Lecture Notes 3
Convergence (Chapter 5)

1 Convergence of Random Variables

Let $X_1, X_2, \ldots$ be a sequence of random variables and let X be another random variable. Let F_n denote the CDF of X_n and let F denote the CDF of X.

Example: A good example to keep in mind is the following. Let $Y_1, Y_2, \ldots$ be a sequence of i.i.d. random variables. Let

$$X_n = \frac{1}{n} \sum_{i=1}^{n} Y_i$$

be the average of the first n of the Y_i's. This defines a new sequence $X_1, X_2, \ldots, X_n$. That is, the sequence of interest $X_1, \ldots, X_n$ might be a sequence of statistics based on some other sequence of random variables.

1. X_n **converges to** X **in probability**, written $X_n \xrightarrow{\text{P}} X$, if, for every $\epsilon > 0$,

$$\mathbb{P}(|X_n - X| > \epsilon) \to 0 \quad \text{as} \quad n \to \infty. \tag{1}$$

   In other words,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \epsilon) = 0$$

   and $X_n - X = o_P(1)$.

2. X_n **converges almost surely to** X, written $X_n \overset{a.s.}{\to} X$, if, for every $\epsilon > 0$,

$$\mathbb{P}(\lim_{n\to\infty} |X_n - X| < \epsilon) = 1. \tag{2}$$

   This is also called *Strong convergence.*

3. X_n **converges to** X **in quadratic mean** (also called convergence in L_2), written $X_n \xrightarrow{\text{qm}} X$, if

$$\mathbb{E}(X_n - X)^2 \to 0 \quad \text{as} \quad n \to \infty. \tag{3}$$

4. X_n **converges to** X **in distribution**, written $X_n \rightsquigarrow X$, if

$$\lim_{n\to\infty} F_n(t) = F(t) \tag{4}$$

   at all t at which F is continuous.

Recall the following definition.

Definition 1 *X has a point mass distribution at a, written as $Z \sim \delta_a$, if $\mathbb{P}(Z = a) = 1$ in which case*

$$F_Z(z) = \delta_a(z) = \begin{cases} 0 & \text{if } z < a \\ 1 & \text{if } z \geq a. \end{cases}$$

and the probability mass function is $f(x) = 1$ for $x = a$ and 0 otherwise.

When the limiting random variable is a point mass, we change the notation slightly. For example,

1. If $\mathbb{P}(X = c) = 1$ and $X_n \xrightarrow{\text{P}} X$ then we write $X_n \xrightarrow{\text{P}} c$.

2. If X_n **convergences to c in quadratic mean**, written $X_n \xrightarrow{\text{qm}} c$, if $\mathbb{E}(X_n - c)^2 \to 0$ as $n \to \infty$.

3. If X_n **convergences to c in distribution**, written $X_n \rightsquigarrow c$, if

$$\lim_{n\to\infty} F_n(t) = \delta_c(t)$$

for all $t \neq c$.

Suppose we are given a probability space $(\Omega, \mathcal{B}, P)$. We say a statement about random elements holds almost surely (a.s.) if there exists an event $N \in \mathcal{B}$ with $P(N) = 0$ such that the statement holds if $\omega \in N^c$. Alternatively, we may say the statement holds for a.a. (almost all) ω. The set N appearing the definition is sometimes called the exception set. Here are several examples of statements that hold a.s.:

1. If $\{X_n\}$ is a sequence of random variables, then $\lim_{n\to\infty} X_n$ exists a.s. means that there exists an event $N \in \mathcal{B}$, such that $P(N) = 0$ and if $\omega \in N^c$ then

$$\lim_{n\to\infty} X_n$$

exists. It also means that for a.a. ω,

$$\limsup_{n\to\infty} X_n(\omega) = \liminf_{n\to\infty} X_n(\omega).$$

We will write $\lim_{n\to\infty} X_n = X \quad a.s.$ or $X_n \overset{a.s.}{\to} X$, or $X_n \to X \quad a.s.$.

