Lecture 6

D Empirical Processes

Consider $X_1, \ldots, X_n$ generated iid μ on $(\mathcal{X}, \Sigma)$

Let $\mu_n(A) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}(X_i \in A)$, $A \in \Sigma$.

Then, $\mathbb{E}[\mu_n(A)] = \mu(A)$.

Furthermore, by Hoeffding $\mathbb{P}\left(|\mu_n(A) - \mu(A)| > \epsilon \right) \le 2 \cdot e^{-2n\epsilon^2}$.

Thus, by Borel-Cantelli lemma, $\mu_n(A) \xrightarrow{a.s.} \mu(A)$ $\forall A$.

<u>Does $\mu_n \to \mu$ in total variation distance?</u>

No. If μ is continuous, $\mu(\mathrm{supp}(\mu_n)) = 0$ $\forall n$.

$$d_{TV}(\mu_n, \mu) = \sup_{A} |\mu_n(A) - \mu(A)|$$

→ For continuous μ, $d_{TV}(\mu_n, \mu) = 1$ $\forall n$.

We shall be interested in $\boxed{\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)|}$

(i) If $\mathcal{A} = \{ (-\infty, x], x \in \mathbb{R} \}$,

$$\sup_{A \in \mathcal{A}} |\mu_n(A) - \mu(A)| = \sup_{x} \underbrace{|F_n(x) - F(x)|}_{\text{Kolmogorov-Smirnov distance b/w } \mu_n \text{ and } \mu}$$

$d_{KS}(\mu_n, \mu) \equiv$

<u>Theorem</u> (Glivenko-Cantelli)

$$d_{KS}(\mu_n, \mu) \to 0 \quad a.s.$$

<u>Proof.</u> Suffices to show : $\mathbb{P}\left(d_{KS}(\mu_n,\mu) \geq \delta_n+\epsilon\right) \leq e^{-nc\epsilon^2}$ for $\delta_n \to 0$.

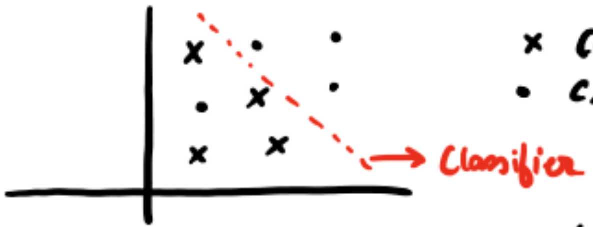
Note that for $\Delta_n(X_1,\dots,X_n) = d_{KS}(\mu_n,\mu)$, by changing only one of X_i we change $|F_n(x)-F(x)|$ for all $x \geq x_i$ by at most $\frac{1}{n}$. Thus, Δ_n satisfies BDP with $\left(\frac{1}{n},\dots,\frac{1}{n}\right)$.

By Mc Diarmid's inequality, we get

$$\mathbb{P}\left(d_{KS}(\mu_n,\mu) \geq \mathbb{E}[d_{KS}(\mu_n,\mu)]+\epsilon\right) \leq e^{-\frac{n\epsilon^2}{2}}.$$

We shall bound $\mathbb{E}[d_{KS}(\mu_n,\mu)]$.

(II) <u>Empirical risk minimization</u> (ERM)



Given a family of classifiers $\mathcal{F}$ (e.g. all linear/affine classifiers), find f^* which performs roughly as well as the best classifier in $\mathcal{F}$.

<u>Bayesian formulation:</u> $\quad P(X^n) = \theta P_1^n(X^n) + \bar{\theta} P_2^n(X^n)$

Find $i \in \{1,2\}$ by observing $X_1,\dots,X_n$.
(MAP is called naive Bayes)

A classifier is specified by an "acceptance" region A where you declare P_1.

<u>Practice</u>: θ, P_1, P_2 are unknown.

