

We will rely on the Rademacher complexity of $\mathcal{F}$,

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\underset{\sim}{X}, \underset{\sim}{\epsilon}} [\sup_{f \in \mathcal{F}} \frac{1}{n} | \sum_{i=1}^{n} \epsilon_i f(X_i)|]$$

Theorem 23.1 *Let $\mathcal{F}$ be a class of functions $(\mathcal{X}, \mathcal{B})$ uniformly bounded by $B > 0$. Then for any data generating distribution P for $(X_1, \dots, X_n)$ and for all $t > 0$,*

$$\mathbb{P}(||P_n - P||_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + t) \leq exp(\frac{-nt^2}{2B^2})$$

<u>Remark</u> If $\mathcal{R}_n(\mathcal{F}) \to 0$ as $n \to \infty$, then $||P_n - P||_{\mathcal{F}} \overset{ae}{\to} 0$ (this is actually a if and only if statement).

Proof. Proof has 2 parts:

1. Show $||P_n - P||_{\mathcal{F}}$ concentrates around $\mathbb{E}[||P_n - P||_{\mathcal{F}}]$ (easy part)
2. Bound $\mathbb{E}[||P_n - P||_{\mathcal{F}}]$ by $2\mathcal{R}_n(\mathcal{F})$ (hard part)

<u>Part 1:</u> Use the <u>Bounded Difference</u> inequality $\to$ off-the-shelf inequality meant for this situation.

Let $(X_1, \dots, X_n) \subset \mathcal{X}$ be arbitrary n-tuple of points. Define the function $G : \mathcal{X}^n \Rightarrow \mathbb{R}$,

$$x_1^n = (x_1, \dots, x_n) \Rightarrow \sup_{f \in \mathcal{F}} \frac{1}{n} | \sum_{i} \bar{f}(x_i)|,$$

where $\bar{f}(x_i) = f(x_i) - \mathbb{E}[f(x_i)]$. Let x_1^n, y_1^n be in $\mathcal{X}^n$ such that $x_i = y_i$ all $i \neq j$ for some $j \in \{1, \ldots, n\}$. For $x_1^n = (x_1, \ldots, x_n)$ and $y_1^n = (y_1, \ldots, y_n)$ differ only along coordinates j, then $|G(x_i^n) - G(y_i^n)| \leq \frac{2B}{n}$. So by the bounded difference inequality,

$$\mathbb{P}(||P_n - P||_{\mathcal{F}} \geq \mathbb{E}[||P_n - P||_{\mathcal{F}}] + t) \leq \exp(\frac{-nt^2}{2B^2})$$

End of part 1 .

Step 2: Need to bound $\mathbb{E}[||P_n - P||_{\mathcal{F}}]$. We will handle this with a general result:

Theorem 23.2 *(Symmetrization Inequalities)*

Let $\mathcal{F}$ be a class of integrable functions on $(\mathcal{X}, \mathcal{B})$. Let $||\mathcal{R}_n||_{\mathcal{F}} = \underset{f \in \mathcal{F}}{sup} \frac{1}{n} |\sum_i^n \epsilon_i f(X_i)|$ where $\underset{\sim}{X} = (X_1, \ldots, X_n) \overset{iid}{\sim} P \perp \underset{\sim}{\epsilon} = (\epsilon_1, \ldots, \epsilon_n) \overset{iid}{\sim}$ Rademacher. Then for any convex non-decreasing function $\phi : \mathbb{R}_+ \to \mathbb{R}_+$,

$$\mathbb{E}[\phi(||P_n - P||_{\mathcal{F}})] \leq \mathbb{E}_{\underset{\sim}{X}, \underset{\sim}{\epsilon}}[\phi(2||\mathcal{R}_n||_{\mathcal{F}})]$$
$$\textit{also } \mathbb{E}_{\underset{\sim}{X}, \underset{\sim}{\epsilon}}[\phi(\frac{1}{2}||\mathbb{R}_n||_{\tilde{\mathcal{F}}})] \leq \mathbb{E}[\phi(||P_n - P||_{\mathcal{F}})]$$
$$\tilde{\mathcal{F}} = \{f - \mathbb{E}[f(X_1)],\ f \in \mathcal{F}\}$$