2. X_n **converges almost surely to a constant** c, written $X_n \overset{a.s.}{\to} c$ if there exists an event $N \in \mathcal{B}$, such that $P(N) = 0$ and if $\omega \in N^c$ then

$$\lim_{n\to\infty} X_n = c.$$

Example 2 *(Almost sure convergence) Let the sample space S be $[0,1]$ with the uniform probability distribution P. If the sample space S has elements denoted by s, then random variables $X_n(s)$ and $X(s)$ are all functions defined on S. Define $X_n(s) = s+s^n$ and $X(s) = s$. For every $s \in [0,1)$, $s^n \to 0$ as $n \to \infty$ and $X_n(s) \to s = X(s)$. However $X_n(1) = 2$ for every n so does not converge to $1 = X(1)$. Since the convergence occurs on the set $[0,1)$ and $P([0,1)) = 1$. $X_n \overset{a.s.}{\to} X$: that is, the function $X_n(s)$ converge to $X(s)$ for all $s \in S$ except for $s \in N = \{1\}$, where $N \subset S$ and $P(N) = 0$.*

See Example CB 5.5.7.

Example 3 Example CB 5.5.8 *Continuing Example 2. Let $S = [0,1]$. Let P be uniform on $[0,1]$. Let $X(s) = s$ and let*

$$\begin{aligned} X_1 &= s + I_{[0,1]}(s), & X_2 &= s + I_{[0,1/2]}(s), & X_3 &= s + I_{[1/2,1]}(s) \\ X_4 &= s + I_{[0,1/3]}(s), & X_5 &= s + I_{[1/3,2/3]}(s), & X_6 &= s + I_{[2/3,1]}(s) \end{aligned}$$

etc. It is straightforward to see that X_n converges to X in probability. As $n \to \infty$, $\mathbb{P}(|X_n - X| > \epsilon)$ is equal to the probability of an interval $[a_n, b_n]$ of s values whose length is going to 0. Then $X_n \xrightarrow{\text{P}} X$. However, X does not converge to X almost surely. Indeed, there is no value of $s \in S$ for which $X_n(s) \to s = X(s)$. For each s, the value $X_n(s)$ alternates between the values of s and $s+1$ infinitely often, that is, $X_n(s)$ does **not** *converge to $X(s)$. That is, no pointwise convergence occurs for this sequence.*

You do not really need to know the following Theorem 4 for this class.

Theorem 4 *$X_n \xrightarrow{\text{as}} X$ if and only if, for every $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(\sup_{m\geq n} |X_m - X| \leq \epsilon) = 1.$$

Example 5 *(Convergence in distribution) Let $X_n \sim N(0, 1/n)$. Intuitively, X_n is concentrating at 0 so we would like to say that X_n converges to 0. Let's see if this is true. Note that $\sqrt{n}X_n \sim N(0,1)$. Let F be the distribution function for a point mass at 0:*

$$\mathbb{P}(X = 0) = 1.$$

Let Z denote a standard normal random variable. For $t < 0$,

$$F_n(t) = \mathbb{P}(X_n < t) = \mathbb{P}(\sqrt{n}X_n < \sqrt{n}t) = \mathbb{P}(Z < \sqrt{n}t) \to 0$$

since $\sqrt{n}t \to -\infty$. For $t > 0$,

$$F_n(t) = \mathbb{P}(X_n < t) = \mathbb{P}(\sqrt{n}X_n < \sqrt{n}t) = \mathbb{P}(Z < \sqrt{n}t) \to 1$$

since $\sqrt{n}t \to \infty$. Hence, $F_n(t) \to F(t)$ for all $t \neq 0$ and so $X_n \rightsquigarrow 0$.

Notice that $F_n(0) = 1/2 \neq F(0) = 1$ so convergence fails at $t = 0$. That doesn't matter because $t = 0$ is not a continuity point of F and the definition of convergence in distribution only requires convergence at continuity points.

