MATH 281C: Mathematical Statistics

Lecture 1

To begin with, we provide a high-level overview of what we plan to cover in this class.

1 Some Aspects of Empirical Process Theory

1.1 Uniform Laws of Large Numbers

The first question concerns uniform (strong) laws of large numbers (ULLN). Let $X_1, X_2, \ldots,$ be independent and identically distributed (iid) random objects taking values in a set $\mathcal{X}$. Denote by P the common distribution on $\mathcal{X}$. For example, $\mathcal{X}$ can be taken as $\mathbb{R}^p$ ($p \geq 1$) or a subset of $\mathbb{R}^p$. Let $\mathcal{F}$ denote class of real-valued functions on $\mathcal{X}$, that is, each $f \in \mathcal{F}$ is a function $\mathcal{X} \to \mathbb{R}$. What can we say about the random quantity

$$\Delta_n = \Delta_n(X_1, \ldots, X_n; \mathcal{F}) := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X_1) \right|. \tag{1.1}$$

Specifically,

1. Does the random variable Δ_n concentrate around its expectation? That is, how large is $|\Delta_n - \mathbb{E}\Delta_n|$? At the very least this difference should vary with n, and depend on the "complexity" of the function class $\mathcal{F}$.

2. Can we provide *finite-sample* bounds (i.e., bounds that hold for every n or all sufficiently large n) for Δ_n in (1.1) in terms of the class of functions $\mathcal{F}$ and the common distribution P of $X_1, X_2, \ldots$?

3. Can we provide conditions on $\mathcal{F}$ such that Δ_n converges to zero *in probability* or *almost surely*? Is this is true, we say that the uniform strong law of large numbers holds.

Empirical process theory provides answers to these questions. Why are these questions relevant to mathematical statistics or even learning theory? The two examples that we shall study in detail are given below.

Before proceeding, we introduce a representative figure in this area – Michel Talagrand, a French mathematician known for his work in functional analysis and probability theory. Among many other awards, Michel Talagrand was awarded the The Shaw Price in Mathematical Sciences 2019 [Press Release] for his work on concentration inequalities, on suprema of stochastic processes and on rigorous results for spin glasses. The first Shaw Prize laureate in Mathematical Science is Shiing Shen Chern (2004), and the only statistician that was awarded is David Donoho (2013).

Example 1.1 (Classification)**.** Consider a pair of random objects X and Y having some joint distribution, where X takes values in a space $\mathcal{X}$ and $Y \in \{-1, 1\}$ is binary. A *classifier* is a function $g : \mathcal{X} \to \{-1, 1\}$. The error of a classifier (classification error) is given by

$$L(g) := \mathbb{P}\{g(X) \neq Y\}.$$

The goal of classification is to construct a classifier with small error based on iid observations $(X_1, Y_1), \ldots, (X_n, Y_n)$ that have the same distribution as (X, Y).

Given a classifier g, its empirical error (i.e., its error on the observed sample) is given by

$$L_n(g) := \frac{1}{n} \sum_{i=1}^{n} I\{g(X_i) \neq Y_i\}.$$

A natural strategy for classification is to select a class of classifiers C, and then to choose the classifier in C which has the smallest empirical error on the observed sample, i.e.,

$$\widehat{g}_n := \operatorname*{argmin}_{g \in C} L_n(g).$$

How good a classifier $\widehat{g}_n$ is, i.e., how small is its error:

$$L(\widehat{g}_n) := \mathbb{P}\{\widehat{g}_n(X) \neq Y | X_1, Y_1, \ldots, X_n, Y_n\}.$$

Two questions are relevant about $L(\widehat{g}_n)$:

1. Is $L(\widehat{g}_n)$ comparable to $\inf_{g \in C} L(g)$, i.e., is the error of $\widehat{g}_n$ comparable to the best achievable error in the class C? In other words,

$$\inf_{g \in C} L(g) \leq L(\widehat{g}_n) \leq \inf_{g \in C} L(g) + \underbrace{\text{excess risk}}_{\text{how large?}}$$

2. Is $L(\widehat{g}_n)$ comparable to $L_n(\widehat{g}_n)$, i.e., is the error of $\widehat{g}_n$ comparable to its "in-sample" empirical error?