Returning to the proof, notice that $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\underset{\sim}{X}, \underset{\sim}{\epsilon}}[||\mathcal{R}_n||_{\mathcal{F}}]$. Take $\phi(t) = t$ (meaning ϕ is identity) to conclude that,

$$\mathbb{E}[||P_n - P||_{\mathcal{F}}] \leq 2\mathbb{E}_{\underset{\sim}{X}, \underset{\sim}{\epsilon}}[||\mathcal{R}_n||_{\mathcal{F}}] = 2\mathcal{R}_n(\mathcal{F})$$

This proves the theorem.

Proof of upper bound of symmetrization inequalities:

$$\mathbb{E}_{\underset{\sim}{X}}[\phi(||P_n - P||_{\mathcal{F}})] = \mathbb{E}_{\underset{\sim}{X}}[\phi(\sup_{f \in \mathcal{F}} \frac{1}{n} |\sum_i^n (f(X_i) - \mathbb{E}[f(X_i)])|)]$$

Let $\underset{\sim}{Y} = (Y_1, \ldots, Y_n) \overset{iid}{\sim} P$ be a "ghost" sample, where $\underset{\sim}{Y} \perp \underset{\sim}{X}$ then $\mathbb{E}[f(X_i)] = \mathbb{E}[f(Y_i)]$ so it

$$= \mathbb{E}_{\underset{\sim}{X}, \underset{\sim}{Y}}[\phi(\sup_{f \in \mathcal{F}} \frac{1}{n} |\sum_i^n (f(X_i) - \mathbb{E}_{Y_i}[f(Y_i)])|)]$$

By Jensen's inequality:

$$\leq \mathbb{E}_{\underset{\sim}{X}, \underset{\sim}{Y}}[\phi(\sup_{f \in \mathcal{F}} \frac{1}{n} |\sum_i^n (f(X_i) - f(Y_i))|)].$$

Next $f(X_i) - f(Y_i) \overset{d}{=} \epsilon_i (f(X_i) - f(Y_i))$ where $\epsilon_i \perp X_i \& Y_i$ is Rademacher. Can now write

$$\leq \mathbb{E}_{\underset{\sim}{X}, \underset{\sim}{Y}, \underset{\sim}{\epsilon}}[\phi(\sup_{f \in \mathcal{F}} \frac{1}{n} |\sum_i^n \epsilon_i (f(X_i) - f(Y_i))|)]$$

Next use the triangle inequality and the fact that ϕ is non-decreasing to get

$$\leq \mathbb{E}_{\underset{\sim}{X}, \underset{\sim}{Y}, \underset{\sim}{\epsilon}}[\phi(\sup_{f \in \mathcal{F}} \frac{1}{n} |\sum_i^n \epsilon_i f(X_i)| + \sup_{f \in \mathcal{F}} \frac{1}{n} |\sum_i^n \epsilon_i f(Y_i)|)]$$

Multiply and divide the two summands on the right-hand side by 2 (constructing a binary random variable),

$$\leq \mathbb{E}_{\underset{\sim}{X},\underset{\sim}{Y},\underset{\sim}{\epsilon}}\left[\phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}\frac{2}{n}\left|\sum_i^n \epsilon_i f(X_i)\right| + \frac{1}{2}\sup_{f\in\mathcal{F}}\frac{2}{n}\left|\sum_i^n \epsilon_i f(Y_i)\right|\right)\right]$$

Then by Jensen's inequality,

$$\begin{aligned}
&\leq \frac{1}{2}\mathbb{E}_{\underset{\sim}{X},\underset{\sim}{Y},\underset{\sim}{\epsilon}}\left[\phi\left(\sup_{f\in\mathcal{F}}\frac{2}{n}\left|\sum_i^n \epsilon_i f(X_i)\right|\right)\right] + \frac{1}{2}\mathbb{E}_{\underset{\sim}{X},\underset{\sim}{Y},\underset{\sim}{\epsilon}}\left[\phi\left(\sup_{f\in\mathcal{F}}\frac{2}{n}\left|\sum_i^n \epsilon_i f(Y_i)\right|\right)\right] \\
&= \mathbb{E}_{\underset{\sim}{X},\underset{\sim}{\epsilon}}\left[\phi\left(\sup_{f\in\mathcal{F}}\frac{2}{n}\left|\sum_i^n \epsilon_i f(X_i)\right|\right)\right] \\
&= \mathbb{E}_{\underset{\sim}{X},\underset{\sim}{\epsilon}}\left[\phi(2||\mathcal{R}_n||_{\mathcal{F}})\right]
\end{aligned}$$

