36-755: Advanced Statistical Theory I	**Fall 2016**

Lecture 14: October 19

Lecturer: Alessandro Rinaldo *Scribes: Enxu Yan*

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

14.1 Uniform Bound via Rademacher Complexity

We are interested in bounding the quantity

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |\mathbb{P}_n f - \mathbb{P} f|$$

where $\mathbb{P}_n f = \frac{1}{n}\sum_{i=1}^n f(X_i)$ and $\mathbb{P} f = \mathbb{E}[f(X)]$ with X and $\{X_i\}_{i=1}^n$ i.i.d. from $\mathbb{P}$. In the following analysis, we assume only boundedness of function $f \in \mathcal{F}$:

Assumption 14.1 *$\mathcal{F}$ is a class of functions $f : \mathcal{X} \to \mathbb{R}$ satisfying $\|f\|_\infty \le b, \forall f \in \mathcal{F}$.*

Given an n-tuple $\boldsymbol{x}^n := (x_1, ..., x_n) \in \mathcal{X}$ and let

$$\mathcal{F}(\boldsymbol{x}^n) = \{(f(x_1), f(x_2), ..., f(x_n)) \in \mathcal{R}^n \mid f \in \mathcal{F}\}.$$

The *empirical Rademacher Complexity* of $\mathcal{F}$ w.r.t. samples $\boldsymbol{x}^n$ is defined as

$$\mathcal{R}_n(\mathcal{F}(\boldsymbol{x}^n)) := \mathbb{E}_{\boldsymbol{\epsilon}}\left[\sup_{f \in \mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)|\right] \tag{14.1}$$

where $\boldsymbol{\epsilon}^n := (\epsilon_1, ..., \epsilon_n)$ are i.i.d. *Rademarcher* random variables (i.e. $\mathbb{P}(\epsilon_i = 1) = \mathbb{P}(\epsilon_i = -1) = 1/2$). $\mathcal{R}_n(\mathcal{F}(\boldsymbol{x}^n))$ computes the expected maximum correlation between n random signs $\boldsymbol{\epsilon}^n$ and points in $\mathcal{F}(\boldsymbol{x}^n)$. The *Rademacher Complexity* of function class $\mathcal{F}$ w.r.t. a distribution $\mathbb{P}$ is then

$$\mathcal{R}_n(\mathcal{F}) := \mathbb{E}_X[\mathcal{R}_n(\mathcal{F}(X^n))] = \mathbb{E}_{X,\boldsymbol{\epsilon}}\left[\sup_{f \in \mathcal{F}} |\frac{1}{n}\sum_{i=1}^n \epsilon_i f(x_i)|\right]. \tag{14.2}$$

Note X^n and $\boldsymbol{\epsilon}^n$ are independent. $\mathcal{R}_n(\mathcal{F})$ is a measure on the "size" of $\mathcal{F}$. If it is large, then $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ could be also large. We want

$$\mathcal{R}_n(\mathcal{F}) \to 0$$

as $n \to \infty$.

14.1.1 Upper Bound

For upper bounding $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$, we have the following theorem.

Theorem 14.2 *Let $\mathcal{F}$ be a class of functions satisfying Assumption 14.1. We have*

$$\mathcal{P}\left(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \geq 2\mathcal{R}_n(\mathcal{F}) + t \right) \leq 2\exp\left\{ -\frac{nt^2}{2b^2} \right\}$$

for all n and $t > 0$.

When $n \to \infty$, from the Theorem, if $\mathcal{R}_n(\mathcal{F}) \to 0$, $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \to 0$ almost surely by Borel-Canteli's lemma. Function class with $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \to 0$ in probability is called *Glivenko-Cantelli* class.

Proof: (Theorem 14.2)

The proof has two parts. Part-(i) shows that $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ converges around its mean $\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}]$. Part-(ii) bounds $\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}]$ by $2\mathcal{R}_n(\mathcal{F})$ (using *symmetrization* technique).

