MATH 281C: Mathematical Statistics

Lecture 2

Let us first start by giving an introduction to Uniform Central Limit Theorems (a topic we will study in detail later). Next, we will talk about measure concentration properties of $(1/n)\sum_{i=1}^n f(X_i) - \mathbb{E}f(X_1)$.

1 Uniform Central Limit Theorems

Recall the uniform empirical process

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

where $F_n(t) = (1/n)\sum_{i=1}^n I(X_i \le t)$ with $X_1, \ldots, X_n \sim \text{Unif}[0,1]$. Also, let $\{U(t), 0 \le t \le 1\}$ be the Brownian bridge on $[0,1]$.

Construction of Brownian bridge. We say the stochastic process $\{W(t), t \ge 0\}$ is a standard *Wiener process* if it satisfies

(i) For $t \ge 0$, $W(t) \sim \mathcal{N}(0,t)$, and $W(0) = 0$;

(ii) For $0 \le t_1 < t_2 < \cdots < t_k < \infty$, $W(t_2) - W(t_1), W(t_3) - W(t_2), \ldots, W(t_n) - W(t_{n-1})$ are independent, and for any $s < t$, $W(t) - W(s)$ is equal in distribution to $W(t-s)$;

(iii) $W(t)$ is continuous in t.

For every $T > 0$, $U(t;T) := W(t) - \frac{t}{T}W(T)$ is a Brownian bridge on $[0,T]$.

What is the meaning of the statement that the sequence of stochastic processes $\{U_n(t), t \in [0,1]\}$ converges in distribution to $\{U(t), t \in [0,1]\}$? To understand this, let us first recall the usual notion of convergence in distribution for sequences of random vectors. We say that a sequence of random vectors $\{Z_n\}$ taking values in $\mathbb{R}^k$ converges in distribution to Z if and only if

$$\mathbb{E}h(Z_n) \to \mathbb{E}h(Z) \quad \text{as } n \to \infty$$

for every **bounded continuous real-valued function** $h : \mathbb{R}^k \to \mathbb{R}$.

We can attempt a direct generalization of this to define convergence of $U_n(\cdot)$ to $U(\cdot)$ as a stochastic process. These processes take values not in $\mathbb{R}^k$ but in the space of all bounded functions on $[0,1]$. Denote this space by $\ell^\infty([0,1])$. This space can be equipped with the supremum metric: $\|g_1 - g_2\|_\infty := \sup_{0 \le t \le 1} |g_1(t) - g_2(t)|$, for $g_1, g_2 \in \ell^\infty([0,1])$. We can then say that U_n converges in distribution to U as a stochastic process provided

$$\mathbb{E}h(U_n) \to \mathbb{E}h(U) \quad \text{as } n \to \infty \tag{1.1}$$

for every bounded and continuous real-valued function $h : \ell^\infty([0,1]) \to \mathbb{R}$. This definition almost makes sense except for one measure-theoretic issue. It turns out that there exist bounded and continuous real-valued functions $h : \ell^\infty([0,1]) \to \mathbb{R}$ for which the random variable $h(U_n)$ is not measurable. To overcome this technical issue, we can replace the left-hand side in (1.1) by its *outer expectation* $\mathbb{E}^* h(U_n)$ (formally defined later).

In this sense, Donsker showed that U_n converges in distribution to Brownian bridge.

Let us now return to the general case. Here we consider the stochastic process

$$G_n(f) := n^{1/2}\left\{\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X_1)\right\} \quad \text{for } f \in \mathcal{F}.$$

Under a simple assumption such as $\sup_{f\in\mathcal{F}} |f(x)| < \infty$ for every $x \in \mathcal{X}$, the function $f \mapsto G_n(f)$ belongs to the space $\ell^\infty(\mathcal{F})$. We then say that the uniform central limit theorem holds over $\mathcal{F}$ if the stochastic process $\{G_n(f), f \in \mathcal{F}\}$ converges in distribution in $\ell^\infty(\mathcal{F})$ to a process $\{G(f), f \in \mathcal{F}\}$ as $n \to \infty$. The limit process $\{G(f), f \in \mathcal{F}\}$ will have the property that, for every $f_1, \ldots, f_k \in \mathcal{F}$, the random vector $(G(f_1), \ldots, G(f_k))^\intercal$ will have a multivariate normal distribution having the same covariance as $(G_n(f_1), \ldots, G_n(f_k))^\intercal$.

