Modern Topics in Statistical Learning Theory Spring 2023

Lecture 3 — Feb. 10, 2023

Prof. Qi Lei *Scribe: Jiajing Chen, Zhengyao Gu, Huyen Nguyen*

1 Last lecture's recap

In the previous lecture, we talked about the conditions for the following inequality to hold:

$$\text{w.h.p. (with high probability)} \geq 1-\delta,\ \left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}[x]\right| \leq \epsilon$$

where ϵ depends on δ and n, i.e. $\epsilon = f_n(\delta)$.

2 Today lecture's topic

In this lecture, we will apply the results from the last lecture to analyze the difference between the population risk $L(\theta)$ and the empirical risk $\hat{L}_n(\theta)$. With $l(h_\theta(x_i), y_i)$ as the per-example loss function between the prediction function $h_\theta(x)$ with parameter θ and y, the empirical and population risk are defined as

$$\hat{L}_n(\theta) := \frac{1}{n}\sum_{i=1}^{n} l(h_\theta(x_i), y_i), \tag{1}$$

$$L(\theta) := \mathop{\mathbb{E}}_{x,y\sim\mathcal{P}(x,y)} l(h_\theta(x), y) \tag{2}$$

respectively.

Specifically, we will show uniform convergence of the empirical risk:

$$\text{w.h.p., } 1-\delta : \sup_{\theta\in\Theta} |\hat{L}_n(\theta) - \text{Ł}(\theta)| \leq \epsilon, \tag{3}$$

where Θ is the set of all prediction functions.
For some examples of what the parametrized prediction function looks like: define the *linear regression prediction function*

$$h_\theta : x \mapsto \theta^T x, \theta \in \Theta = \mathbb{R}^d \tag{4}$$

where $\Theta = \{\mathbf{v} \in \mathbb{R}^d \mid \|\mathbf{v}\| \leq 1\}$.

h_θ can also be a two-layered Neural Network (NN):

$$h_\theta : x \mapsto \sum_{i=1}^{m} a_i \sigma(w_i^T x), \tag{5}$$

where $a_i \in \mathbb{R}^m$, $w_i \in \mathbb{R}^{md}$, and σ is the activation function. A popular choice for σ is the Rectified Linear Unit (ReLU):

$$\sigma(x) = \text{ReLU}(x) = (x)_+ = \begin{cases} x, & x \leq 0 \\ 0, & \text{otherwise} \end{cases} \tag{6}$$

3 Motivation

In this lecture we are interested in how the empirical risk $\hat{L}_n$ generalizes to the population risk, i.e. we would like the magnitude of the *excess risk* $R(\hat{\theta}) = \hat{L}_n(\hat{\theta}) - L(\hat{\theta})$ to be small, where $\hat{\theta} = \underset{\theta \in \Theta}{\text{argmin}} \hat{L}_n(\theta)$. We prove bounds on the size of the excess risk using **uniform convergence**.

First note that we can decompose the excess risk by the **telescoping**.

$$R(\hat{\theta}) = \underbrace{L(\hat{\theta}) - \hat{L}_n(\hat{\theta})}_{①} + \underbrace{\hat{L}_n(\hat{\theta}) - \hat{L}_n(\theta^* s)}_{②} + \underbrace{\hat{L}_n(\theta^*) - L(\theta^*)}_{③}, \tag{7}$$

where $\theta^* = \underset{\theta \in \Theta}{\text{argmin}} L(\theta)$ and $\hat{\theta} = \underset{\theta \in \Theta}{\text{argmin}} \hat{L}_n(\theta)$. ① is the excess risk with respect to the empirical risk minimizer $\hat{\theta}$, and ③ is the population risk with respect to the optimal estimator θ^*, whereas, by definition,

$$② := \hat{L}_n(\hat{\theta}) - \hat{L}_n(\theta^*) \leq 0. \tag{8}$$

Also note that

$$|L(\theta) - \hat{L}_n(\theta)| \leq \epsilon, \forall \theta \in \Theta \tag{9}$$

$$\Rightarrow |L(\hat{\theta}) - \hat{L}_n(\hat{\theta})| \leq \epsilon \tag{10}$$

$$\Leftrightarrow \underset{\theta \in \Theta}{sup} |L(\theta) - \hat{L_n}(\theta)| \leq \epsilon \tag{11}$$

With respect to equation 7, we then have:

$$R(\hat{\theta}) := \underbrace{L(\hat{\theta}) - \hat{L}_n(\hat{\theta})}_{①} + \underbrace{\hat{L}_n(\hat{\theta}) - \hat{L}_n(\hat{\theta^*})}_{②} + \underbrace{\hat{L}_n(\theta^*) - L(\theta^*)}_{③} \tag{12}$$

$$\leq |L(\hat{\theta}) - \hat{L}_n(\hat{\theta})| + |\hat{L}_n(\theta^*) - L(\theta^*)| \tag{13}$$

$$\leq 2 \times \sup_{\theta \in \Theta} |L(\theta) - \hat{L}_n(\theta)| \tag{14}$$

The inequality above showed that excess risk is bounded by two times the uniform convergence gap, which admits the intuitive interpretion — **"uniform convergence implies generalization"**.