Now convergence in probability follows from Theorem 6 (c):

$$X_n \xrightarrow{\text{P}} 0.$$

Here we also provides a direct proof. For any $\epsilon > 0$, using Markov's inequality,

$$\mathbb{P}(|X_n| > \epsilon) \quad = \quad \mathbb{P}(|X_n|^2 > \epsilon^2) \leq \frac{\mathbb{E}(X_n^2)}{\epsilon^2} = \frac{\frac{1}{n}}{\epsilon^2} \to 0$$

as $n \to \infty$. ■

Theorem 6 *The following relationships hold:*

(a) $X_n \xrightarrow{\text{qm}} X$ implies that $X_n \xrightarrow{\text{P}} X$.

(b) $X_n \xrightarrow{\text{P}} X$ implies that $X_n \rightsquigarrow X$.

(c) If $X_n \rightsquigarrow X$ and if $\mathbb{P}(X = c) = 1$ for some real number c, then $X_n \xrightarrow{\text{P}} X$.

(d) $X_n \xrightarrow{\text{as}} X$ implies $X_n \xrightarrow{\text{P}} X$.

In general, none of the reverse implications hold except the special case in (c).

We will show proof of Theorem 6(a)– (c) in a second after seeing another example.

Example 7 *Let $X_1, \ldots, X_n \sim \text{Uniform}(0, 1)$. Let $X_{(n)} = \max_i X_i$. First we claim that $X_{(n)} \xrightarrow{\text{P}} 1$. This follows since*

$$\mathbb{P}(|X_{(n)} - 1| > \epsilon) = \mathbb{P}(X_{(n)} \leq 1 - \epsilon) = \prod_i \mathbb{P}(X_i \leq 1 - \epsilon) = (1 - \epsilon)^n \to 0.$$

Also

$$\mathbb{P}(n(1 - X_{(n)}) \leq t) = 1 - \mathbb{P}(X_{(n)} \leq 1 - (t/n)) = 1 - (1 - t/n)^n \to 1 - e^{-t}.$$

So $n(1 - X_{(n)}) \rightsquigarrow \text{Exp}(1)$.

Proof of Theorem 6. We start by proving (a). Suppose that $X_n \xrightarrow{\text{qm}} X$. Fix $\epsilon > 0$. Then, using Markov's inequality,

$$\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(|X_n - X|^2 > \epsilon^2) \leq \frac{\mathbb{E}|X_n - X|^2}{\epsilon^2} \to 0.$$

Proof of (b). Fix $\epsilon > 0$ and let x be a continuity point of F. Then

$$\begin{aligned} F_n(x) &= \mathbb{P}(X_n \leq x) = \mathbb{P}(X_n \leq x, X \leq x + \epsilon) + \mathbb{P}(X_n \leq x, X > x + \epsilon) \\ &\leq \mathbb{P}(X \leq x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon) \\ &= F(x + \epsilon) + \mathbb{P}(|X_n - X| > \epsilon). \end{aligned}$$

Also,

$$\begin{aligned} F(x-\epsilon) &= \mathbb{P}(X \le x-\epsilon) = \mathbb{P}(X \le x-\epsilon, X_n \le x) + \mathbb{P}(X \le x-\epsilon, X_n > x) \\ &\le F_n(x) + \mathbb{P}(|X_n - X| > \epsilon). \end{aligned}$$

Hence,

$$F(x-\epsilon) - \mathbb{P}(|X_n - X| > \epsilon) \le F_n(x) \le F(x+\epsilon) + \mathbb{P}(|X_n - X| > \epsilon).$$

Take the limit as $n \to \infty$ to conclude that

$$F(x-\epsilon) \le \liminf_{n\to\infty} F_n(x) \le \limsup_{n\to\infty} F_n(x) \le F(x+\epsilon).$$

This holds for all $\epsilon > 0$. Take the limit as $\epsilon \to 0$ and use the fact that F is continuous at x and conclude that $\lim_n F_n(x) = F(x)$.

Proof of (c). Fix $\epsilon > 0$. Then,

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

Warning!

