Lecture 15: Empirical Risk Minimization

Isabella Zhu

3 April 2025

Reading: Non-Asymptotic Stats sections 4.1-4.3.

§1 Empirical Process Theory

Recall that the setup is as follows: we have a class of functions $f \in F$ and data $\{X_i\}_{i=1}^n$ which is (usually) i.i.d but not always. Define the **empirical process** as $\{Z_f | f \in F\}$

$$f \to Z_f := \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]$$

Glivenko-Cantelli class refers to where

$$\sup_{f\in F} |Z_f| \to 0$$

as $n \to 0$. Last class, we showed that $F = \{f_t | t \in \mathbb{R}\}$ is Glivenko-Cantelli.

Theorem 1.1

(Donsker). We have that

$$\sqrt{n} Z_f \to N(0, \text{var}(f(x)))$$

which generalizes CLT, the same way Glivenko-Cantelli generalizes LLN.

Remark 1.2. We will *not* be going over Donsker in class.

Theorem 1.3

Suppose F is uniformly b-bounded, which means that

$$||f||_\infty = \sup_x |f(x)| \leq b$$

We proved last time that

$$\sup_{f\in F} |Z_f| \leq 2\mathbb{E}_X \mathbb{E}_\epsilon \left[\sup_{f\in F} \left| \frac{1}{n}\sum_{i=1}^{n} \epsilon_i f(X_i) \right| + \delta \right]$$

with probability $1 - e^{-n\delta^2/2b^2}$ where ϵ are the independent Rademacher RVs.

Proof. We first show concentration around the mean. We have

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

which satisfies the bounded difference property with parameter $\frac{b}{n}$.

We plug into bounded difference inequality to get the result $1 - e^{-n\delta^2/2b^2}$.

Next, we follow with a symmetrization argument. Note that we didn't exploit any structure regarding F. We showed instead that

$$\mathbb{E}[\sup_{f\in F}|Z_f|] \leq 2R_n(F)$$

where in class we defined $R_n(F)$ to be

$$R_n(F) = \mathbb{E}_X\mathbb{E}_\epsilon\left[\sup_{f\in F}\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(X_i)\right| + \delta\right]$$

This is when we exploit properties of F to calculate tighter bounds.

§2 Empirical Risk Minimization

We will talk about prediction problems specifically. We have two types of random variables, $x \in X$ (features, covariates) and $y \in \mathbb{R}$ (responses, outputs). A **predictor** is a function g that outputs $\hat{y} = g(x) \approx y$.

We have some loss function

$$L : x \times y \to \mathbb{R}, \;\; (x, y) \to L(g(x), y) \in \mathbb{R}$$

One example of loss is the least-squares loss, where

$$L(g(x), y) = (y - g(x))^2$$

Observe i.i.d samples from an unknown $\mathbb{P}$ on $X \times Y$. Samples are $Z_i = (X_i, Y_i)$. Study the procedure

$$\hat{g} \in \arg\min_{g\in G}\frac{1}{n}\sum_{i=1}^{n} L(g(X_i), Y_i) \sim \hat{R}_n(g)$$

Ideal object: define **population risk** as

$$\bar{R}(g) := \mathbb{E}[L(g(X), Y)]$$

which is nonrandom. Define

$$g^+ \in \arg\min_{g\in G}\bar{R}(g)$$

The **excess risk** is

$$\bar{R}(\hat{g}_n) - \inf_{g\in G}\bar{R}(g)$$

We care about (i) if excess risk goes to 0 and (ii) how quickly it goes to 0.

Define the G-induced loss class

$$L \circ G = \{(x, y) \to L(g(x), y)\}$$

Proposition 2.1

We have that

$$\mathbb{E}_{X_1^n,Y_1^n}[\bar{R}(\hat{g}_n)] - R(g^+) \leq 4R_n(L \circ G(x_1^n, y_1^n)) = 4\mathbb{E}_{X_1^n,Y_1^n}\mathbb{E}_{\epsilon_1^n} \sup_{g\in G} \left| \frac{1}{n}\sum_{i=1}^n \epsilon_i L(g(x_i), y_i) \right|$$

Note that $\bar{R}(\hat{g}_n)$ is a random variable because $\hat{g}_n$ is random.

Proof. We have

$$\bar{R}(\hat{g}_n) - \bar{R}(g^+) = \overbrace{\{\bar{R}(\hat{g}_n) - \hat{R}_n(\hat{g}_n)\}}^{T_1} + \overbrace{\{\hat{R}_n(\hat{g_n}) - \hat{R}_n(g^+)\}}^{T_2} + \overbrace{\{\hat{R}_n(g^+) - \bar{R}(g^+)\}}^{T_3}$$

Note that $T_2 \leq 0$ by definition and $T_3 \to 0$ by LLN (since this doesn't depend on the data). The interesting character is T_1, which measures overfitting. We have that

$$\bar{R}(\hat{g}_n) - \bar{R}(g^+) \leq 2\mathbb{E}_{X_1^n,Y_1^n} \sup_{g\in G} \left| \frac{1}{n}\sum_{i=1}^n L(g(X_i), Y_i) - \mathbb{E}[L(g(X), Y)] \right|$$

because we have $\hat{R}(g) = \frac{1}{n}\sum_{i=1}^n L(g(X_i), Y_i)$ and $\bar{R}(g) = \mathbb{E}[L(g(X), y)]$. We have essentially upper bounded each of T_1 and T_2. We get the final result by using the same symmetry argument as earlier.