LECTURE 38

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

$$= \mathbb{E}_X\left|\underbrace{\mathbb{E}_{X'}}_{\text{Jensen ineq.}}\left(\sum_i (X_i - X_i')\right)\right| \le \mathbb{E}_X\mathbb{E}_{X'}\left|\sum_i \underbrace{(X_i - X_i')}_{\overset{\text{dist}}{=}\ \varepsilon_i(X_i - X_i')}\right|$$

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

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

same distr.

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

Proof of ULLN (stated in lec. 37).

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

$$\le 2\cdot \underbrace{\frac{1}{n}\mathbb{E}\sup_{f\in\mathcal{F}}\left|\sum_{i=1}^n \varepsilon_i f(X_i)\right|}_{\text{"Rademacher complexity" of } \mathcal{F}}$$

by a 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 (lec): $\forall f \in \mathcal{F}$,

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

$a_i \in \{0,1\}$ (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:

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

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

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

- Integral identity:

$$\mathbb{E}Z = \int_0^\infty \mathbb{P}\{Z \geq t\}\,dt \leq \int_0^{t_0} \underbrace{\mathbb{P}\{Z \geq t\}}_{\leq 1}\,dt + \int_{t_0}^\infty \mathbb{P}\{Z \geq t\}\,dt$$

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

QED

Example (a): half-lines $\mathcal{F} = \{ \mathbb{1}_{(-\infty, a]} : a \in \mathbb{R} \}$. $vc(\mathcal{F}) = 1$.

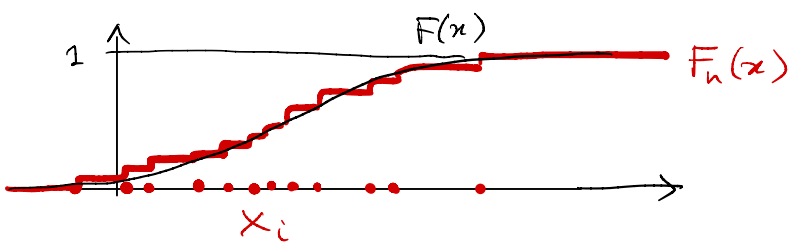
$\forall f \in \mathcal{F}$:

$\mathbb{E} f(X) = \mathbb{P}\{X \le a\} =: F(a)$, the cumulative distribution function (CDF) of X

$\frac{1}{n}\sum_i^n f(X_i) = \frac{\#\{X_i \le a\}}{n} =: F_n(a)$, the empirical CDF of X.

$\Rightarrow$ ULLN yields $\mathbb{E} \sup_{a \in \mathbb{R}} |F_n(a) - F(a)| \le C\sqrt{\frac{\log n}{n}}$, i.e.:

> **Glivenko-Cantelli Thm** $\forall$ random variable X, the CDF of X can be uniformly estimated from a sample of n pts:
>
> $$\mathbb{E}\|F_n - F\|_\infty \le C\sqrt{\frac{\log n}{n}}.$$



All quantiles of X can be estimated from one sample $X_1, \dots, X_n$.

Example (b): axis-aligned rectangles $\mathcal{F} = \{ \mathbb{1}_R : R = [a,b] \times [c,d] \}$

$vc(\mathcal{F}) = 4$

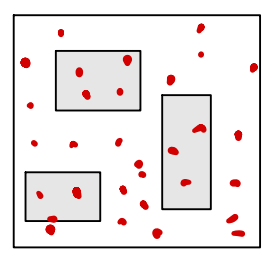
let $X \sim \mathrm{Unif}\,[0,1]^2 \Rightarrow$

$\mathbb{E} f(X) = \mathrm{Area}(R)$; $\frac{1}{n}\sum_{i=1}^n f(X) = \frac{\#\{X_i \in R\}}{n} \xrightarrow{\text{ULLN}}$

$$\boxed{\mathbb{E} \sup_R \left| \frac{\#\{X_i \in R\}}{n} - \mathrm{Area}(R) \right| \le C\sqrt{\frac{\log n}{n}}}$$

uniformly over all rectangles

discrepancy



Like in lec.

Example (c): A similar result holds for other common shapes: triangles, circles, ellipses, etc. since $vc(\mathcal{F}) < \infty$.

BACK TO MACHINE LEARNING

- Recall ERM from lec. 35: we want to predict Y from X.

Risk (test error): $R(h) := \mathbb{E}\,\ell(h(X), Y)$, $\qquad h^* = \operatorname{argmin}_{h \in \mathcal{H}} R(h)$.

Empirical risk (training error): $R_n(h) := \frac{1}{n} \sum_{i=1}^{n} \ell(h(X_i), Y_i)$, $\qquad h_n^* := \operatorname{argmin}_{h \in \mathcal{H}} R_n(h)$

- Generalization bound:

$$R(h_n^*) \le R(h^*) + 2 \sup_{h \in \mathcal{H}} \underbrace{|R_n(h) - R(h)|}_{\parallel}$$

$$\frac{1}{n} \sum_{i=1}^{n} \underbrace{\ell(h(X_i), Y_i)}_{\parallel} - \mathbb{E}\,\ell(h(X), Y)$$

$f(X_i)$, apply ULLN $\Rightarrow$

THM (VC Generalization bound)

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

$$\mathbb{E}\, R(h_n^*) \le R(h^*) + C \sqrt{\frac{d \log n}{n}}$$

generalization error

$\Rightarrow$ generalizes well from a sample of size

$$\boxed{n \sim vc(\mathcal{H})}$$

Remark: $\log n$ can be removed (see my book)