- Convergence in probability does not imply convergence in quadratic mean.

  Let $U \sim \text{Unif}(0,1)$ and let $X_n = \sqrt{n} I_{(0,1/n)}(U)$. Then

  $$\mathbb{P}(|X_n| > \epsilon) = \mathbb{P}(\sqrt{n} I_{(0,1/n)}(U) > \epsilon) = \mathbb{P}(0 \le U < 1/n) = 1/n \to 0.$$

  Hence, $X_n \xrightarrow{\text{P}} 0$. But

  $$\mathbb{E}(X_n^2) = n \int_0^{1/n} du = 1$$

  for all n so X_n does not converge in quadratic mean.

- Convergence in distribution does not imply convergence in probability.

  Let $X \sim N(0,1)$. Let $X_n = -X$ for $n = 1, 2, 3, \ldots$; hence $X_n \sim N(0,1)$. X_n has the same distribution function as X for all n so, trivially, $\lim_n F_n(x) = F(x)$ for all x. Therefore, $X_n \rightsquigarrow X$. But $\mathbb{P}(|X_n - X| > \epsilon) = \mathbb{P}(|2X| > \epsilon) = \mathbb{P}(|X| > \epsilon/2) \ne 0$. So X_n does not converge to X in probability.

- One might conjecture that if $X_n \xrightarrow{\text{P}} b$, then $\mathbb{E}(X_n) \to b$. This is not true.
  - Let X_n be a random variable defined by $\mathbb{P}(X_n = n^2) = 1/n$ and $\mathbb{P}(X_n = 0) = 1 - (1/n)$.
  - Now, $\mathbb{P}(|X_n| < \epsilon) = \mathbb{P}(X_n = 0) = 1 - (1/n) \to 1$. Hence, $X_n \xrightarrow{\text{P}} 0$.
  - However, $\mathbb{E}(X_n) = [n^2 \times (1/n)] + [0 \times (1 - (1/n))] = n$.
  - Thus, $\mathbb{E}(X_n) \to \infty$.

2 Review on Limit Theorems

Some convergence properties are preserved under transformations.

Theorem 8 *Let X_n, X, Y_n, Y be random variables. Let g be a continuous function.*
(a) If $X_n \xrightarrow{\text{P}} X$ and $Y_n \xrightarrow{\text{P}} Y$, then $X_n + Y_n \xrightarrow{\text{P}} X + Y$.
(b) If $X_n \xrightarrow{\text{qm}} X$ and $Y_n \xrightarrow{\text{qm}} Y$, then $X_n + Y_n \xrightarrow{\text{qm}} X + Y$.
(c) If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, then $X_n + Y_n \rightsquigarrow X + c$.
(d) If $X_n \xrightarrow{\text{P}} X$ and $Y_n \xrightarrow{\text{P}} Y$, then $X_n Y_n \xrightarrow{\text{P}} XY$.
(e) If $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow c$, then $X_n Y_n \rightsquigarrow cX$.
(f) If $X_n \xrightarrow{\text{P}} X$, then $g(X_n) \xrightarrow{\text{P}} g(X)$.
(g) If $X_n \rightsquigarrow X$, then $g(X_n) \rightsquigarrow g(X)$.

- Parts (c) and (e) are know as **Slutzky's theorem**
- Parts (f) and (g) are known as **The Continuous Mapping Theorem**.
- It is worth noting that $X_n \rightsquigarrow X$ and $Y_n \rightsquigarrow Y$ does not in general imply that $X_n + Y_n \rightsquigarrow X + Y$.

3 The Law of Large Numbers

The LLN says that the mean of a large sample is close to the mean of the distribution. For example, the proportion of heads of a large number of tosses of a fair coin is expected to be close to 1/2. We now make this more precise.

Let $X_1, X_2, \ldots$ be an IID sample, let $\mu = \mathbb{E}(X_1)$ and $\sigma^2 = \text{Var}(X_1)$. Recall that the sample mean is defined as $\overline{X}_n = n^{-1} \sum_{i=1}^n X_i$ and that $\mathbb{E}(\overline{X}_n) = \mu$ and $\text{Var}(\overline{X}_n) = \sigma^2/n$.