Our <u>goal now</u> (to use this result), is to upper bound,

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{\underset{\sim}{X},\underset{\sim}{\epsilon}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_i^n \epsilon_i f(X_i)\right|\right]$$

23.2 VC Theory

Definition 23.3 *A class F of functions on $\mathcal{X}$ has <u>polynomial</u> discrimination with parameter $\nu \geq 1$ if for each $n \in \mathbb{N}$ and each n-tuple $x_1^n = (x_1, \ldots, x_n) \subset \mathcal{X}$ then set $\mathcal{F}(x_i) = \{(f(x_1), \ldots, f(x_n)) \in \mathbb{R}^n,\ f \in \mathcal{F}\} \subset \mathbb{R}^n$ has cardinality no larger than $(n+1)^\nu$.*

This is a purely combinatorial property of class $\mathcal{F}$, nothing stochastic, always have n-tuple, then doing this for each function in the class, and get this bound.

Lemma 23.4 *Assume $\mathcal{F}$ has polynomial discrimination with parameter ν, then $\forall n \in \mathbb{N}$ and for each $x_1^n \in \mathcal{X}^n$,*

$$\mathbb{E}_{\underset{\sim}{\epsilon}}\left[\sup_{f\in\mathcal{F}}\frac{1}{n}\left|\sum_i^n \epsilon_i f(X_i)\right|\right] \leq 2D(x_1^n)\sqrt{\frac{\nu \log(n+1)}{n}}$$

where $D(x_1^n) = \sup_{f\in\mathcal{F}}\sqrt{\frac{\sum_i^n f(x_i)^2}{n}}$.

Proof. (Exercise) Use $\sum \epsilon_i f(x_i) \sim SG(D(x_1^n)^2)$.

Corollary 23.5 *If F is a class uniformly bounded (by $B > 0$), then $\mathcal{R}_n(\mathcal{F}) \leq 2B\sqrt{\frac{\nu \log(n+1)}{n}}$*

<u>Example:</u> Let $\mathcal{F} = \{I\{(-\infty, z]\},\ z \in \mathbb{R}\}$,

$$||P_n - P||_{\mathcal{F}} = \sup_{z\in\mathbb{R}}|\hat{F}_n(z) - F_X(z)|$$

where $\hat{F}_n(z)$ is the empirical CDF (this is the maximal difference between the CDF and empirical CDF). The $\mathcal{F}$ has polynomial discrimination with parameter $\nu = 1$. Fix $x_1^n \subset \mathbb{R}$, let's order then

$$x_{(1)} \leq x_{(2)} \leq ... \leq x_{(n)}$$

obtain n^{th} intervals

$$(-\infty, x_{(1)}], (x_{(1)}, x_{(2)}], \ldots, (x_{(n)}, \infty)$$

as z varies over $\mathbb{R}$ the function $z \Rightarrow I\{(-\infty, z]\}$ will be 1 or 0 depending on how many interval z crosses in total , there are only $(n+1)$ possible realizations

$$\mathcal{F}(x_1^n) \leq n+1, \text{ all } x_1^n$$

$$\Rightarrow \mathbb{P}(||P_n - P||_{\mathcal{F}} \geq 4\sqrt{\frac{\log n}{n}} + t) \leq \exp\frac{-nt^2}{2}$$

Even stronger result <u>DKW:</u> (1990 Massart)

$$\mathbb{P}(\sup_{z\in\mathbb{R}}|\hat{F}_n(z) - F(z)| \geq t) \leq 2\exp(\frac{-nt^2}{2})$$

This is a sharper result as the $\log n$ term is gone, but this is only for the empirical CDF.