Part i) To apply *bounded differences inequality* on the function

$$G(x_1, ..., x_n) := \sum_{f \in \mathcal{F}} |\frac{1}{n} \sum_{i=1}^{n} \bar{f}(x_i)|,$$

where $\bar{f}(X) = f(X) - \mathbb{E}_X[f(X)]$, we need to verify that $G(.)$ has bounded difference when varying each single coordinate. Since $G(.)$ is invariant to permutation of $(x_1, .., x_n)$, without loss of generality, let $\boldsymbol{y}$ and $\boldsymbol{x}$ differ by only J-th coordinate $y_J \neq x_J$. For any $f \in \mathcal{F}$, we have

$$\begin{aligned} |\frac{1}{n}\sum_{i=1}^{n} \bar{f}(x_i)| - \sup_{h \in \mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} \bar{h}(y_i)| &\leq |\frac{1}{n}\sum_{i=1}^{n} \bar{f}(x_i)| - |\frac{1}{n}\sum_{i=1}^{n} \bar{f}(y_i)| \\ &\leq \frac{1}{n}|\bar{f}(x_J) - \bar{f}(y_J)| \\ &\leq \frac{2b}{n}. \end{aligned}$$

Then taking supremum over $f \in \mathcal{F}$ on both sides, we have $G(\boldsymbol{x}) - G(\boldsymbol{y}) \leq 2b/n$. Similarly, we can obtain $G(\boldsymbol{y}) - G(\boldsymbol{x}) \leq 2b/n$ by the same argument. Then since we have verified $|G(\boldsymbol{x}) - G(\boldsymbol{y})| \leq 2b/n$ when $\boldsymbol{x}$, $\boldsymbol{y}$ differ by a single coordinate, applying the *bounded differences inequality* yields

$$\mathbb{P}\left(\left| \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} - \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] \right| \geq t \right) = \mathbb{P}\left(\left| G(\boldsymbol{x}^n) - \mathbb{E}[G(\boldsymbol{x}^n)] \right| \geq t \right) \leq 2\exp\left\{ -\frac{nt^2}{2b^2} \right\}$$

as desired.

Part ii) Using the *symmetrization* argument, we have

$$\begin{aligned} \mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] &= \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} (f(X_i) - \mathbb{E}_{Y_i}[f(Y_i)]) \right| \right] \\ &= \mathbb{E}_X \left[\sup_{f \in \mathcal{F}} \left| \mathbb{E}_{Y_i}[\frac{1}{n}\sum_{i=1}^{n} (f(X_i) - f(Y_i))] \right| \right] \\ &\leq \mathbb{E}_{X,Y} \left[\sup_{f \in \mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^{n} (f(X_i) - f(Y_i)) \right| \right] \end{aligned}$$

where the last is Jensen's inequality. Now let $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)$ be i.i.d. Rademacher random variables. The distribution of $\epsilon_i(f(X_i) - f(Y_i))$ is exactly the same as $f(X_i) - f(Y_i)$. Therefore,

$$\begin{aligned}\mathbb{E}[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}] &\leq \mathbb{E}_{X,Y,\boldsymbol{\epsilon}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f(X_i) - f(Y_i))\right|\right] \\ &= 2\mathbb{E}_{X,\boldsymbol{\epsilon}}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right] = 2\mathcal{R}_n(\mathcal{F}).\end{aligned}$$

■

14.1.2 Lower Bound

To show that Rademacher complexity gives a tight enough bound on

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

the following theorem gives a more general result on bounding the expectation of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ with the expectation of its symmetrized version

$$\|\mathcal{R}_n\|_{\mathcal{F}} := \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|.$$

Theorem 14.3 *For any convex, non-decreasing function $\Phi : \mathbb{R} \to \mathbb{R}$, we have*

$$\mathbb{E}_{X,\boldsymbol{\epsilon}}\left[\Phi(\frac{1}{2}\|\mathcal{R}_n\|_{\bar{\mathcal{F}}})\right] \leq \mathbb{E}_X\left[\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})\right] \leq \mathbb{E}_{X,\boldsymbol{\epsilon}}\left[\Phi(2\|\mathcal{R}_n\|_{\mathcal{F}})\right], \tag{14.3}$$

where $\bar{\mathcal{F}} = \{f - \mathbb{E}[f] \mid f \in \mathcal{F}\}$.

