36-710: Advanced Statistical Theory	**Spring 2019**

Lecture 25: April 23

Lecturer: Alessandro Rinaldo *Scribes: Addison J. Hu*

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

25.1 Uniform Law of Large Numbers

Suppose $\mathcal{F} : \mathcal{X} \to \mathbf{R}^d$, a set of real-valued functions on some space.

Definition 25.1 *Rademacher complexity of $\mathcal{F}$.*

$$R_n(\mathcal{F}) = \mathbf{E}_{X_1,\dots,X_n \sim P, \varepsilon_1,\dots,\varepsilon_n \sim \text{Rad}} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \varepsilon_i f(X_i) \right| \right\}$$

where $\underline{X} = X_1, \dots, X_n \overset{\text{iid}}{\sim} P$ *on* $\mathcal{X}$.

Remark. The Rademacher complexity of Ff may be thought of a "measure" of the "size" of $\mathcal{F}$: "how well can functions from $\mathcal{F}$ fit to random noise"?

Theorem 25.2 *Let $\mathcal{F}$ be a class of real-valued functions on $\mathcal{X}$ such that $\|f\|_\infty \le b$ for all $f \in \mathcal{F}$. Then:*

$$\mathbf{P}\left\{ \|P_n - P\|_{\mathcal{F}} \ge 2R_n(\mathcal{F}) + t \right\} \le \exp\left\{ -\frac{nt^2}{2b^2} \right\}$$

for all $t > 0$. Here $\|P_n - P\|_{\mathcal{F}}$ denotes the supremum of an empirical process, i.e.,

$$\|P_n - P\|_{\mathcal{F}} \triangleq \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \left(f(X_i) - \mathbf{E}\, f(X_i) \right) \right|$$

Proof: The proof of this theorem is done in two parts. In **Part I**, we control the variation of $\|P_n - P\|_{\mathcal{F}}$ about its mean (show that it concentrates). In **Part II**, we control the mean of $\|\P_n - P\|_{\mathcal{F}}$ by bounding its supremum.

Part I. For $f \in \mathcal{F}$, denote $\bar{f}(X) = f(X) - \mathbf{E}\, f(X)$. Then we may write:

$$\|P_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(X_i) \right|$$

Now, fix $(X_1, \ldots, X_n)$ to $x_1^n = x_1, \ldots, x_n$ and let:

$$G(x_1^n) \triangleq \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^{n} \bar{f}(X_i) \right|$$

We want to say that if we apply G to a random sequence, it concentrates about its mean. We may do so using the **bounded differences inequality**.

We now show that $G(\cdot)$ satisfies the bounded differences property. Let $x_1^n = (x_1, \ldots, x_n), y_1^n = y_1, \ldots, y_n$ be fixed sequences such that $x_i = y_i$ for all $i \neq j$, and $x_j \neq y_j$ for some j, i.e., they differ only on one coordinate. Then, for any given function $f \in \mathcal{F}$:

$$\begin{aligned}
\frac{1}{n} \left| \sum_{i=1}^{n} \bar{f}(x_i) \right| - \sup_{h \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^{n} \bar{h}(y_i) \right| &\leq \frac{1}{n} \left| \sum_{i=1}^{n} \bar{f}(x_i) \right| - \frac{1}{n} \left| \sum_{i=1}^{n} \bar{f}(y_i) \right| \\
&\leq \frac{1}{n} \sum_{i=1}^{n} \left| \bar{f}(x_i) - \bar{f}(y_i) \right| \\
&\leq \frac{1}{n} \left| \bar{f}(x_j) - \bar{f}(y_j) \right| \\
&\leq \frac{1}{n} \left| \bar{f}(x_j) \right| + \frac{1}{n} \left| \bar{f}(y_j) \right| \\
&\leq \frac{2b}{n}
\end{aligned}$$

This bound is independent of choices x_1^n, y_1^n, f. Therefore we may conclude that:

$$G(x_1^n) - G(y_1^n) \leq \frac{2b}{n}$$

Reverse the roles of x_1^n, y_1^n to obtain:

$$|G(x_1^n) - G(y_1^n)| \leq \frac{2b}{n}$$

This bound works no matter how complicated G is, but leans heavily on the uniform boundedness condition.