We shall characterize convergence in distribution in $\ell^\infty(\mathcal{F})$, and then see some sufficient conditions on $\mathcal{F}$ that ensure that the Uniform CLT holds.

The following are some statistical applications of Uniform CLTs.

Example 1.1 (Goodness-of-fit testing)**.** Suppose we observe iid observations $X_1, \ldots, X_n$ from a CDF F, and we want to test the null hypothesis $H_0 : F = F_0$ versus the alternative hypothesis $H_1 : F \neq F_0$. Here F_0 is a fixed (known) distribution function.

Kolmogorov recommended testing this hypothesis via the statistic

$$D_n := n^{1/2} \sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)|,$$

where F_n is the empirical CDF of the data $X_1, \ldots, X_n$. The idea is to reject H_0 when D_n is large. To calculate the p-value of this test, the null distribution *(i.e., the distribution of D_n under H_0) needs to be determined. An interesting property of the null distribution of D_n is that the null distribution does not depend on F_0 as long as F_0 is continuous.

Assume F_0 is continuous. Then under the null, $F_0(X_1), \ldots, F_0(X_n)$ are iid following the uniform distribution Unif $[0,1]$, so that

$$\sup_{x\in\mathbb{R}} |F_n(x) - F_0(x)| = \sup_{t\in[0,1]} \left|\frac{1}{n}\sum_{i=1}^{n} I\{F_0(X_i) \le t\} - t\right| \qquad \text{(by change of variable } t = F_0(x)).$$

Therefore, we can compute the null distribution of D_n assuming that F_0 is the uniform distribution on $[0,1]$. In this case, we can write

$$D_n = \sup_{0\le t\le 1} |U_n(t)|,$$

where $U_n(\cdot)$ is the uniform empirical process.

The fact that $\{U_n(t), t \in [0,1]\}$ converges in distribution to a Brownian bridge $\{U(t), t \in [0,1]\}$ as $n \to \infty$ allows one to claim that

$$\lim_{n\to\infty} \mathbb{P}(D_n \le x) = \mathbb{P}\left\{\sup_{0\le t\le 1} |U(t)| \le x\right\}, \quad \text{for every } x > 0.$$

The latter probability can be computed exactly; see, for example, Proposition 12.3.4 in Dudley (2002). Thus the uniform central limit theorem gives a way of computing asymptotically valid p-values for goodness-of-fit via the Kolmogorov statistic.

The same argument can be used for many related goodness-of-fit statistics such as

(1) Cramér-von Mises statistic:

$$W_n := n \int \{F_n(x) - F_0(x)\}^2 \mathrm{d}F_0(x).$$

(2) Anderson-Darling statistic:

$$A_n := n \int \frac{\{F_n(x) - F_0(x)\}^2}{F_0(x)\{1 - F_0(x)\}} \mathrm{d}F_0(x).$$

(3) Smirnov statistic:

$$D_n^+ := n^{1/2} \sup_x \{F_n(x) - F_0(x)\} \quad \text{and} \quad D_n^- = n^{1/2} \sup_x \{F_0(x) - F_n(x)\}.$$

The asymptotic null distribution of all these statistics can be computed from Brownian bridge, and this will be validated by the uniform CLT.

Example 1.2 (Asymptotic Distribution of MLE)**.** Suppose $X_1, \ldots, X_n$ are iid from an unknown density p_{θ_0} belonging to a known class $\{p_\theta : \theta \in \Theta \subseteq \mathbb{R}^k\}$. Let $\widehat{\theta}_n$ denote the MLE of θ_0, defined as the maximizer of

$$\frac{1}{n} \sum_{i=1}^n \log p_\theta(X_i) \quad \text{over } \theta \in \Theta.$$

A classical result is that, under some smoothness assumptions, $n^{1/2}(\widehat{\theta}_n - \theta_0)$ converges in distribution to $\mathcal{N}(0, I(\theta_0)^{-1})$, where $I(\theta_0)$ denotes the $k \times k$ Fisher information matrix defined as

$$I(\theta_0) := \mathbb{E}\{\nabla_\theta \log p_\theta(X) \{\nabla_\theta \log p_\theta(X)\}^\intercal\}\Big|_{\theta=\theta_0},$$

and the expectation is taken with respect to the density p_θ.

