Lecture 16: More ERM

Isabella Zhu

8 April 2025

§1 ERM

As a reminder (yet again), for prediction problems (X, Y), goal is to minimize loss over datapoints

$$\hat{g}_n \in \operatorname{argmin} \frac{1}{n} \sum_{i=1}^{n} L(g(X_i), Y_i)$$

As n gets large, converges to

$$g^+ = \operatorname*{argmin}_{g \in G} \mathbb{E}[L(g(X), Y))] \triangleq \bar{R}(g)$$

We care about bounding

$$\mathbb{E}\bar{R}(\hat{g}_n) - \bar{R}(g^+)$$

We can also define a function g^* that might not even be in G

$$g^* = \arg\min_g \bar{R}(g)$$

So we would have

$$\bar{R}(g^*) \leq \bar{R}(g^+) \leq \bar{R}(\hat{g}_n)$$

§1.1 Estimation and Approximation Error

Definition 1.1. $\bar{R}(g^+) - \bar{R}(g^*)$ is called **approximation error** and is nonrandom and measures how rich G is.

Definition 1.2. $\bar{R}(\hat{g}_n) - \bar{R}(g^+)$ is called **estimation error**. Estimation error is random. This is where Rademacher complexity comes in.

$$\mathbb{E}[\bar{R}(\hat{g}_n)] - \bar{R}(g^*) = \text{approximation error} + \text{estimation error}$$

We know that

$$\begin{aligned}\mathbb{E}[\bar{R}(\hat{g}_n)] - \bar{R}(g^*) &= \{\mathbb{E}[\bar{R}(\hat{g}_n)] - \bar{R}(g^+)\} + \{\bar{R}(g^+) - \bar{R}(g^*)\} \\ &\leq 2\mathbb{E}\left[\sup_{g \in G} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i L(g(X_i), Y_i)\right] + \min_{g \in G}(\bar{R}(g) - \bar{R}(g^*))\end{aligned}$$

§1.2 Sieves and Regularization

A **sieve** is a chain of function classes

$$G_1 \subseteq G_2 \subseteq G_3 \dots$$

The approximation term $\min_{g \in G_k} \bar{R}(g) - R(g^*)$ is non-increasing. In many cases, however, this should be decreasing at a nice rate (we should choose a sieve that makes the error goes to 0).

§2 Tools for Rademacher Complexity

The essential problem becomes bounding

$$R'_n(\Pi) = \mathbb{E}_\epsilon \sup_{t \in \Pi} \left[\frac{1}{n} \sum_{i=1}^{n} \epsilon_i t_i \right]$$

and

$$R_n(\Pi) = \mathbb{E}_\epsilon \sup_{t \in \Pi} \left[\left| \frac{1}{n} \sum_{i=1}^{n} \epsilon_i t_i \right| \right]$$

where $t_i = L(g(X_i), Y_i)$ and $\Pi = \{t_1, \dots, t_n\} \subseteq \mathbb{R}^n$.

§2.1 Readings (Ch 4)

Vapnik-Chernovenkis Theory: VC theory applies when $t \in \{0,1\}^n \subseteq$ boolean hypercube.

§2.2 Ledoux-Talagrand Contraction

Applies to Lipschitz functions, which means we have

$$|\phi_i(t_i) - \phi_i(\tilde{t}_i)| \leq L|t_i - \tilde{t}_i|$$

Intuitively, Lipschitz functions lets us get rid of the L and Y_i in $\epsilon_i L(g(X_i), Y_i)$ so we only look at $\epsilon_i g(X_i)$.

Define a **contraction** as $\phi(\Pi) = \{(\phi_1(t_1), \dots \phi_n(t_n)) \mid t \in \Pi\}$.

Lemma 2.1

We have

(a) $R'_n(\phi(\Pi)) \leq L R'_n(\Pi)$ for any L-Lipschitz ϕ.

(b) $R_n(\phi(\Pi)) \leq 2L R_n(\Pi)$

Proof (a). It suffices to show

$$nR'_n((\phi_1(t_1), \dots, \phi_n(t_n) \mid t \in \Pi) \leq nR'_n((Lt_1, \phi_2(t_2), \dots, \phi_n(t_n)) \mid t \in \Pi)$$

We have

$$\mathbb{E}_{\epsilon_2^n} \left[\frac{1}{2} \sup_{t \in \Pi} (\phi(t_1) + S(t_2^n)) + \frac{1}{2} \sup_{t \in \Pi} (-\phi(t_1) + S(\tilde{t_2}^n)) \right]$$

$$= \frac{1}{2}\mathbb{E}_{\epsilon_2^n}\left[\sup t, \tilde{t}\{\phi(t_1) - \phi(\tilde{t}_1) + S(t_2^n) + S(\tilde{t}_2^n)\}\right]$$

We know that $\phi(t_1) - \phi(\tilde{t}_1) \leq L|t_1 - \tilde{t}_1|$ by Lipshictz. In fact, we can say $\phi(t_1) - \phi(\tilde{t}_1) \leq L(t_1 - \tilde{t}_1)$ by symmetry. This implies the result.