We have shown that $G(\cdot)$ satisfies the bounded differences property. Therefore, by McDiarmid's inequality:

$$|\|P_n - P\|_{\mathcal{F}} - \mathbf{E} \|P_n - P\|_{\mathcal{F}}| \leq t$$

with probability at least $1 - 2\exp\{-nt^2/2b^2\}$.

Part II. We now control the mean of $\|P_n - P\|_{\mathcal{F}}$. We will do so by taking the supremum over the function class $\mathcal{F}$. Unfortunately, this class may consist of infinitely many functions. To handle this, we turn to Rademacher complexity.

Theorem 25.3 *Symmetrization. Let $\mathcal{F}$ be a class of integrable functions, i.e., $\mathbf{E}_{X \sim P} |f(X)| < \infty$. Further, denote:*

$$\|R_n\|_{\mathcal{F}} = \sup f \in \mathcal{F} \frac{1}{n} \left| \sum_{i=1}^{n} \varepsilon_i f(X_i) \right|$$

Then, for any nondescending and convex function $\phi : \mathbf{R}_+ \to \mathbf{R}_+$:

$$\mathbf{E} \left\{ \phi(\|P_n - P\|_{\mathcal{F}}) \right\} \leq \mathbf{E}_{X,\varepsilon} \left\{ \phi(2 \|R_n\|_{\mathcal{F}}) \right\}$$

Moreover,

$$\mathbf{E}\left\{\phi(\|P_n - P\|_{\mathcal{F}})\right\} \geq \mathbf{E}_{X,\varepsilon}\left\{\phi(1/2\,\|R_n\|_{\bar{\mathcal{F}}})\right\}$$

where $\bar{\mathcal{F}} = \{f - \mathbf{E}\,f, f \in \mathcal{F}\}$.

Notice that $R_n(\mathcal{F}) = \mathbf{E}_{X,\varepsilon}\|R_n\|_{\mathcal{F}}$. Then if $\phi(x) = x$, we obtain:

$$\mathbf{E}\,\|P_n - P\| \leq 2R_n(\mathcal{F})$$

proving the main theorem. ■

Proof: Of Theorem 25.3. We only provide the proof of the upper bound; the lower bound follows similarly.

Suppose $\underline{X} = (X_1, \ldots, X_n) \perp\!\!\!\perp \underline{\varepsilon} = (\varepsilon_1, \ldots, \varepsilon_n)$. Further suppose a "ghost sample" $\underline{Y} = (Y_1, \ldots, Y_n)$. Then:

$$\begin{aligned}
&\mathbf{E}_{\underline{X}}\left\{\phi(\|P_n - P\|_{\mathcal{F}})\right\} \\
&\quad= \mathbf{E}_{\underline{X}}\left\{\phi\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n}\left(f(X_i) - \mathbf{E}\,f(X_i)\right)\right|\right)\right\} \\
&\quad= \mathbf{E}_{\underline{X}}\left\{\phi\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n}\left(f(X_i) - \mathbf{E}\,f(Y_i)\right)\right|\right)\right\} \\
&\quad\leq \mathbf{E}_{\underline{X},\underline{Y}}\left\{\phi\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n}\left(f(X_i) - f(Y_i)\right)\right|\right)\right\} && \text{(Jensen)} \\
&\quad= \mathbf{E}_{\underline{\mathrm{X}},\underline{\mathrm{Y}},\underline{\varepsilon}}\left\{\phi\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n}\varepsilon_i\left(f(X_i) - f(Y_i)\right)\right|\right)\right\} && (f(X_i) - f(Y_i) \overset{d}{=} \varepsilon_i(f(X_i) - f(Y_i))) \\
&\quad\leq \mathbf{E}_{\underline{\mathrm{X}},\underline{\mathrm{Y}},\underline{\varepsilon}}\left\{\phi\left(\sup_{f\in\mathcal{F}} \frac{1}{n}\left|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right| + \left|\sum_{i=1}^{n}\varepsilon_i - f(Y_i)\right|\right)\right\} && \text{(Triangle)} \\
&\quad\leq \mathbf{E}_{\underline{\mathrm{X}},\underline{\mathrm{Y}},\underline{\varepsilon}}\left\{\frac{1}{2}\phi\left(\sup_{f\in\mathcal{F}} \frac{2}{n}\left|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right) + \frac{1}{2}\phi\left(\sup_{f\in\mathcal{F}} \frac{2}{n}\left|\sum_{i=1}^{n}\varepsilon_i - f(Y_i)\right|\right)\right\} && \text{(Convexity)} \\
&\quad= \mathbf{E}_{\underline{\mathrm{X}},\underline{\varepsilon}}\left\{\phi\left(\sup_{f\in\mathcal{F}} \frac{2}{n}\left|\sum_{i=1}^{n}\varepsilon_i f(X_i)\right|\right)\right\} \\
&\quad= \mathbf{E}_{\underline{\mathrm{X}},\underline{\varepsilon}}\left\{\phi(2\,\|R_n\|_{\mathcal{F}})\right\}
\end{aligned}$$