ERM procedure: Select $$f_n^* = \arg\min_{f \in \mathcal{F}} \mu_n\big(f(X) \neq Y\big)$$
$$= \arg\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}\big(f(X_i)\neq Y_i\big)$$
$$\equiv \arg\min_{A\in\mathcal{A}} \frac{1}{n}\sum_{i=1}^{n}\Big[\mathbb{1}\big(X_i \notin A, Y_i = 1\big) + \mathbb{1}\big(X_i \in A, Y_i \neq 1\big)\Big]$$

Vapnik and Chervonenkis showed that
$$\Big(\mu\big(f_n^*(X)\neq Y\big) - \min_{f\in\mathcal{F}} \mu\big(f(X)\neq Y\big)\Big)_+ \to 0 \quad a.s.$$

Proof: (a) McDiarmid for conc. (b) Bound for $\mathbb{E}\Big[\max_{A\in\mathcal{A}} |\mu_n(A) - \mu(A)|\Big]$

For $\mathcal{A}$, denote
$$\Delta_n(\mathcal{A}) := \sup_{A\in\mathcal{A}} \big|\mu_n(A) - \mu(A)\big|$$

Our goal is to bound $\mathbb{E}[\Delta_n(\mathcal{A})]$.

(I) <u>Method 1</u>: Using Vapnik Chervonenkis Theory
$$\mathbb{E}[\Delta_n(\mathcal{A})] = \mathbb{E}\Big[\sup_{A\in\mathcal{A}}\Big|\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i\in A\}} - \mu(A)\Big|\Big]$$

<u>Step 1. Symmetrization</u>

Let $X_1', \dots, X_n'$ be an indep. copy of $X_1, \dots, X_n$.

4

Then, conditioned on $X=(x_1,\ldots,x_n)$,

$$\sup_{A\in\mathcal{A}} \left| \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{\{x_i\in A\}} - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n}\mathbb{1}_{\{X_i'\in A\}}\right]\right|$$

$$\leq \sup_{A\in\mathcal{A}} \mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\{x_i\in A\}} - \mathbb{1}_{\{X_i'\in A\}}\right)\right|\right]$$

(Jensen's inequality)

$$\leq \mathbb{E}\left[\sup_{A\in\mathcal{A}}\left|\frac{1}{n}\sum_{i=1}^{n}\left(\mathbb{1}_{\{x_i\in A\}} - \mathbb{1}_{\{X_i'\in A\}}\right)\right|\right]$$

Therefore,

$$\mathbb{E}[\Delta_n(\mathcal{A})] \leq \mathbb{E}\left[\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n}\left(\mathbb{1}_{\{X_i\in A\}} - \mathbb{1}_{\{X_i'\in A\}}\right)\right|\right]$$

Step 2: Radamachar Sums

Let $\sigma_1,\ldots,\sigma_n$ be iid unif $\{-1,+1\}$.

Then, R.H.S. above equals

$$\mathbb{E}\Bigg[\underbrace{\sup_{A\in\mathcal{A}} \frac{1}{n}\left|\sum_{i=1}^{n}\sigma_i\left(\mathbb{1}_{\{X_i\in A\}} - \mathbb{1}_{\{X_i'\in A\}}\right)\right|}_{n^{-1}\cdot Z(X_1,\ldots,X_n,X_1',\ldots,X_n')}\Bigg]$$

$$\mathbb{E}[Z \mid X=x, X'=x']$$

$$= \mathbb{E}\left[\sup_{A\in\mathcal{A}}\left|\sum_{i=1}^{n}\sigma_i\left(\mathbb{1}_{\{x_i\in A\}} - \mathbb{1}_{\{x_i'\in A\}}\right)\right|\right] \quad (1)$$

Recall: $\mathbb{E}\left[\max_{1\le i\le n} X_i\right] \leq \sigma\sqrt{2\log n}$ if $X_i \in \mathcal{G}(\sigma^2)$

If we apply this bound naively to (1), we will be hit by $|\mathcal{A}|$, which may even be ∞. 5

Step 3: Shatter coefficients

Definition (Shatter coefficient, VC dimension)

The n^{th} shatter coefficient of $\mathcal{A}$, $S_{\mathcal{A}}(n)$, is given by $S_{\mathcal{A}}(n) = \sup_{x_1,\ldots,x_n} \left|\left\{\{x_1,\ldots,x_n\} \cap A,\ A \in \mathcal{A}\right\}\right|$.

