36-710: Advanced Statistical Theory II **Fall 2019**

Lecture 1: August 12

Lecturer: Alessandro Rinaldo *Scribe: David Zhao*

Note: *LaTeX template courtesy of UC Berkeley EECS dept.*

In this lecture, we present several motivations for studying what is known as the **supremum of the empirical process**. This object of interest will occupy us for the first half of the semester.

1.1 Uniform Law of Large Numbers

Reference notes can be found in Chapter 4 of Wainwright's textbook.

Let $X_1, \dots, X_n \overset{iid}{\sim} P$ with common mean μ. Recall that by the Law of Large Numbers,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{P} \mu$$

as $n \to \infty$. We can say more under additional assumptions. E.g. if the X_i's are $SG(\sigma^2)$, then

$$\mathbb{P}\Big(|\bar{X}_n - \mu| \geq \epsilon\Big) \leq 2\exp\Big(-\frac{n\epsilon^2}{\sigma^2}\Big), \quad \forall \epsilon > 0$$

But sometimes this is not enough. For example, let $X_1, \dots, X_n \overset{iid}{\sim} P$ with CDF F, fix $t \in \mathbb{R}$, and let $\hat{F}_n$ be the empirical CDF defined as

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \leq x)$$

Note that

$$\mathbb{E}[\hat{F}_n(t)] = F(t) = \mathbb{P}(X_i \leq t)$$

Also, we can write

$$\hat{F}_n(t) \overset{d}{=} \frac{Bin(F(t), n)}{n}$$

since the empirical CDF is just $\frac{1}{n}$ times the sum of i.i.d. Bernoulli(F(t)) random variables.

It easily follows, using Hoeffding's inequality, that

$$\hat{F}_n(t) \xrightarrow{P} F(t)$$

The difficulty arises when we wish to study

$$\sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)|$$

Note that the Chernoff bound and similar techniques hold for a fixed t, not over all $t \in \mathbb{R}$. The union bound doesn't help either, since $\mathbb{R}$ is uncountable. We seek a LLN that holds uniformly over all $t \in \mathbb{R}$.

1.2 General Setup

Let us now set up the general problem, and show that deriving a uniform LLN is just a special case of studying the supremum of the empirical process.

In general, let $\mathcal{X}$ be a set and P be a probability on it. (We can think of $\mathcal{X}$ as $\mathbb{R}^d$ for most of our purposes.) Let $X_1, \ldots, X_n \overset{i.i.d.}{\sim} P$. Let P_n be the empirical probability measure associated with this sample, which defines a mapping from any measurable set to a number in $[0, 1]$:

$$A \subseteq \mathcal{X} \mapsto P_n(A) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}(X_i \in A)$$

Let $\mathcal{F}$ be a class of functions on $\mathcal{X}$ taking values in $\mathbb{R}$. Assume that

$$\sup_{f\in\mathcal{F}} \sup_{x\in\mathcal{X}} |f(x)| \leq b$$

for some $b > 0$. In other words, we assume the class of functions is uniformly bounded, which is a strong but useful assumption. We also introduce some notation. If $f \in \mathcal{F}$, we define:

$$Pf \equiv \mathbb{E}[f(X)]$$

$$P_n f \equiv \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

where $X \sim P$.

Now we arrive at our main object of interest, the **supremum of the empirical process**:

$$\|P_n - P\| = \sup_{f\in\mathcal{F}} \frac{1}{n}\Big|\sum_{i=1}^{n}\Big(f(X_i) - \mathbb{E}[f(X_i)]\Big)\Big|$$

Returning to our uniform LLN example, note that if $\mathcal{X} = \mathbb{R}$ and $\mathcal{F} = \{(-\infty, x], x \in \mathbb{R}\}$, then for $f_t = (-\infty, t] \in \mathcal{F}$, we have

$$Pf_t = \mathbb{E}[f_t(X)] = \mathbb{P}(X \leq t) = F(t)$$

and similarly,

$$P_n f_t = \hat{F}_n(t)$$

It follows that the object we need to bound in order to derive a uniform LLN is just a special case of the supremum of the empirical process:

$$\|P_n - P\| = \sup_{f\in\mathcal{F}} |F(t) - \hat{F}_n(t)|$$

As another example, we briefly show that in covariance matrix estimation, the operator norm of the difference between the empirical and true covariance matrices.

