Lecture 18

Instructor: Quan Zhou

For more details about the materials covered in this note, see Chapters 8.2 to 8.6 and 9.6 of Resnick [3] and Chapter 3.2 of Durrett [2].

18.1 More about convergence in distribution

Theorem 18.1 (Skorohod's representation theorem). *If $\{X_n\}_{n\geq 0}$ is a sequence of random variables defined on $(\Omega, \mathcal{F}, \mathsf{P})$ and $X_n \overset{D}{\to} X_0$. Then there exist random variables $\{Y_n\}_{n\geq 0}$ defined on $([0,1], \mathcal{B}([0,1]), m)$ where m denotes the Lebesgue measure such that for every $n \geq 0$, $X_n \overset{D}{=} Y_n$ (i.e. X_n, Y_n have the same distribution) and $Y_n \overset{a.s.}{\to} Y_0$.*

Proof. See the textbook. □

Example 18.1 (Delta method). Given i.i.d. observations $X_1, X_2, \dots$ with mean μ and finite variance σ^2, by central limit theorem $\sqrt{n}(\bar{X}_n - \mu)/\sigma \overset{D}{\to} N(0,1)$ where $\bar{X}_n$ denotes the average of the first n observations and $N(0,1)$ is a standard normal random variable. Skorohod's representation theorem can be used to prove the following result which is widely used in statistics. If g is a Borel function with non-zero derivative at μ, then

$$\frac{\sqrt{n}\left[g(\bar{X}_n) - g(\mu)\right]}{\sigma g'(\mu)} \overset{D}{\to} N(0,1).$$

Theorem 18.2. *$X_n \overset{D}{\to} X$ if and only if for every bounded continuous function h, we have $E[h(X_n)] \to E[h(X)]$.*

Proof. By Skorohod's representation theorem, we may let $\{Y_n\}$ be a sequence of random variables such that $Y_n \overset{a.s.}{\to} Y$, $X_n \overset{D}{=} Y_n$ and $X \overset{D}{=} Y$. By the continuous mapping theorem for almost sure convergence, we have $h(Y_n) \overset{a.s.}{\to} h(Y)$, and the bounded convergence theorem implies that

$$E[h(X_n)] = E[h(Y_n)] \to E[h(Y)] = E[h(X)].$$

To prove the converse, fix an arbitrary $x \in \mathbb{R}$ and let

$$h_{x,\epsilon}(y) = \begin{cases} 1 & y \le x \\ 0 & y \ge x + \epsilon \\ \text{linear} & x \le y \le x + \epsilon, \end{cases}$$

which is continuous and bounded and thus $E[h_{x,\epsilon}(X_n)] \to E[h_{x,\epsilon}(X)]$. So,

$$\limsup_{n\to\infty} \mathsf{P}(X_n \le x) \le \limsup_{n\to\infty} E[h_{x,\epsilon}(X_n)] = E[h_{x,\epsilon}(X)] \le \mathsf{P}(X \le x + \epsilon).$$

Letting $\epsilon \downarrow 0$, we get $\limsup_{n\to\infty} \mathsf{P}(X_n \le x) \le \mathsf{P}(X \le x)$. Similarly,

$$\liminf_{n\to\infty} \mathsf{P}(X_n \le x) \ge \liminf_{n\to\infty} E[h_{x-\epsilon,\epsilon}(X_n)] = E[h_{x-\epsilon,\epsilon}(X)] \ge \mathsf{P}(X \le x - \epsilon).$$

Letting $\epsilon \downarrow 0$, we get $\liminf_{n\to\infty} \mathsf{P}(X_n \le x) \ge \mathsf{P}(X \le x)$ for any x at which F_X (the distribution function of X) is continuous. Combining the lim sup and lim inf inequalities, we conclude that $X_n \xrightarrow{D} X$. □

Theorem 18.3 (Continuous mapping theorem). *Let $g : \mathbb{R} \to \mathbb{R}$ be a measurable function and denote the set of discontinuity points by $D_g = \{x \in \mathbb{R} \colon g$ is discontinuous at $x\}$. If $X_n \xrightarrow{D} X$ and $\mathsf{P}(X \in D_g) = 0$, then $g(X_n) \xrightarrow{D} g(X)$. If in addition g is bounded, then $E[g(X_n)] \to E[g(X)]$.*