What smoothness assumptions need to be imposed on p_θ, $\theta \in \Theta$ for this result to hold? Because the result involves the information matrix $I(\theta_0)$ that depends on gradients, a minimal assumption seems to be that $\theta \mapsto \log p_\theta(x)$ is differentiable with respect to θ. Also, because of the presence of the expectation in the definition of $I(\theta_0)$, it should be okay if the derivative w.r.t. θ does not exist on sets of measure zero under p_{θ_0} (think about the density $p_\theta(x) = \exp(-|x - \theta|/2)$).

The classical proofs of this result assume, however, that this map allows derivatives of order two, and sometimes even three. Using uniform central limit theorems, we will present later a proof using a minimal differentiability assumption, called *Differentiability in Quadratic Mean* (DQM).

Example 1.3 (Asymptotic Distribution of M-estimators)**.** Uniform central limit theorems can be used to derive limiting distributions of other M-estimators as well. For example, consider the sample median defined as

$$\widehat{\theta}_n := \operatorname*{argmin}_{\theta \in \mathbb{R}} \frac{1}{n} \sum_{i=1}^n |X_i - \theta|.$$

Assuming that the CDF F of the observations is differentiable at its median θ_0 with positive density $f(\theta_0)$, it can be proved that $n^{1/2}(\widehat{\theta}_n - \theta_0)$ converges in distribution to $\mathcal{N}(0, (4f^2(\theta_0))^{-1})$.

For the mode estimator defined as $\operatorname{argmax}_{\theta\in\mathbb{R}} \sum_{i=1}^n m_\theta(X_i)$ with $m_\theta(x) = I(|x-\theta| \le 1)$ and $\Theta = \mathbb{R}$, the asymptotic distribution is much more complicated. The objective function is not even continuous. The result is that

$$n^{1/3}(\widehat{\theta}_n - \theta_0)$$

converges in distribution to

$$\operatorname*{argmax}_{h\in\mathbb{R}} \{aZ(h) - bh^2\},$$

where $\{Z(h), h \in \mathbb{R}\}$ is a standard two-sided Brownian motion (Wiener process) starting from 0,

$$a = p(\theta_0 + 1) - p(\theta_0 - 1) \quad \text{and} \quad b = \frac{1}{2}\{p'(\theta_0 - 1) - p'(\theta_0 + 1)\}.$$

Here $p(\cdot)$ represents the the density of the observations and it is assumed that p is unimodal and symmetric w.r.t. mode θ_0, i.e., $p'(x) > 0$ for $x < \theta_0$ and $p'(x) < 0$ for $x > \theta_0$. This result is stated here just to illustrate that the limiting distributions of even simple-looking M-estimators can be quite complicated. Later we will see how to prove these results via Uniform CLTs.

2 Concentration Results

Let us now start with our discussion of uniform laws of large numbers. The key object of study is

$$\Delta_n = \Delta_n(X_1, \ldots, X_n) := \sup_{f\in\mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}f(X_1) \right|, \tag{2.1}$$

where $X_1, \ldots, X_n$ are iid random objects taking values in a space $\mathcal{X}$, and $\mathcal{F}$ is a collection of real-valued functions on $\mathcal{X}$. We will argue that Δ_n concentrates around its expectation. The is fairly easy to prove when it is assumed that all the functions in $\mathcal{F}$ are bounded by a constant $B > 0$:

$$\sup_{x\in\mathcal{X}} |f(x)| \le B \quad \text{for every } f \in \mathcal{F}. \tag{2.2}$$

Under the above assumption, we will prove a concentration result for Δ_n. This in fact is a direct consequence of the *bounded differences inequality*.

Theorem 2.1 (Bounded Differences Inequality)**.** Suppose $X_1, \ldots, X_n$ are independent random variables taking values in a set $\mathcal{X}$. Suppose $g : \mathcal{X} \times \cdots \times \mathcal{X} \to \mathbb{R}$ is a function satisfying the following "bounded differences" assumption:

$$\sup_{x_1,\ldots,x_n,x_i'\in\mathcal{X}} \left| g(x_1, \ldots, x_n) - g(x_1, \ldots, x_{i-1}, x_i', x_{i+1}, \ldots, x_n) \right| \le c_i \tag{2.3}$$

for $i = 1, \ldots, n$. Then, for every $t \ge 0$,

$$\mathbb{P}\{g(X_1, \ldots, X_n) \ge \mathbb{E}g(X_1, \ldots, X_n) + t\} \le \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2} \right) \tag{2.4}$$

and

$$\mathbb{P}\{g(X_1, \ldots, X_n) \le \mathbb{E}g(X_1, \ldots, X_n) - t\} \le \exp\left(\frac{-2t^2}{\sum_{i=1}^n c_i^2} \right). \tag{2.5}$$