Let $X_1, \ldots, X_n \overset{iid}{\sim} P$ on $\mathbb{R}^d$ with mean 0 and covariance matrix $\Sigma = \mathbb{E}[XX^T]$. Let $\hat{\Sigma}_n = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^T$ be the empirical covariance matrix. Then we are interested in

$$\|\hat{\Sigma}_n - \Sigma\|_{op} = \max_{\nu\in\mathbb{S}^{d-1}} |\nu^T(\hat{\Sigma}_n - \Sigma)\nu|$$

where we define the unit sphere in $\mathbb{R}^d$

$$\mathbb{S}^{d-1} = \{\nu \in R^d : \ \|\nu\| = 1\}$$

For each $\nu \in \mathbb{S}^{d-1}$, we define $f_\nu : \mathbb{R}^d \to \mathbb{R}$ as

$$f_\nu(X) = \nu^T X X^T \nu$$

Then letting $\mathcal{F} = \{f_\nu, \nu \in \mathbb{S}^{d-1}\}$, we see that

$$\|\hat{\Sigma}_n - \Sigma\|_{op} = \|P_n - P\|_{\mathcal{F}}$$

So the operator norm is another familiar quantity we can express in terms of our main object of interest.

As a side note, what exactly do we mean by supremum of the empirical process? The empirical process is just a stochastic process over $\mathcal{F}$. For every function in this function class we have

$$f \in \mathcal{F} \mapsto P_n(f) - P(f)$$

In future lectures, our goal will be to show that $\|P_n - P\|_{\mathcal{F}} \xrightarrow{P-a.s.} 0$ as $n \to \infty$.

1.3 Excess Risk

Reference notes can be found in Chapter 4.2.1 of Wainwright's textbook.

Another motivation for studying the supremum of the empirical process is the decision-theoretic concern with **excess risk**.

Let $\{P_\theta : \theta \in \Omega\}$ be a collection of probability distributions on $\mathcal{X}$ indexed by some parameter $\theta \in \Omega$. Let $X_1, \ldots, X_n \overset{iid}{\sim} P_{\theta^*}$ where P_{θ^*} is in the collection. We define a **loss function** to measure the discrepancy between x and θ:

$$(x, \theta) \in \mathcal{X} \otimes \Omega \longrightarrow \mathcal{L}_\theta(x) \in \mathbb{R}_+$$

For example, we could have

$$\mathcal{L}_\theta(x) = \|x - \theta\|$$

or

$$\mathcal{L}_\theta(x) = |x - \theta|^2, \ \mathcal{X} = \Omega = \mathbb{R}$$

We can then define the **risk**:

$$R(\theta, \theta^*) = \mathbb{E}_{X \sim P_{\theta^*}}[\mathcal{L}_\theta(X)], \ \theta \in \Omega$$

and the **empirical risk**:

$$\hat{R}(\theta, \theta^*) = \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_\theta(X_i), \ \theta \in \Omega$$

This leads to the notion of the empirical risk minimzer:

$$\hat{\theta} = \arg\min_{\theta \in \Omega} \hat{R}(\theta, \theta^*)$$

For example, assume each probability distribution P_θ has a density f_θ, and define the loss function to be the log-likelihood ratio:

$$\mathcal{L}_\theta(x) = \log \frac{f_{\theta^*}(x)}{f_\theta(x)}$$

Then $\hat{\theta}$ is the MLE (maximum likelihood estimator) of θ^*, so that the minimizer of risk is the maximizer of likelihood. In this case, we also have that $R(\theta, \theta^*) = KL(P_\theta, P_{\theta^*})$.