Proof. Let $\{Y_n\}$ be a sequence of random variables such that $Y_n \xrightarrow{a.s.} Y$, $X_n \stackrel{D}{=} Y_n$ and $X \stackrel{D}{=} Y$. Define two events $A = \{\omega \colon Y_n(\omega) \to Y(\omega)\}$ and $B = \{\omega \colon Y(\omega) \in D_g^c\}$. Then, $\mathsf{P}(A) = 1$, and $\mathsf{P}(B) = \mathsf{P}(Y \in D_g^c) = (\mathsf{P} \circ Y^{-1})(D_g^c) = (\mathsf{P} \circ X^{-1})(D_g^c) = \mathsf{P}(X \in D_g^c) = 1$. By the union bound, $\mathsf{P}(A \cap B) = 1$. But note that for any $\omega \in A \cap B$, $g(Y_n(\omega)) \to g(Y(\omega))$. Thus, $g(X_n) \stackrel{D}{=} g(Y_n) \xrightarrow{a.s.} g(Y) \stackrel{D}{=} g(X)$. The second conclusion follows from bounded convergence theorem. □

Proposition 18.1. *If $X_n \xrightarrow{D} c$ where c is a constant, then $X_n \xrightarrow{P} c$ (provided that $X_1, X_2, \dots$ are defined on the same space.)*

Proof. Since $X_n \xrightarrow{D} c$, we have $\mathsf{P}(X_n \le x) \to 0$ if $x < c$ and $\mathsf{P}(X_n \le x) \to 1$ if $x > c$ (note that we may not have the convergence at $x = c$.) This implies, for any $\epsilon > 0$,

$$\begin{aligned} \mathsf{P}(|X_n - c| > \epsilon) &= \mathsf{P}(X_n > c + \epsilon) + \mathsf{P}(X_n < c - \epsilon) \\ &\le 1 - \mathsf{P}(X_n \le c + \epsilon) + \mathsf{P}(X_n \le c - \epsilon) \to 0. \end{aligned}$$

That is, $X_n \xrightarrow{P} c$. □

Theorem 18.4 (Slutsky's theorem). *Suppose* $X_n \xrightarrow{D} X$ *and* $Y_n \xrightarrow{D} c$ *where* c *is a constant. Then* $X_n + Y_n \xrightarrow{D} X + c$, $X_n Y_n \xrightarrow{D} cX$ *and* $X_n / Y_n \xrightarrow{D} X/c$ *(if* $c \neq 0$*).*

Proof. See the textbook. □

18.2 Limits of distribution functions

Theorem 18.5 (Helly's selection theorem). *Let* $\{F_n\}_{n \geq 1}$ *be a sequence of distribution functions. There exists a subsequence* $\{F_{n(k)}\}_{k \geq 1}$ *and a right-continuous non-decreasing function* F *such that* $\lim_{k \to \infty} F_{n(k)}(y) = F(y)$ *at all continuity points* y *of* F*.*

Proof. See Resnick [3, Theorem 9.6.1]. □

Theorem 18.6 (Tightness). *Let* $\{F_n\}_{n \geq 1}$ *be a sequence of distribution functions. Then every subsequential limit is the distribution function of a probability measure if and only if* $\{F_n\}$ *is tight, i.e. for any* $\epsilon > 0$*, there exists* $M_\epsilon < \infty$ *such that* $\liminf_{n \to \infty} \mu_n([-M_\epsilon, M_\epsilon]) \geq 1 - \epsilon$*, where* μ_n *denotes the distribution corresponding to* F_n*.*

Proof. See Resnick [3, Theorem 9.6.2]. □

Example 18.2. If $X_n \xrightarrow{D} X$, then the sequence $\{F_n\}$ is tight where F_n is the distribution function of X_n. To prove this, first note that since a random variable is real-valued, for any ϵ, there exists C_ϵ such that $\mathsf{P}(|X| > C_\epsilon) < \epsilon/2$. Since there are at most countably many discontinuity points of F, we can pick $M_\epsilon \geq C_\epsilon$ such that $\pm M_\epsilon$ are both continuity points of F and thus $\mathsf{P}(|X_n| > M_\epsilon) \to \mathsf{P}(|X| > M_\epsilon) < \epsilon/2$, i.e. there exists $N < \infty$ such that $\inf_{n \geq N} \mathsf{P}(|X_n| \leq M_\epsilon) \geq 1 - \epsilon$.

Example 18.3. Let μ_n be the distribution of a normal random variable with mean n and variance 1. Then $\mu_n([-M, M]) < 1/2$ for any $n > M$. That is, $\{\mu_n\}_{n \geq 1}$ is not a tight collection of measures.

18.3 More about the convergence of probability measures

Theorem 18.7 (Portmanteau theorem). *Let* $\mu_1, \mu_2, \dots$ *and* μ *be distributions on* $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$*. The following statements are equivalent.*

(i) μ_n *converges weakly to* μ.