Theorem 9
The Weak Law of Large Numbers (WLLN).
If $X_1, \ldots, X_n$ are IID, then $\overline{X}_n \xrightarrow{\text{P}} \mu$. Thus, $\overline{X}_n - \mu = o_P(1)$.

Interpretation of the WLLN: The distribution of $\overline{X}_n$ becomes more concentrated around μ as n gets large.

Proof. Assume that $\sigma < \infty$. This is not necessary but it simplifies the proof. Using Chebyshev's inequality,

$$\mathbb{P}\left(|\overline{X}_n - \mu| > \epsilon\right) \leq \frac{\mathrm{Var}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

which tends to 0 as $n \to \infty$. ■

Example 10 *Consider flipping a coin for which the probability of heads is p. Let X_i denote the outcome of a single toss (0 or 1). Hence, $p = P(X_i = 1) = E(X_i)$. The fraction of heads after n tosses is $\overline{X}_n$.*

According to the law of large numbers, $\overline{X}_n$ converges to p in probability. This does not mean that $\overline{X}_n$ will numerically equal p. It means that, when n is large, the distribution of $\overline{X}_n$ is tightly concentrated around p.

Suppose that $p = 1/2$. How large should n be so that $P(.4 \leq \overline{X}_n \leq .6) \geq .7$? First, $\mathbb{E}(\overline{X}_n) = p = 1/2$ and $\mathrm{Var}(\overline{X}_n) = \sigma^2/n = p(1-p)/n = 1/(4n)$. From Chebyshev's inequality,

$$\begin{aligned}\mathbb{P}(.4 \leq \overline{X}_n \leq .6) &= \mathbb{P}(|\overline{X}_n - \mu| \leq .1) \\ &= 1 - \mathbb{P}(|\overline{X}_n - \mu| > .1) \\ &\geq 1 - \frac{1}{4n(.1)^2} = 1 - \frac{25}{n}.\end{aligned}$$

The last expression will be larger than .7 if $n = 84$. ■

Theorem 11 The Strong Law of Large Numbers. *We have*

$$\overline{X}_n \xrightarrow{\text{as}} \mu.$$

The Central Limit Theorem.

The law of large numbers says that the distribution of $\overline{X}_n$ piles up near μ. This isn't enough to help us approximate probability statements about $\overline{X}_n$. For this we need the central limit theorem.

Suppose that $X_1, \ldots, X_n$ are IID with mean μ and variance σ^2. The central limit theorem (CLT) says that $\overline{X}_n = n^{-1}\sum_i X_i$ has a distribution which is approximately Normal with mean μ and variance σ^2/n. This is remarkable since nothing is assumed about the distribution of X_i, except the existence of the mean and variance. For instance, the CLT applies even if X is a coin toss. Although a Bernoulli distribution is far from normal, the mean of a sequence of Bernoulli experiments is normally distributed (for a large number of tosses).

Theorem 12 (The Central Limit Theorem (CLT)) *Let $X_1, \ldots, X_n$ be* IID *with mean μ and variance σ^2. Let $\overline{X}_n = n^{-1} \sum_{i=1}^{n} X_i$. Then*

$$Z_n \equiv \frac{\overline{X}_n - \mu}{\sqrt{\mathrm{Var}(\overline{X}_n)}} = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \rightsquigarrow Z$$

where $Z \sim N(0,1)$. In other words,

$$\lim_{n\to\infty} \mathbb{P}(Z_n \le z) = \Phi(z) = \int_{-\infty}^{z} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx.$$

> Interpretation: Probability statements about $\overline{X}_n$ can be approximated using a Normal distribution. It's the probability statements that we are approximating, not the random variable itself.