The VC dimension of $\mathcal{A}$, denoted $V(\mathcal{A})$, is the largest n such that $S_{\mathcal{A}}(n) = 2^n$. (Sauer: $S_{\mathcal{A}}(n) \le (n+1)^{V(\mathcal{A})}$)

How does it apply to our quantity in (1)?

Note that the maximum number of distinct sequences

$$b(A) = \left(\mathbb{1}_{\{x_1 \in A\}}, \ldots \mathbb{1}_{\{x_n \in A\}}, \mathbb{1}_{\{x_1' \in A\}}, \ldots, \mathbb{1}_{\{x_n' \in A\}}\right) \in \{0,1\}^{2n}$$

that we see for a fixed (x, x') and different $A \in \mathcal{A}$ is bounded above by $S_{\mathcal{A}}(2n)$.

Thus, there exists $\hat{\mathcal{A}}$ with $|\hat{\mathcal{A}}| \le S_{\mathcal{A}}(2n)$ such that the right-side of (1) equals

$$\mathbb{E}\left[\sup_{A \in \hat{\mathcal{A}}} \left|\sum_{i=1}^n \sigma_i \left(\mathbb{1}_{\{x_i \in A\}} - \mathbb{1}_{\{x_i' \in A\}}\right)\right|\right]$$

$$\leq \sqrt{2v \log 2 S_{\mathcal{A}}(2n)} \qquad (2)$$

6

if

$$\sum_{i=1}^{n} \sigma_i \left(\mathbb{1}_{\{x_i \in A\}} - \mathbb{1}_{\{x_i' \in A\}} \right) \in \mathcal{G}(v). \quad (3)$$

By Hoeffding's lemma, (3) holds with $v = n$.

Therefore, (2) and (1) imply

$$\mathbb{E}[\Delta_n(\mathcal{A})] \leq \sqrt{\frac{2}{n} \log 2 S_{\mathcal{A}}(2n)}$$

Finally, note that $S_{\mathcal{A}}(2n) \leq S_{\mathcal{A}}(n)^2$ to get

$$\mathbb{E}[\Delta_n(\mathcal{A})] \leq \sqrt{\frac{4 \log S_{\mathcal{A}}(n) + 2 \log 2}{n}}$$

<u>Example</u>.

For $\mathcal{A} = \{(-\infty, \theta), \theta \in \mathbb{R}\}$, for any fixed sequence $(x_1, \dots, x_n) \in \mathbb{R}^n$,

$$|\{A \cap \{x_1, \dots, x_n\}, A \in \mathcal{A}\}| \leq n+1$$

since if we arrange $x_1, \dots, x_n$ in an increasing order, the only possible $b(\underline{x})$ are $(0, \dots, 0)$, $(1, 0, \dots, 0)$, $(1, 1, 0, \dots, 0)$, $\dots$, $(1, \dots, 1)$)

Thus, for this case, ⑦

$$\mathbb{E}[\Delta_n(A)] = \mathbb{E}[d_{KS}(\mu_n, \mu)]$$
$$\leq \sqrt{\frac{4\log(n+1) + 2\log 2}{n}}$$

Combining this with McDiarmid we get

$$\mathbb{P}\left(d_{KS}(\mu_n, \mu) > \sqrt{\frac{8\log n}{n}} + \epsilon\right) \leq e^{-\frac{n\epsilon^2}{2}}.$$

By using the Borel-Cantelli lemma:

$d_{KS}(\mu_n, \mu) \longrightarrow 0$ a.s. (Glivenko-Cantelli) ∎