LECTURE 27

EMPIRICAL PROCESSES

- Let X be a r.v. taking values in $\forall$ set $\mathcal{X}$, $X_1, \dots, X_n$ be independent copies of X.
- Law of large numbers $\Rightarrow$ $\forall$ Boolean function $f: \mathcal{X} \to \{0,1\}$:

$$\frac{1}{n}\sum_1^n f(X_i) \xrightarrow{a.s.} \mathbb{E}f(X) \quad \text{as } n \to \infty$$

- Deviation:

$$\mathbb{E}\left|\frac{1}{n}\sum_1^n f(X_i) - \mathbb{E}f(X)\right| \underset{L^1 \le L^2}{\le} \left(\mathbb{E}|\cdots|^2\right)^{1/2} = \mathrm{Var}\left(\frac{1}{n}\sum_1^n f(X_i)\right)^{1/2}$$

$$= \left[\frac{1}{n^2}\sum_1^n \underbrace{\mathrm{Var}(f(X_i))}_{\le 1 \ (f \text{ Boolean})}\right]^{1/2} = \frac{1}{\sqrt{n}} \quad \text{("weak LLN")}$$

- Is this true uniformly over all Boolean functions f?

$$\mathbb{E}\sup_{\forall f}\left|\frac{1}{n}\sum_1^n f(X_i) - \mathbb{E}f(X)\right| \to 0 \quad \text{as } n \to \infty \ ?$$

NO For $f = \mathbb{1}_{\{X_1,\dots,X_n\}}$, $\frac{1}{n}\sum_1^n f(X_i) = 1$ but $\mathbb{E}f(X) = 0$

- But it is true uniformly over $f \in \mathcal{F}$ whenever $\mathrm{vc}(\mathcal{F}) < \infty$:

THM (Uniform Law of Large Numbers)

If $\mathcal{F}$ is a Boolean class with $d = \mathrm{vc}(\mathcal{F}) < \infty$, then

$$\mathbb{E}\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)\right| \le C\sqrt{\frac{d\log n}{n}}.$$

Remarks ① Same rate $O(1/\sqrt{n})$ as in WLLN!

② $\left(\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}f(X)\right)_{f\in\mathcal{F}}$ is called an empirical process.

The proof uses a new tool:

The proof of the Uniform Law of Large Numbers is based on:

Symmetrization Lemma [Gine-Zinn]

Let $X, X_1, \dots, X_n$ be iid r.v's;
let $\varepsilon_1, \dots, \varepsilon_n$ be independent ± 1 with prob $1/2$.
Then $\mathbb{E}\left|\sum_{i=1}^n (X_i - \mathbb{E}X_i)\right| \le 2\,\mathbb{E}\left|\sum_{i=1}^n \varepsilon_i X_i\right|$

Proof

$$\mathbb{E}\left|\sum_i X_i - \mathbb{E}\left(\sum_i X_i\right)\right| = \mathbb{E}_X\left|\sum_i X_i - \mathbb{E}_{X'}\left(\sum_i X_i'\right)\right|$$

where X_i' are independent copies of X_i

$$= \mathbb{E}_X\left|\mathbb{E}_{X'}\left(\sum_i (X_i - X_i')\right)\right| \le \mathbb{E}_X \mathbb{E}_{X'}\left|\sum_i (X_i - X_i')\right|$$

Jensen ineq.

$X_i - X_i' \overset{\text{dist}}{=} X_i' - X_i$ by symmetry: $\varepsilon_i(X_i - X_i')$

$$= \mathbb{E}_X \mathbb{E}_{X'} \mathbb{E}_\varepsilon \left|\sum_i \varepsilon_i (X_i - X_i')\right| \le \mathbb{E}\left|\sum_i \varepsilon_i X_i\right| + \mathbb{E}\left|\sum_i \varepsilon_i X_i'\right|$$

$\mathbb{E}$; $\sum_i \varepsilon_i X_i + \sum_i \varepsilon_i X_i'$; same distr.

$$= 2\,\mathbb{E}\left|\sum_i \varepsilon_i X_i\right|. \qquad \text{QED.}$$

Proof of ULLN

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

$$\le 2 \cdot \frac{1}{n}\,\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|$$

"Rademacher complexity" of $\mathcal{F}$

by a "uniform" version of Symmetrization Lemma (DIY)

- Condition on r.v's X_i (treat them as fixed #'s). Randomness remains in ε_i.

Use Hoeffding inequality: $\forall f \in \mathcal{F}$,

$$\mathbb{P}\left\{ \frac{1}{n} \left| \sum_{i=1}^n \varepsilon_i f(x_i) \right| \geq t \right\} \leq 2\exp\left(- \frac{(nt)^2}{2\sum_{i=1}^n a_i^2} \right) \leq 2\exp\left(-\frac{nt^2}{2}\right).$$

($f(x_i) = a_i$; $a_i \in \{0,1\}$ (Boolean class))

- Restriction: $\sup_{f\in\mathcal{F}} \left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right|$ is determined by the values of $f \in \mathcal{F}$ on the sample $x_1, \dots, x_n$ only.

$\Rightarrow$ we can replace the class $\mathcal{F}$ by its restriction

$$\mathcal{F}_n := \{ f|_{\{x_1,\dots,x_n\}} : f \in \mathcal{F} \}.$$

- Sauer-Shelah Lemma $\Rightarrow$

$$|\mathcal{F}_n| \leq \sum_{k=1}^d \binom{n}{k} \leq (en)^d \quad \text{where } d = vc(\mathcal{F}).$$

- Union Bound:

$$\mathbb{P}\left\{ \sup_{f\in\mathcal{F}} \left| \frac{2}{n} \sum_{i=1}^n \varepsilon_i f(x_i) \right| \geq t \right\} \leq |\mathcal{F}_n| \cdot 2\exp\left(-\frac{t^2 n}{2}\right)$$

$$\leq 2\exp\left(d\log(en) - \frac{t^2 n}{2}\right)$$

$$\leq 2\exp\left(-\frac{t^2 n}{4}\right) \qquad \forall\, t \geq 10\sqrt{\frac{d\log n}{n}} =: t_0$$

(Z)

- Integral identity:

$$\mathbb{E}Z = \int_0^\infty \mathbb{P}\{Z \geq t\}\, dt \leq \int_0^{t_0} \mathbb{P}\{Z \geq t\}\, dt + \int_{t_0}^\infty \mathbb{P}\{Z \geq t\}\, dt$$

($\mathbb{P}\{Z \geq t\} \leq 1$)

$$\leq t_0 + \int_{t_0}^\infty 2\exp\left(-\frac{t^2 n}{4}\right) dt \leq 2t_0.$$

QED