In addition to $Z_n \rightsquigarrow N(0,1)$, there are several forms of notation to denote the fact that the distribution of Z_n is converging to a Normal. They all mean the same thing. Here they are:

$$\begin{aligned}
Z_n &\approx N(0,1) \\
\overline{X}_n &\approx N\left(\mu, \frac{\sigma^2}{n}\right) \\
\overline{X}_n - \mu &\approx N\left(0, \frac{\sigma^2}{n}\right) \\
\sqrt{n}(\overline{X}_n - \mu) &\approx N\left(0, \sigma^2\right) \\
\frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} &\approx N(0,1).
\end{aligned}$$

Recall that if X is a random variable, its moment generating function (MGF) is $\psi_X(t) = \mathbb{E}e^{tX}$. Assume in what follows that the MGF is finite in a neighborhood around $t = 0$.

Lemma 13 *Let $Z_1, Z_2, \ldots$ be a sequence of random variables. Let ψ_n be the* MGF *of Z_n. Let Z be another random variable and denote its* MGF *by ψ. If $\psi_n(t) \to \psi(t)$ for all t in some open interval around 0, then $Z_n \rightsquigarrow Z$.*

Recall

- n'th moment: $E\left(X^n\right)$
- Central moments: $E\left((X-\mu)^n\right)$.

Proof of the central limit theorem. Let

$$Y_i = (X_i - \mu)/\sigma.$$

Then,

$$Z_n = n^{-1/2} \sum_i Y_i.$$

Let $\psi(t)$ be the MGF of Y_i. The MGF of $\sum_i Y_i$ is $(\psi(t))^n$ and MGF of Z_n is $[\psi(t/\sqrt{n})]^n \equiv \xi_n(t)$. Now

$$\begin{aligned}\psi'(0) &= \mathbb{E}(Y_1) = 0\\ \psi''(0) &= \mathbb{E}(Y_1^2) = \mathrm{Var}(Y_1) = 1.\end{aligned}$$

So,

$$\begin{aligned}\psi(t) &= \psi(0) + t\psi'(0) + \frac{t^2}{2!}\psi''(0) + \frac{t^3}{3!}\psi'''(0) + \cdots\\ &= 1 + 0 + \frac{t^2}{2} + \frac{t^3}{3!}\psi'''(0) + \cdots\\ &= 1 + \frac{t^2}{2} + \frac{t^3}{3!}\psi'''(0) + \cdots\end{aligned}$$

Now,

$$\begin{aligned}\xi_n(t) &= \left[\psi\left(\frac{t}{\sqrt{n}}\right)\right]^n\\ &= \left[1 + \frac{t^2}{2n} + \frac{t^3}{3!n^{3/2}}\psi'''(0) + \cdots\right]^n\\ &= \left[1 + \frac{\frac{t^2}{2} + \frac{t^3}{3!n^{1/2}}\psi'''(0) + \cdots}{n}\right]^n\\ &\to e^{t^2/2}\end{aligned}$$

which is the MGF of a N(0,1). The result follows from the previous Theorem. In the last step we used the fact that if $a_n \to a$ then

$$\left(1 + \frac{a_n}{n}\right)^n \to e^a. \quad \blacksquare$$

The central limit theorem tells us that
$Z_n = \sqrt{n}(\overline{X}_n - \mu)/\sigma$ is approximately N(0,1).
However, we rarely know σ. Later, we will see that we can estimate σ^2 from $X_1, \ldots, X_n$ by

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2.$$

This raises the following question: if we replace σ with S_n, is the central limit theorem still true? The answer is yes.

Theorem 14 *Assume the same conditions as the CLT. Then,*

$$T_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{S_n} \rightsquigarrow N(0,1).$$

PROOF. We have that

$$T_n = Z_n W_n$$

where

$$Z_n = \frac{\sqrt{n}(\overline{X}_n - \mu)}{\sigma} \text{ and } W_n = \frac{\sigma}{S_n}.$$

Now $Z_n \rightsquigarrow N(0,1)$ and $W_n \xrightarrow{\text{P}} 1$. The result follows from Slutzky's theorem. ■