Proof: The proof for the upper bound is similar to that for Theorem 14.2. First, by applying symmetrization and Jensen's inequality, we have

$$\begin{aligned}\mathbb{E}_X\left[\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})\right] &= \mathbb{E}_X\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(Y_i)]\right|\right)\right] \\ &\leq \mathbb{E}_{X,Y}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - f(Y_i)\right|\right)\right] \\ &\leq \mathbb{E}_{X,Y,\boldsymbol{\epsilon}}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i(f(X_i) - f(Y_i))\right|\right)\right]\end{aligned}$$

Then by triangular inequality and Jensen's inequality (again!), we have

$$\begin{aligned}\mathbb{E}_X\left[\Phi(\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}})\right] &\leq \mathbb{E}_{X,Y,\boldsymbol{\epsilon}}\left[\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right| + \left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(Y_i)\right|\right)\right] \\ &\leq \frac{1}{2}\mathbb{E}_{X,\boldsymbol{\epsilon}}\left[\Phi\left(2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right)\right] + \frac{1}{2}\mathbb{E}_{Y,\boldsymbol{\epsilon}}\left[\Phi\left(2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(Y_i)\right|\right)\right] \\ &= \mathbb{E}_{X,\boldsymbol{\epsilon}}\left[\Phi\left(2\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right|\right)\right],\end{aligned}$$

which proves the upper bound in (14.3). Now for the lower bound, using Jensen's inequality and symmetrization equality, we have

$$\begin{aligned}\mathbb{E}_{X,\boldsymbol{\epsilon}}\left[\Phi\left(\frac{1}{2}\|\mathcal{R}_n\|_{\bar{\mathcal{F}}}\right)\right] &= \mathbb{E}_{X,\boldsymbol{\epsilon}}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(X_i)-\mathbb{E}_Y[f(Y_i)])\right|\right)\right]\\ &\le \mathbb{E}_{X,Y,\boldsymbol{\epsilon}}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \epsilon_i(f(X_i)-f(Y_i))\right|\right)\right]\\ &= \mathbb{E}_{X,Y}\left[\Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i)-f(Y_i))\right|\right)\right]\end{aligned}$$

By triangular inequality and by convexity of $\Phi(.)$, we have

$$\begin{aligned}&\Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i)-f(Y_i))\right|\right)\\ &\le \Phi\left(\frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i)-\mathbb{E}[f(X)])\right| + \frac{1}{2}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(Y_i)-\mathbb{E}[f(X)])\right|\right)\\ &\le \frac{1}{2}\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(X_i)-\mathbb{E}[f(X)])\right|\right) + \frac{1}{2}\Phi\left(\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n (f(Y_i)-\mathbb{E}[f(X)])\right|\right).\end{aligned}$$

Taking expectation on both sides and noticing that Y_i and X_i are identically distributed give the lower bound in (14.3). ■

Note that the lower bound in (14.3) takes norm w.r.t. $\bar{\mathcal{F}}$ instead of $\mathcal{F}$. The following corollary gives a lower bound of $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ in terms of *Rademacher complexity*. It follows directly from the lower bound in Theorem 14.3.

Corollary 14.4 *For a function class $\mathcal{F}$ satisfying assumption 14.1 and any $\delta \ge 0$,*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \ge \frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f]|}{2\sqrt{n}} - \delta \tag{14.4}$$

with probability at least $1 - e^{-\frac{n\delta^2}{2b^2}}$.

The lower bound (14.4) indicates the when $n \to \infty$, if the *Rademacher Complexity* $\mathcal{R}_n(\mathcal{F})$ does not converge to 0, $\|\mathbb{P}_n - \mathbb{P}\|$ will also not go to 0. In other words, the convergence of *Rademacher Complexity* is a **necessary** and **sufficient** condition for $\mathcal{F}$ to be a *Glivenko-Cantelli* class.

In the next lecture, we will focus on how to bound the Rademacher complexity $\mathcal{R}_n(\mathcal{F})$ to get an actual uniform concentration bound for the function class $\mathcal{F}$.