It is relatively easy to relate these two questions to the size of

$$\sup_{g \in C} |L_n(g) - L(g)|.$$

Let $g^* := \operatorname{argmin}_{g \in C} L(g)$ be the best classifier. Then

$$\begin{aligned} L(\widehat{g}_n) &= L(\widehat{g}_n) - L_n(\widehat{g}_n) + L_n(\widehat{g}_n) - L(g^*) + L(g^*) \\ &\leq L(\widehat{g}_n) - L_n(\widehat{g}_n) + L_n(g^*) - L(g^*) + L(g^*) \\ &\leq L(g^*) + 2 \sup_{g \in C} |L_n(g) - L(g)|. \end{aligned}$$

Also,

$$L(\widehat{g}_n) = L(\widehat{g}_n) - L_n(\widehat{g}_n) + L_n(\widehat{g}_n) \leq L_n(\widehat{g}_n) + \sup_{g \in C} |L_n(g) - L(g)|.$$

Thus the key quantity to answering the above questions is $\sup_{g \in C} |L_n(g) - L(g)|$, which is a special case of (1.1) when $\mathcal{F}$ is taken to be

$$\mathcal{F} = \{(x, y) \to I\{g(x) \neq y\} : g \in C\}.$$

To be precise, the X_i's in (1.1) should be replaced by (X_i, Y_i)'s.

Sometimes, the two inequalities above can sometimes be quite loose. Later, we shall see more sharper inequalities which utilize a technique known as "localization".

Example 1.2 (Consistency and Rates of Convergence of M-estimators)**.** Many problems in statistics are concerned with estimators of the form

$$\widehat{\theta_n} := \operatorname*{argmax}_{\theta \in \Theta} \underbrace{\frac{1}{n} \sum_{i=1}^{n} m_\theta(X_i)}_{=: M_n(\theta)} \tag{1.2}$$

for iid observations $X_1, \ldots, X_n$ taking values in $\mathcal{X}$. Here Θ denotes the parameter space and, for each $\theta \in \Theta$, m_θ denotes a real-valued function (known as a loss or criterion function) on $\mathcal{X}$. Such an estimator $\widehat{\theta_n}$ is called an M-estimator as it is obtained by maximizing an objective function. The most standard examples of M-estimators are:

- **Maximum likelihood estimators**: These correspond to $m_\theta(x) = \log p_\theta(x)$ for a class of densities $\{p_\theta, \theta \in \Theta\}$ on $\mathcal{X}$.
- **Location estimators**:
  (a) Mean: $m_\theta(x) = -(x - \theta)^2$;
  (b) Median: $m_\theta(x) = -|x - \theta|$;
  (c) Model: $m_\theta(x) = I(|x - \theta| \leq 1)$.

The target quantity for the estimator $\widehat{\theta_n}$ is

$$\theta_0 := \operatorname*{argmax}_{\theta \in \Theta} \underbrace{\mathbb{E} m_\theta(X_1)}_{=: M(\theta)}.$$

The main question of interest while studying M-estimators concerns the accuracy of $\widehat{\theta_n}$ for estimating θ_0. In the asymptotic framework $n \to \infty$, the two key questions are

1. Is $\widehat{\theta_n}$ consistent for estimating θ_0, i.e., does $d(\widehat{\theta_n}, \theta_0)$ converge to zero *almost surely* or *in probability* as $n \to \infty$? Here $d(\cdot, \cdot)$ is a metric on Θ. Most of the time we will use the Euclidean metric.
2. What is the rate of convergence of $d(\widehat{\theta_n}, \theta_0)$? For example, is it $O_{\mathbb{P}}(n^{-1/2})$ or $O_{\mathbb{P}}(n^{-1/3})$?

To answer these questions, one must investigate the closeness of $(1/n) \sum_{i=1}^{n} m_\theta(X_i)$ to $\mathbb{E} m_\theta(X_1)$ in some sort of uniform sense over θ, which leads to investigation of (1.1) for appropriate subclasses $\mathcal{F}$ of $\{m_\theta, \theta \in \Theta\}$.

Recall from the early studies in Math 281A, given $\epsilon > 0$, we have

$$\mathbb{P}\{d(\widehat{\theta}_n, \theta_0) \geq \epsilon\} \leq \mathbb{P}\left\{ \sup_{\theta \in \Theta : d(\theta, \theta_0) \geq \epsilon} M_n(\theta) \geq M_n(\theta_0) \right\}, \tag{1.3}$$

We then bound the right-hand of (1.3) by

$$\mathbb{P}\left(\sup_{\theta \in \Theta : d(\theta, \theta_0) \geq \epsilon} [M_n(\theta) - M(\theta) - \{M_n(\theta_0) - M(\theta_0)\}] \geq \underbrace{- \sup_{\theta \in \Theta : d(\theta, \theta_0) \geq \epsilon} \{M(\theta) - M(\theta_0)\}}_{\text{deterministic, strictly positive}} \right).$$

Empirical process results provide bounds for the above probability (some assumptions on the relation between M and the metric d will be needed).

The high-level strategy for controlling Δ_n in (1.1) is as follows (we will mostly focus on the case when $\mathcal{F}$ is a uniformly bounded class of functions):

(i) The key observation is that the random variable Δ_n in (1.1) "concentrates" around its mean (or expectation).