4 Generalization Bound for Finite Hypothesis Classes

In this section, we demonstrate how one can derive generalization bounds from concentration theorems when assuming a finite hypothesis space. We stop using parameters $\theta \in \Theta$ to denote the range of prediction functions available to the optimization. Instead we use the non-parametric notation $h \in H$, where H is the *hypothesis space.*

Theorem 1 (Union Bound)**.** *For a series of events $E_1, E_2, \ldots, E_K$,*

$$\mathbb{P}r(\bigcup_{i=1}^{K} E_i) \leq \sum_{k=1}^{K} \mathbb{P}r(E_i). \tag{15}$$

Apply the union-bound claim to $\underset{\theta \in \Theta}{sup} |L(\theta) - \hat{L}_n(\theta)|$, we have:

$$\mathbb{P}r(|L(\theta) - \hat{L}_n(\theta)| \geq \epsilon, \forall \theta \in \Theta) \tag{16}$$

$$\leq \sum_{\theta \in \Theta} \mathbb{P}r(|L(\theta) - \hat{L}_n(\theta)| \geq \epsilon). \tag{17}$$

The next theorem is derived from the Hoeffding concentration theorem from the last lecture. It demonstrates how combining concentration theorems and union bound, one can derive generalization bounds.

Theorem 2. *Suppose $H \leq \infty$, l is bounded in [0,1], i.e., $0 \leq l(h(x), y) \leq 1 \quad \forall h \in H, \forall x, y$. Then $\forall \delta$ such that $0 \leq \delta \leq \frac{1}{2}$, we have, w.h.p. $\geq 1 - \delta$,*

$$|L(h) - \hat{L}_n(h)| \leq \sqrt{\frac{\ln|H| + \ln(\frac{2}{\delta})}{2n}} \tag{18}$$

Corollary 3. *As a corollary, we have the up-scaling for generalization bound $|L(\hat{h}) - L(h^*)|$ as follows:*

$$|L(\hat{h}) - \hat{L}_n(\hat{h})| \leq \sqrt{\frac{2(ln|H| + ln(\frac{2}{\delta}))}{n}} \tag{19}$$

Proof Sketch.

1. Use concentration inequality to prove the bound for each fixed $h \in H$.
2. Use union bound across all $h \in H$.
   (a) Fix an $\epsilon > 0$, apply Hoeffding's inequality on $l(h(x_i), y_i) \in [0, 1]$ as follows, with respect to $l(h(x_i), y_i) \in [a_i, b_i]$, that is $a_i = 0, b_i = 1$.

$$\mathbb{P}r(|\hat{L}_n(h) - L(h)| \geq \epsilon) \quad \text{(concentration inequality)} \tag{20}$$

$$\leq 2\exp\left(-\frac{2n^2\epsilon^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right) \tag{21}$$

$$= 2\exp(-2n\epsilon^2). \tag{22}$$

(b) Apply union bound for each event:

$$E_h := |\hat{L}_n(h) - L(n)| \geq \epsilon \tag{23}$$

$$\mathbb{P}r(E_h) \leq 2\exp(-2n\epsilon^2) \tag{24}$$

Thus, we have:

$$\mathbb{P}r(\exists h \in H, s.t. |\hat{L}_n(h) - L(h)| \geq \epsilon) \tag{25}$$

$$\leq \sum_{h \in H} \mathbb{P}r(E_h) \tag{26}$$

$$\leq |H| \times 2\exp(-2n\epsilon^2) \quad \text{(generalization bound)} \tag{27}$$

Using the notation from lecture 1 we derive the specific form of $\epsilon = f_n(\delta)$ (see Section 1):

$$\delta = |H| 2\exp(-2n\epsilon^2) \Leftrightarrow \epsilon = \sqrt{\frac{\ln|H| + \ln(\frac{2}{\delta})}{2n}} \tag{28}$$

□

5 Generalization Bound for Infinite Hypothesis Class

The Union Bound theorem can only be applied to a countable set of events $E_1, E_2, \ldots$. One cannot apply the union bound to broadcast the concentration bound to all parameters when the hypotheses space is infinite. In the infinite case, one can construct a finite collection of balls whose union covers the whole hypothesis space, given that the hypotheses space is bounded:

$$H = \{h_\theta : \theta \in \mathbb{R}^d, ||\theta||_2 \leq B\}. \tag{29}$$

One then derive an error bound for each ball and apply union bound to pool all the error bounds into the final generalization bound.

Now, define the notion of an ϵ-cover.

