Lecture 20: ERM with Least Squares

Isabella Zhu

22 April 2025

Reading: chapter 13, Wai 19.

§1 ERM with Least-Squares

We have (X_i, Y_i) pairs i.i.d. We have population risk

$$\bar{R}(f) = \mathbb{E}_{X,Y}(Y - f(X))^2$$

where $x \to f(x)$ is hopefully a reasonable approximation of y. We have

$$\hat{g} = \operatorname*{argmin}_{g \in G} \frac{1}{n} \sum_{i=1}^{n} (y_i - g(x_i))^2$$

for some function class G.

We can analyze this in a fixed design setting, assuming that $\{X_i\}$ are fixed deterministic (for example: timeseries). Then, we know that

$$y_i = f^*(x_i) + w_i$$

where f^* is the regression function

$$f^*(x) = \mathbb{E}[Y|X = x]$$

and we have that weights are zero-mean, i.e. $\mathbb{E}[w_i|x_i] = 0$. We can define

$$g^\dagger = \operatorname*{argmin}_{g \in G} \frac{1}{n} \sum_{i=1}^{n} (g(x_i) - f^*(x_i))^2 \triangleq ||g - f^*||_n^2$$

The value $||g^\dagger - f^*||_n^2$ is deterministic (for non-random x_i) and is called the **approximation error**. The value $||\hat{g} - g^\dagger||_n^2$ is the **estimation error** and is random.

§2 Localization and LS Over Coverings

§2.1 Covering Estimator

Let F be the original function space that f^* belongs to. Choose $G = \{g^1, \ldots g^N\} \subset F$ where we are choosing G to be a δ-covering in the $||.||_n$-norm.

Then, there is some $g^j = g^\dagger$ that is closest to f^*, which is less than δ away by construction. The procedure is to return

$$\hat{g} = \underset{g^1, g^2, \ldots g^N}{\operatorname{argmin}} = \frac{1}{n} \sum_{i=1}^{n} ||y - g||_n^2$$

Here, we have

$$\log N = \log N(\delta; F, ||.||_n)$$

which is the δ-covering number.

Lemma 2.1

We have

$$||\hat{g} - f^*||_2^2 \leq \overbrace{4\langle w, \hat{g} - g^\dagger \rangle_n - \frac{1}{2}||\hat{g} - g^\dagger||_n^2}^{\text{estimator error (localized)}} + 3||g^\dagger - f^*||_n^2$$

Lemma 2.2

For any finite set $\{\Delta^1, \ldots \Delta^N\} \subseteq \mathbb{R}^n$ with w_i 1-subgaussian i.i.d, we have

$$\mathbb{E}\left[\max_{j=1,\ldots N} \langle w, \Delta^j \rangle - \gamma ||\Delta||_n^2\right] \leq \frac{1}{2\gamma} \frac{\log N}{n}$$

Proof. We have

$$\begin{aligned} \mathbb{E}\left[\max_{j=1,\ldots N} \langle w, \Delta^j \rangle - \gamma ||\Delta^j||_n^2\right] &= \frac{1}{\lambda} \mathbb{E}\left[\max_{j=1,\ldots,N} \log \exp\left\{\lambda \langle w, \Delta^j \rangle - \lambda\gamma ||\Delta^j||_n^2\right\}\right] \\ &\leq \frac{1}{\lambda} \log \mathbb{E}\left(\sum_{j=1}^{N} e^{\lambda \langle w, \Delta^j \rangle - \lambda\gamma ||\Delta^j||_n^2}\right) \quad \text{(Jensen)} \end{aligned}$$

We have that

$$\mathbb{E}\left[\exp\left(\frac{\lambda}{n} \sum_{i=1}^{n} w_i \Delta^j(x_i)\right)\right] \leq e^{\frac{\lambda^2}{2n} ||\Delta^j||_n^2}$$

by subgaussianity. So, we can factor out

$$\frac{1}{\lambda} \log \mathbb{E}\left(\sum_{j=1}^{N} e^{\lambda \langle w, \Delta^j \rangle - \lambda\gamma ||\Delta^j||_n^2}\right) \leq \frac{1}{\lambda} \log \mathbb{E}\left(\sum_{j=1}^{N} e^{(\lambda^2/2n - \lambda\gamma) ||\Delta^j||_n^2}\right)$$

We can choose $\lambda = 2\gamma n$ to get the desired result.

Theorem 2.3

Suppose each w_i is 1-subgaussian. Then,

$$\mathbb{E}||\hat{g} - f^*||_n^2 \leq c\left\{\delta^2 + \frac{\log N(\delta; F, ||.||_n)}{n}\right\}$$

where c is some universal constant.

Proof. We have

$$\mathbb{E}||\hat{g} - f^*||_n^2 \leq \mathbb{E}\left[4 \max_{j=1,\dots N} \langle w, g^j - g^+ \rangle - \frac{1}{2}||g^j - g^\dagger||_n^2\right] + 3 \underbrace{||g^\dagger - f^*||_n^2}_{\leq \delta^2 \text{ by construction}}$$
$$\leq c' \frac{\log N}{n} + 3\delta^2$$

where the first inequality follows from Lemma 1 and the second from Lemma 2.

Example 2.4

(Parametric entropy) For parametric function class, we have $\log N(\delta) \approx d \log(1/\delta)$. Recall that last week, global Rademacher bounds gives $\sqrt{d/n \cdot \log n/d}$, plus Dudley give us $\sqrt{d/n}$.

For the covering estimator, we solve

$$\delta^2 \approx \frac{d \log(1/\delta)}{n}$$

We will obtain the bound

$$\frac{d}{n} \log \frac{n}{d} << \sqrt{\frac{d}{n} \log \frac{n}{d}}$$

Example 2.5

Consider Lipschitz functions $f : [0, 1] \to \mathbb{R}$. We have

$$\log N(\delta, F, ||.||_\infty) \approx \frac{1}{\delta}$$

Note the different norms!! $||.||_\infty$ and $||.||_n$ are in fact different norms. However,

$$||f - g||_n \leq ||f - g||_\infty$$

Intuitively, $||f - g||_\infty$ is a much stronger norm. Thus, we have

$$\log N(\delta; F, ||.||_n) \leq \log N(\delta; F, ||.||_\infty) \lesssim \frac{1}{\delta}$$

Recall global Rademacher bounds gives

$$\delta \approx \sqrt{\log N(\delta)/n}, \quad \delta_n = (1/n)^{1/3}$$

while our theorem for the covering estimator gives $\lesssim \delta_n^2 = (1/n)^{2/3}$ which is optimal over all estimators.