(ii) Because of *concentration*, it suffices to control the mean of Δ_n. The mean will be bounded by a quantity called *Rademacher complexity* of $\mathcal{F}$ via a technique called *symmetrization*.

(iii) The Rademacher complexity involves the expected supremum over a *sub-Gaussian process*. This is further controlled via a technique known as *chaining*. In the process, we shall also encounter a quantity known as the *Vapnik-Chervonenkis dimension*.

A comprehensive reference for these topics is the book Boucheron, Lugosi and Massart (2013). In this course we will take the nonasymptotic viewpoint, where bounds to be established hold for every n. The more classical viewpoint is the asymptotic one (as in 281A) where statements are made that hold as $n \to \infty$. In the asymptotic regime, it is said that the class $\mathcal{F}$ is *Glivenko-Cantelli* provided Δ_n in (1.1) converges almost surely as $n \to \infty$. Using our nonasymptotic bounds, it will be possible to put appropriate conditions on $\mathcal{F}$ under which $\mathcal{F}$ becomes a Glivenko-Cantelli class.

1.2 Uniform Central Limit Theorems

Let us now describe the second fundamental question that is addressed by empirical process theory.

By the classical Central Limit Theorem (CLT), we have as $n \to \infty$ that

$$n^{1/2} \left\{ \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X_i) \right\}$$

converges in distribution to a normal distribution with mean zero and variance $\mathrm{var}(f(X_1))$. This statement is true for every $f \in \mathcal{F}$. Does this convergence hold uniformly over f in the class $\mathcal{F}$ in a reasonable sense? To illustrate this, let us look at the following example.

Example 1.3. Suppose that $X_1, \ldots, X_n$ are iid observations from the uniform distribution on $[0, 1]$. Also suppose that $\mathcal{F}$ consists of all indicator functions $\{x \to I_{(-\infty, t]}(x) : t \in \mathbb{R}\}$. In this case, for $f = I_{(-\infty, t]}$, the quantity

$$\frac{1}{n} \sum_{i=1}^{n} f(X_i) = F_n(t),$$

where $F_n(\cdot)$ is the empirical cumulative distribution function (ECDF) of the observations $X_1, \ldots, X_n$. Define

$$U_n(t) := n^{1/2}\{F_n(t) - t\}, \quad t \in [0,1]. \tag{1.4}$$

Here $U_n(t)$ represents a collection of random variables as t varies in $[0,1]$. The *stochastic process* $\{U_n(t), t \in [0,1]\}$ is known as the "Uniform Empirical Process". It is easy to see that every realization of $\{U_n(t), t \in [0,1]\}$, viewed as a function on $[0,1]$, is piecewise linear with jump discontinuities at the n data points $X_1, \ldots, X_n$. Also $U_n(0) = U_n(1) = 0$ for every n.

The CLT states that, for each $t \in [0,1]$ fixed, the sequence of real-valued random variables $\{U_n(t)\}_{n \geq 1}$ converges in distribution to $\mathcal{N}(0, t - t^2)$ (why?) as $n \to \infty$. Moreover, the multivariate CLT states that, for every finite collection of points $t_1, \ldots, t_k$, the sequence of random vectors $(U_n(t_1), \ldots, U_n(t_k))^\intercal$ converges in distribution to a multivariate normal distribution with zero mean and covariance matrix given by

$$\Sigma_k := (\min(t_i, t_j) - t_i t_j)_{1 \leq i,j \leq k}.$$

At this point, let us introduce an object called *Brownian Bridge*. The Brownian Bridge on $[0,1]$ is a stochastic process $\{U(t), 0 \leq t \leq 1\}$ that is characterized by the following two requirements:

(i) Every realization is a continuous function on $[0,1]$ with $U(0)$ and $U(1)$ always fixed to be 0.

(ii) For every fixed $t_1, \ldots, t_k$ in $[0,1]$, the random vector $(U(t_1), \ldots, U(t_k))^\intercal$ has a multivariate normal distribution with zero mean and covariance matrix given by Σ_k as defined above.

We therefore see that the "finite dimensional distributions" of the process $\{U_n(t), 0 \leq t \leq 1\}$ converge to the "finite dimensional distributions" of $\{U(t), 0 \leq t \leq 1\}$. It is natural to ask here if we can claim anything beyond finite-dimensional convergence here. Does the entire process $\{U_n(t), 0 \leq t \leq 1\}$ converges to $\{U(t), 0 \leq t \leq 1\}$, and in what sense? This was first conjectured by Doob and rigorously proved by Donsker. We shall discuss Donsker's result in the next class.

References

Boucheron, S., Lugosi, G. and Massart, P. (2013). *Concentration Inequalities: A Nonasymptotic Theory of Independence.* Oxford Univ. Press, Oxford.

Dudley, R. M. (2002). *Real Analysis and Probability.* Cambridge Univ. Press, Cambridge.