36-755 – Advanced Statistical Theory I **Fall 2017**

Lecture 20: November 8

Lecturer: Alessandro Rinaldo *Scribes: Minshi Peng*

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

20.1 ULLN via Rademacher complexity

Theorem 20.1 *Let $\mathcal{F}$ be a class of real valued functions on $\mathcal{X}$ (i.e. $\mathbb{R}^d$), s.t. $\forall f \in \mathcal{F}$, $\|f\|_\infty \leq b$ for some $b > 0$. Then $\forall t > 0$,*

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

where $X = (X_1, \ldots, X_n) \overset{i.i.d}{\sim} \mathcal{P}$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \overset{i.i.d}{\sim}$ Radmacher, ϵ independent of X,

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

and

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

Actually, $\|P_n - P\|_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + C\sqrt{\frac{\log n}{n}}$ with probability $1 - \frac{1}{n}$

Proof:

1) . Bounded difference inequality applied to $\|P_n - P\|_{\mathcal{F}}$
2) . Symmetrization inequality

Lemma 20.2 *Let $\mathcal{F}$ be a class of integrable (w.r.t $\mathcal{P}$) real valued functions on $\mathcal{X}$ and let*

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

where $X = (X_1, \ldots, X_n) \overset{i.i.d}{\sim} \mathcal{P}$ and $\epsilon = (\epsilon_1, \ldots, \epsilon_n) \overset{i.i.d}{\sim}$ Radmacher, ϵ independent of X. Then for any convex, non-decreasing $\phi : \mathbb{R}_+ \to \mathbb{R}_+$

$$\mathbb{E}_{X,\epsilon}\Big[\phi(\frac{1}{2}\|\mathcal{R}_n\|_{\bar{\mathcal{F}}})\Big] \leq \mathbb{E}_X\Big[\phi(\|P_n - P\|_{\mathcal{F}})\Big] \leq \mathbb{E}_{X,\epsilon}\Big[\phi(2\mathcal{R}_n(\mathcal{F}))\Big]$$

where$\bar{\mathcal{F}} = \{f - \mathbb{E}[f(X)], f \in \mathcal{F}\}$.

Remark:

1) . $\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_{X,\epsilon}\Big[\|\mathcal{R}_n\|_{\mathcal{F}}\Big]$

2) . Take $\phi(x) = x$ to prove the theorem.

Proof of Symmetrization lemma:
By using ghost samples $Y = (Y_1, \dots, Y_n) \overset{i.i.d}{\sim} \mathcal{P}$ where Y independent of X, ϵ and the convexity of ϕ

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

The first inequality is because $f(X_i) - f(Y_i) \overset{d}{=} \epsilon_i(f(X_i) - f(Y_i)), \forall i$ where $\epsilon = (\epsilon_1, \dots, \epsilon_n) \overset{i.i.d}{\sim}$ Radmacher. This concludes the proof of upper bound. The proof of lower bound is similar (refers to Proposition 4.1. in the book). ■

We have seem that $\|P_n - P\|_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + t$ with probability $1 - e^{-\frac{nt^2}{2b^2}}$. Using the lower bound in the symmetrization inequality you can show that

$$\|P_n - P\|_{\mathcal{F}} \geq \frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{\sup_{f\in\mathcal{F}} |\mathbb{E}[f(X)]|}{2\sqrt{n}} - t$$

with probability at least $1 - e^{-\frac{nt^2}{2b^2}}$. It shows that class $\mathcal{F}$ is Glwenko Cantelli w.r.t. $\mathcal{P}$, since $\|P_n - P\|_{\mathcal{F}} \xrightarrow{P} 0$ iff $\mathcal{R}_n(\mathcal{F}) \to 0$ as $n \to \infty$. Thus our task is to control

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

Definition 20.3 *A class $\mathcal{F}$ of real valued functions on $\mathcal{X}$ has polynomail discrimination with parameter $\nu \geq 1$, if $\forall n$and for each n-tuple $xn = (x_1, \dots, x_n)$ of points in $\mathcal{X}$, the set $\mathcal{F}(x_1^n) = \Big\{(f(x_1), \dots, f(x_n)) \in \mathbb{R}^n, f \in \mathcal{F}\Big\}$ has cardinality $\leq (n+1)^\nu$.*

Example $\mathcal{F} = \Big\{1_{(-\infty,x]}, x \in \mathbb{R}\Big\}$ has polynomial discrimination with parameter $\nu = 1$. This is because fix an n-tuple $x_1^n = (x_1, \dots, x_n)$, it splits real line into $n+1$ intervals

$$(-\infty, x_{(1)}], (x_{(2)}, x_{(3)}], \cdots, (x_{(n-1)}, x_{(n)}], (x_{(n)}, \infty),$$

where $x'_{(i)}s$ are order statistics $x_{(1)} \leq x_{(2)} \leq \cdots, \leq x_{(n)}$. The function $1_{(\infty,z]}$ is 1 for all i s.t. $x_{(i)} \leq z$. Thus $|\mathcal{F}(x_1^n)| \leq n+1$.