(ii) $X_n \xrightarrow{D} X$ *where* X_n *has distribution* μ_n *and* X *has distribution* μ.

(iii) For any bounded and continuous function f, $\int f d\mu_n \to \int f d\mu$.

(iv) For any bounded and Lipschitz function f, $\int f d\mu_n \to \int f d\mu$.

(v) For any open sets G, $\liminf_{n\to\infty} \mu_n(G) \geq \mu(G)$.

(vi) For any closed sets K, $\limsup_{n\to\infty} \mu_n(K) \leq \mu(K)$.

(vii) For any Borel sets A *with* $\mu(\partial A) = 0$, $\lim_{n\to\infty} \mu_n(A) = \mu(A)$.

Proof. See Durrett [2, Theorem 3.2.1]. □

Definition 18.1. The total variation distance between two probability measures μ, ν on $(\Omega, \mathcal{F})$ is defined as

$$d_{\mathrm{TV}}(\mu, \nu) = ||\mu - \nu||_{\mathrm{TV}} = \sup_{A \in \mathcal{F}} |\mu(A) - \nu(A)|.$$

We say μ_n converges to μ in total variation distance if $||\mu_n - \mu||_{\mathrm{TV}} \to 0$.

Proposition 18.2. *Let* $\{\mu_n\}$ *and* μ *be probability measures on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. *If* $||\mu_n - \mu||_{\mathrm{TV}} \to 0$, *then* μ_n *converges weakly to* μ.

Proof. This follows from the definition. □

Example 18.4. Let $\mu_n = \delta_{1/n}$, i.e. a unit point mass on the point n^{-1}. It is easy to see that the corresponding distribution function converges pointwise to $F(x) = \mathbb{1}_{[0,\infty)}(x)$ except at the discontinuity point $x = 0$. Hence, μ_n converges weakly to $\mu = \delta_0$. However, $||\mu_n - \mu||_{\mathrm{TV}} = 1$ for every n.

Theorem 18.8 (Scheffe's lemma)**.** *Let* $\{\mu_n\}$ *and* μ *be probability distributions absolutely continuous w.r.t. some measure* λ *on* $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. *Let* f_n *and* f *be the corresponding Radon-Nikodym densities. If* $f_n \to f$ λ*-a.e., then* $\int |f_n - f| d\lambda \to 0$ *and* $||\mu_n - \mu||_{\mathrm{TV}} \to 0$.

Proof. First, since Radon-Nikodym derivatives are non-negative by definition,

$$\int |f_n - f| d\lambda \leq \int (f_n + f) d\lambda = 2$$

That is, by letting $g_n = f_n + f$, we have $g_n \geq |f_n - f|$ and $\int g_n d\lambda \to \int \lim_{n\to\infty} g_n \, d\lambda$. So, by DCT,[1]

$$\int |f_n - f| d\lambda \to \int \lim_{n\to\infty} |f_n - f| d\lambda = 0.$$

The second conclusion follows from Theorem 14.4. We replicate that proof here (just in different notation). For any $A \in \mathcal{B}(\mathbb{R})$,

$$|\mu_n(A) - \mu(A)| = \left| \int_A (f_n - f) d\lambda \right| \leq \int_A |f_n - f| d\lambda \leq \int |f_n - f| d\lambda.$$

Hence, $\int |f_n - f| d\lambda \to 0$ implies that $|\mu_n(A) - \mu(A)| \to 0$ uniformly over all Borel sets A. That is, $||\mu_n - \mu||_{\mathrm{TV}} \to 0$. □

Proposition 18.3. *Let $\{\mu_n\}$ and μ be probability measures on $(\mathbb{Z}, \mathcal{P}(\mathbb{Z}))$ where $\mathbb{Z}$ denotes the set of all integers. Then $||\mu_n - \mu||_{\mathrm{TV}} \to 0$ if and only if μ_n converges weakly to μ.*

Proof. Use Scheffe's lemma. □

References

[1] Patrick Billingsley. *Probability and measure.* John Wiley & Sons, 2017.

[2] Rick Durrett. *Probability: Theory and Examples*, volume 49. Cambridge university press, 2019.

[3] Sidney Resnick. *A Probability Path.* Springer, 2019.

[1] Here we are applying a general version of DCT where we have a sequence of dominating functions $\{g_n\}$ which converges pointwise and $\int g_n d\lambda \to \int (\lim g_n) d\lambda$. You can check that the proof is almost the same as that for the original one (Theorem 5.3).