The lower bound follows the same argument. ■

Due to the lower bound, we obtain the following corollary.

Corollary 25.4 *If* $\|\mathcal{F}\|_\infty \leq b$, *then*

$$\|P_n - P\|_{\mathcal{F}} \geq \frac{1}{2}R_n(\mathcal{F}) - \frac{\sup_{f\in\mathcal{F}}|\mathbf{E}\,f(X)|}{2\sqrt{n}} - t$$

with probability at least $1 - \exp\{-nt^2/2b^2\}$.

It follows immediately that $\|P_n - P\|_{\mathcal{F}} \xrightarrow{p} 0 \Leftrightarrow R_n(\mathcal{F}) \overset{n\to\infty}{\to} 0$. If $\mathcal{F}$ is the Glivenko-Cantelli class of functions, then we satisfy $R_n(\mathcal{F}) \overset{n\to\infty}{\to}$

25.2 Polynomial Discrimination

Our concern now becomes controlling $R_n(\mathcal{F})$.

Definition 25.5 *Polynomial discrimination. A class $\mathcal{F}$ of $f : \mathcal{X} \to \mathbf{R}$ has polynomial discrimination with parameter $\nu \geq 1$ if, for all $n, x_1^n = x_1, \ldots, x_n \in \mathcal{X}$:*

$$\mathcal{F}(x_1^n) = \{(f(x_1), \ldots, f(x_n)) \in \mathbf{R}^n, f \in \mathcal{F}\} \subseteq \mathbf{R}^n$$

has cardinality $\leq (n+1)^\nu$.

Lemma 25.6 *If $\mathcal{F}$ has polynomial discrimination with parameter ν, then for any $x_1^n = x_1, \ldots, x_n$:*

$$\mathbf{E}_{\underline{\varepsilon}} \left\{ \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n \varepsilon_i f(x_i) \right| \right\} \leq D(x_1^n) \sqrt{2\nu \frac{\log(n+1)}{n}}$$

where $D(x_1^n) = \sup_{f \in \mathcal{F}} \sqrt{\frac{1}{n} \sum_{i=1}^n f^2(x_i)}$.

Remark. This does not require the boundedness assumption. If $\|f\|_\infty \leq b$ for all $f \in \mathcal{F}$, then $D_(x_1^n) \leq b$.

Example. Consider the function class:

$$\mathcal{F} = \{\mathbf{1}_{(-\infty, z)}(\cdot), z \in \mathbf{R}\}$$

Observe that $\mathbf{E} f = \mathbf{P}\{X \leq z\}$.

This class of functions has polynomial discrimination with parameter $\nu = 1$. To see this, let $x_1, \ldots, x_n \in \mathbf{R}^n$. This splits $\mathbf{R}$ into at most $n+1$ intervals:

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

Therefore, we may bound the empirical process:

$$\mathbf{P} \left\{ \left\| \hat{F}_n - F \right\|_\infty \geq 4 \sqrt{\frac{\log(n+1)}{n}} + t \right\} \leq \exp\left\{ -\frac{nt^2}{2} \right\}$$

Remark. For a sharper bound, use the DKW inequality.