Definition 4 (ϵ-cover)**.** *Let $\epsilon > 0$. An ϵ-cover of a set S with respect to a distance metric ρ is a subset $C \subseteq S$, s.t. $\forall x \in S$, $\exists x_0 \in C$ s.t. $\rho(x, x_0) \leq \epsilon$, or equivalently:*

$$S \subseteq \bigcup_{x_0 \in C} Ball(x_0, \epsilon, \rho) \tag{30}$$

$$Ball(x_0, \epsilon, \rho) := \{x : \rho(x, x_0) \leq \epsilon\} \tag{31}$$

The following lemma establishes that for a bounded set of hypotheses, one can construct an ϵ-cover with finite cardinality.

Lemma 5 (ϵ-cover of an l_2 ball)**.** *Let $B, \epsilon > 0$ and $S = \{x \in \mathbb{R}^p \mid ||x||_2 \leq B\}$. Then there exists an ϵ-cover of S with respect to the l_2 norm with*

$$|S_\epsilon| \leq \max\left\{\left(\frac{3B\sqrt{p}}{\epsilon}\right)^p, 1\right\} \tag{32}$$

With Lemma 5, we can demonstrate how one would derive generalization bounds for infinite hypothesis classes.

Proof Sketch.

1. Find ϵ-cover (ϵ-net) $S\epsilon$ of H.
2. Prove uniform convergence over all ϵ-balls of S_ϵ.
   Similar to the previous section, we have:

$$|L(\theta_0) - \hat{L}_n(\theta_0)| \leq \eta \quad \forall \theta_0 \in S_\epsilon \tag{33}$$

3. For $\theta \in \Theta$, find the closest $\theta_0 \in S_\epsilon$, with some Lipschitz condition of $l(h_\theta(x), y)$, we have:

$$|L(\theta) - L(\theta_0)| \leq \eta_2, \text{ when } ||\theta - \theta_0|| \leq \epsilon \tag{34}$$

$$\tag{35}$$

   besides,

$$|\hat{L}_n(\theta) - \hat{L}_n(\theta_0)| \leq \eta_2 \tag{36}$$

   Together, we again derive the generalization bound by telescoping:

$$|L(\theta) - \hat{L}_n(\theta)| \leq |L(\theta) - L(\theta_0)| + |L(\theta_0) - \hat{L}_n(\theta_0)| + |\hat{L}_n(\theta_0) - \hat{L}_n(\theta)|| \tag{37}$$

□

Detailed proof.

1. By Lemma 5 we can construct an ϵ-cover of the hypotheses space.
2. Fix η, ϵ For ϵ-cover, it satisfies:

$$\begin{aligned} \mathbb{P}r(|\hat{L}_n(\theta_0) - L(\theta_0)| \leq \eta, \forall Ball(\theta_0, \epsilon, \rho) \in S_\epsilon) &\geq 1 - 2|S_\epsilon| \exp(-2n\eta^2) && (38) \\ &= 1 - 2\exp(\ln |S_\epsilon| - 2n\eta^2) && (39) \\ &\geq 1 - 2\exp(p \cdot \ln(3BP/\epsilon)) - 2n\eta^2) && (40) \end{aligned}$$

   Step (40) is derived by applying Equation (32).

3. [κ-Lipschitz] Let $\kappa > 0, || \cdot ||$ be a norm on the domain D. A function $L : D \to \mathbb{R}^d$ is called κ-Lipschitz with respect to $|| \cdot ||$, if $\forall \theta, \theta' \in D$, we have $L(\theta) - L(\theta')| \leq \kappa ||\theta - \theta'||$.

   If l is κ-Lipschitz, then we know $\hat{L}_n$ and L are κ-Lipschitz, i.e. if $|\theta - \theta_0| \leq \epsilon$,

$$|L(\theta) - L(\theta_0)| \leq \kappa \cdot |\theta - \theta_0| \leq \kappa\epsilon \tag{41}$$

$$|\hat{L}_n(\theta) - \hat{L}_n(\theta_0)| \leq \kappa \cdot \epsilon \tag{42}$$

Then, we have

$$|L(\theta) - \hat{L}_n(\theta)| \leq |L(\theta) - L(\theta_0)| + |L(\theta_0) - \hat{L}_n(\theta_0)| + |\hat{L}_n(\theta_0) - \hat{L}_n(\theta)| \tag{43}$$
$$= 2\kappa \cdot \epsilon + \eta \tag{44}$$

Plug $\epsilon = \frac{\eta}{2\kappa}$ and $\eta = \sqrt{\frac{36 \cdot \ln(\kappa B n)}{n}}$ into inequality (43) and (40), then we have

$$|L(\theta) - \hat{L}_n(\theta)| \leq 2\eta \tag{45}$$

and

$$\mathbb{P}r(|\hat{L_n}(\theta) - L(\theta)| \leq 2\eta, \forall \theta \in \Theta) \geq 1 - 2\exp(-p). \tag{46}$$

Additionally, note that:

$$|\hat{L}_n(\theta) - L(\theta)| \leq \mathcal{O}\left(\sqrt{\frac{p \cdot \ln(\kappa B n)}{n}}\right) \tag{47}$$

for large enough n.

□