Lemma 20.4 *If $\mathcal{F}$ has polynomial discrimination with parameter ν, then for any n-tuple x_1^n*

$$\mathbb{E}_{X,\epsilon}\Big[\sup_{f\in\mathcal{F}}\Big|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\Big|\Big] \le 2D_{\mathcal{F}}(x_1^n)\sqrt{\frac{\nu\log(n+1)}{n}}$$

where $D_{\mathcal{F}}(x_1^n) = \sup_{f\in\mathcal{F}}\sqrt{\frac{\sum_{i=1}^n f^2(x_i)}{n}}$ is the L_2 diameter of $\mathcal{F}$.

Example: $\mathcal{F} = \Big\{1_{(-\infty,z]}, z\in\mathbb{R}\Big\}$ so

$$\|P_n - P\|_{\mathcal{F}} = \sup_{z\in\mathbb{R}}\big|F(z) - F_n(z)\big| = \|F - F_n\|_\infty$$

where $F(z)$ is c.d.f. of $\mathcal{P}$ and $F_n(z)$ is empirical c.d.f.

Corollary 20.5

$$\mathbb{P}\Big(\|F - F_n\|_\infty \ge 4\sqrt{\frac{\log n}{n}} + t\Big) \le \exp\Big\{-\frac{nt^2}{2b^2}\Big\}$$

Which means $\|F - F_n\|_\infty \lesssim \sqrt{\frac{\log n}{n}}$ with probability at least $1 - \frac{1}{n}$.

The sharpest result is **DKW Iequality**

$$\mathbb{P}(\|F - F_n\|_\infty \ge t) \le 2\exp\Big\{-\frac{nt^2}{2}\Big\}$$

The constants are due to (Massart 1990).

20.2 VC Theory

For now assume $\mathcal{F}$ consists of binary 0-1-functions. Such that $\big|\mathcal{F}(x_1^n)\big| \le 2^n$. But we want $\big|\mathcal{F}(x_1^n)\big| \le (n+1)^\nu$.

Definition 20.6 *We say that the n-tuple x_1^n is shattered by $\mathcal{F}$ if $|\mathcal{F}(x_1^n)| = 2^n$. The VC-dimension of $\mathcal{F}$ is the largest integer n for which some n-tuple $x_1^n = (x_1, \ldots, x_n) \subset \mathcal{X}$ is shattered by $\mathcal{F}$ (If $\mathcal{F}$ has VC-dimension ν, then if $n > \nu$, no n-tuple is shattered by $\mathcal{F}$)*

Notation change Let $\mathcal{A}$ be a collection of subsets of $\mathcal{X}$ and $\mathcal{F}$ is the set of indicator functions of sets in $\mathcal{A}$. $A \in \mathcal{A} \iff f_A \in \mathcal{F}, f_A(x) = 1_A(x)$ Then we may speak of VC-dimension of $\mathcal{A}$.

$$\mathcal{F}(x_1^n) \iff \mathcal{A}(x_1^n) = \Big\{A \cap x_1^n, A \in \mathcal{A}\Big\}$$

If $|\mathcal{A}(x_1^n)| = 2^n$, $\mathcal{A}$ picks out all subsets of coordinates of x_1^n

Examples

1) $\mathcal{F} = \Big\{1_{(-\infty,z]}, z\in\mathbb{R}\Big\} \iff \mathcal{A} = \Big\{(-\infty, z], z \in \mathbb{R}\Big\}$. $|\mathcal{A}(x_1^n)| \le n+1$. The VC-dimension is 1.

2) $\mathcal{A} = \Big\{(b,a], b < a\Big\}$. Observe that $|\mathcal{A}(x_1^n)| \leq (x+1)^2$ because each x_1^n splits $\mathbb{R}$ into $n+1$ intervals so we have up to (n+1) choices of for a and up to $(n+1)$ choices for b.

Suppose VC-dimension of $\mathcal{A}$ is ν. Then for $n > \nu$ $|\mathcal{A}(x_1^n)| < 2^n$ for all n-tuples x_1^n. Surprising result is that $|\mathcal{A}(x_1^n)|$ grows polynomially in n (polynomial discrimination).

Lemma 20.7 *If $\mathcal{A}$ has VC-dimension ν. Then for each $x_1^n \subset \mathcal{X}$*

$$\big|\mathcal{A}(x_1^n)\big| = \Big|\big\{A \cap x_1^n, A \in \mathcal{A}\big\}\Big| \leq \sum_{i=0}^{\nu} \binom{n}{i} \leq (n+1)^{\nu}$$

for $\forall n \geq 1$ and $\leq \left(\frac{en}{\nu}\right)^{\nu}$ for $n \geq \nu$.