As a concrete example, consider binary classification. We have n i.i.d. pairs $(X_i, Y_i) \in \mathbb{R}^d \times \{-1, 1\}$. We can write the joint distribution of the data as

$$P_{X,Y} = P_{Y|X} P_X$$

using Bayes' rule. We typically are not concerned with P_X. The conditional distribution $P_{Y|X}$ can be specified, via a 1-to-1 mapping, by the likelihood ratio:

$$x \in \mathbb{R}^d \mapsto \psi(x) = \frac{\mathbb{P}(Y = 1|X = x)}{\mathbb{P}(Y = -1|X = x)}$$

In this example, $\mathcal{X} = \mathbb{R}^d \times \{-1, 1\}$ is the abstract space, and Ω is the set of all classification functions.

Our goal is to estimate a function $f : \mathbb{R}^d \to \{-1, 1\}$ that minimizes $P_{X,Y}(f(X) \neq Y)$. We define the loss function

$$\mathcal{L}_f((x, y)) = \begin{cases} 1, & f(x) \neq y \\ 0, & else \end{cases}$$

Suppose that unconditionally, $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1)$. Then the canonical example of the Bayes classifier, $f^*(x)$, is the optimal classifier for this problem:

$$f^*(x) = \begin{cases} 1, & \psi(x) \geq 1/2 \\ -1, & else \end{cases}$$

Now, we come to the notion of **excess risk**:

$$\delta R(\hat{\theta}, \theta^*) = R(\hat{\theta}, \theta^*) - \inf_{\theta \in \Omega} R(\theta, \theta^*)$$

We can rewrite this as

$$\delta R(\hat{\theta}, \theta^*) = R(\hat{\theta}, \theta^*) - \hat{R}(\hat{\theta}, \theta^*) + \hat{R}(\hat{\theta}, \theta^*) - \hat{R}(\theta_0, \theta^*) + \hat{R}(\theta_0, \theta^*) - R(\theta_0, \theta^*) = T_1 + T_2 + T_3$$

where θ_0 is such that

$$R(\theta_0, \theta^*) = \inf_{\theta \in \Omega} R(\theta, \theta^*)$$

and

$$T_1 = R(\hat{\theta}, \theta^*) - \hat{R}(\hat{\theta}, \theta^*)$$
$$T_2 = \hat{R}(\hat{\theta}, \theta^*) - \hat{R}(\theta_0, \theta^*)$$
$$T_3 = \hat{R}(\theta_0, \theta^*) - R(\theta_0, \theta^*)$$

Note that $T_2 \leq 0$ since $\hat{\theta}$ is the ERM that minimizes $\hat{R}$. So we have

$$\delta R(\hat{\theta}, \theta^*) = T_1 + T_2 + T_3 \leq T_1 + T_3$$

The term T_3 is also easily dealt with, as we can just use a standard concentration inequality because both θ_0 and θ^* are fixed.

The term T_1 is the difficult one, since $\hat{\theta}$ is random and data-dependent. We basically need to bound

$$\begin{aligned} T_1 &\leq \sup_{\theta \in \Omega} \frac{1}{n} \Big| \sum_{i=1}^{n} \Big(\mathcal{L}_\theta(x_i) - \mathbb{E}[\mathcal{L}_\theta(x_i)] \Big) \Big| \\ &= \|P_n - P\|_{\mathcal{F}} \end{aligned}$$

where we define the function class

$$\mathcal{F} = \{\mathcal{L}_\theta(\cdot),\ \theta \in \Omega\}$$

Observe that yet again, we need to "sup out", and yet again our difficult problem reduces to a special case of bounding the supremum of an empirical process.

References

[W01] M. Wainwright, "High-Dimensional Statistics: A Non-Asymptotic Viewpoint," 2019.