Remark 2.1. The bounded differences condition (2.3) is equivalent to the following:

$$\left|g(x_1,\dots,x_n) - g(z_1,\dots,z_n)\right| \le c_i$$

whenever $(x_1,\dots,x_n)$ and $(z_1,\dots,z_n)$ differ in exactly the i-th coordinate. It is also equivalent to the following condition:

$$\left|g(x_1,\dots,x_n) - g(z_1,\dots,z_n)\right| \le \sum_{i=1}^{n} c_i I(x_i \neq z_i) \quad \text{for all } x_1,\dots,x_n,z_1,\dots,z_n \in \mathcal{X}.$$

Theorem 2.1 can be seen as a quantification of the following qualitative statement of Talagrand: '*A random variable that depends on (in a "smooth" way) the influence of many independent variables (but not too much on any of them) is essentially constant (i.e. concentrates)*' (Talagrand, 1996). We will prove this result in the next class.

Let us argue here that Theorem 2.1 implies a concentration inequality for $\Delta_n(X_1,\dots,X_n)$, defined in (2.1), under condition (2.2). Let

$$g(x_1,\dots,x_n) = \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(x_i) - \mathbb{E}f(X_1)\right|.$$

By replacing x_i with x_i', we see that

$$\begin{aligned} g(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n) &= \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j\neq i} f(x_j) + \frac{1}{n}f(x_i') - \mathbb{E}f(X_1)\right| \\ &= \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n} f(x_j) - \mathbb{E}f(X_1) + \frac{1}{n}f(x_i') - \frac{1}{n}f(x_i)\right| \\ &\le \sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{j=1}^{n} f(x_j) - \mathbb{E}f(X_1)\right| + \frac{2B}{n}. \end{aligned}$$

Taking supremum over $f \in \mathcal{F}$ on both sides, we obtain

$$g(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n) \le g(x_1,\dots,x_{i-1},x_ix_{i+1},\dots,x_n) + \frac{2B}{n}.$$

Interchanging the roles of x_i and x_i' yields

$$\left|g(x_1,\dots,x_{i-1},x_i',x_{i+1},\dots,x_n) - g(x_1,\dots,x_{i-1},x_ix_{i+1},\dots,x_n)\right| \le \frac{2B}{n},$$

so that (2.3) holds with $c_i = 2B/n$. Theorem 2.1, specifically inequality (2.4), then gives

$$\mathbb{P}(\Delta_n \ge \mathbb{E}\Delta_n + t) \le \exp\left(\frac{-nt^2}{2B^2}\right), \quad \text{valid for any } t \ge 0.$$

Setting

$$\delta = \exp\exp\left(\frac{-nt^2}{2B^2}\right) \quad \text{so that} \quad t = B\sqrt{\frac{2\log(1/\delta)}{n}},$$

we obtain that with probability at least $1-\delta$,

$$\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i)-\mathbb{E}f(X_1)\right| \le \mathbb{E}\left\{\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i)-\mathbb{E}f(X_1)\right|\right\} + B\sqrt{\frac{2\log(1/\delta)}{n}}. \tag{2.6}$$

This inequality implies that $\mathbb{E}\Delta_n$ is usually the dominating term for understanding the behavior of Δ_n. This is because typically $\mathbb{E}\Delta_n$ dominates the last term on the right-hand side of (2.6). Indeed, for every $f \in \mathcal{F}$,

$$\mathbb{E}\Delta_n \ge \mathbb{E}\left|\frac{1}{n}\sum_{i=1}^{n} f(X_i)-\mathbb{E}f(X_1)\right|. \tag{2.7}$$

Because

$$\mathbb{E}\left\{\frac{1}{n}\sum_{i=1}^{n} f(X_i)-\mathbb{E}f(X_1)\right\}^2 = \frac{1}{n}\mathrm{var}(f(X_1)),$$

it is reasonable to believe that the right-hand of (2.7) will typically be of order $\sqrt{\mathrm{var}(f(X_1))/n}$. Unless $\mathrm{var}(f(X_1))$ is much smaller compared to B^2 for every $f \in \mathcal{F}$, the first term on the right-hand side of (2.6) usually dominates the second, and hence in order to control the random quantity Δ_n, it suffices to focus on the expectation $\mathbb{E}\Delta_n$.

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.

Talagrand, M. (1996). A new look at independence. *The Annals of Probability* **24** 1–34.