Note 08. Modes of convergence

Alex Fu

Fall 2022

Now that we are able to characterize individual and finite collections of random variables, let us consider countably infinite collections — sequences of random variables. The first and foremost ideas that arise in the study of sequences are the *limit* and the related *convergence*.

Definition 1 (Limit; convergence; pointwise convergence).

A sequence of real numbers $(x_n)_{n=1}^{\infty}$ is said to **converge** to a **limit** $x \in \mathbb{R}$, denoted $x_n \to x$, if "eventually, any deviation can be made arbitrarily small." That is, for any $\varepsilon > 0$,

$$\text{there exists } N \text{ such that for all } n \geq N,\ |x_n - x| < \varepsilon.$$

The classic mode of convergence of a sequence of real-valued functions $(f_n)_{n=1}^{\infty}$ is **pointwise convergence**: "at every point t in the domain, $f_n(t) \to f(t)$ as real numbers."

Throughout, let $\varepsilon > 0$, let $(\Omega, \mathcal{F}, \mathbb{P})$ be a shared probability space, and let $(X_n)_{n=1}^{\infty}$ be a sequence of real-valued random variables. $X_n : \Omega \to \mathbb{R}$ do have a sense of possible pointwise convergence, but here, we will explore some more common and more interesting **modes of convergence** of random variables which involve probability:

$$\text{almost surely} \implies \text{in probability} \implies \text{in distribution}.$$

1 Almost sure convergence

Definition 2 (Almost sure convergence).

$(X_n)_{n=1}^{\infty}$ converges **almost surely** (a.s.) to X if

$$\mathbb{P}\Big(\lim_{n\to\infty} X_n = X\Big) = \mathbb{P}\Big(\Big\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big\}\Big) = 1.$$

This is denoted $X_n \overset{\text{a.s.}}{\to} X$. Equivalently, $X_n \overset{\text{a.s.}}{\to} X$ if $\mathbb{P}(\lim_{n\to\infty} X_n \neq X) = 0$.

Example 1 (Convergence almost surely but not pointwise).

Let $\Omega = [0,1]$, and let $\mathbb{P}$ be uniform, so the probability of an event is its length. Then

$$X_n(\omega) = \begin{cases} n & \text{if } \omega \in [0, \frac{1}{n}] \\ 0 & \text{otherwise} \end{cases}$$

converges almost surely to the constant 0. Show that $X_n(\omega)$ converges to $X(\omega)$ pointwise on the set of outcomes $(0,1]$, which has probability 1.

Almost sure convergence is usually the strongest form of convergence we will work with, in line with the philosophy of "caring up to almost surely," or up to "convergence with probability 1." However, it may be difficult to show a.s. convergence from its definition in general, so we present two commonly used results.

Theorem 1 (Strong Law of Large Numbers).

If $(X_n)_{n=1}^{\infty}$ are i.i.d. random variables with finite mean $\mu = \mathbb{E}(X_1)$, then the sample mean $\bar{X}_n$ converges almost surely to the true mean. That is,

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mu.$$

The SLLN is celebrated due to its very weak assumptions and its very strong conclusion of a.s. You may find a proof of a slightly weaker form of the SLLN assigned to you as homework, involving fourth moments and the following general result:

Definition 3 (Infinitely often).

Let $\{A_n\}_{n=1}^{\infty}$ be a countably infinite collection of events. Then the event that A_n happens **infinitely often** is

$$A_n \text{ i.o.} := \limsup_{n\to\infty} A_n = \bigcap_{n=1}^{\infty} \bigcup_{k\geq n} A_k.$$

Equivalently, $\omega \in A_n$ i.o. if for every $n \geq 1$, there is a greater index $k \geq n$ such that $\omega \in A_k$. So $\omega \notin A_n$ i.o. if there exists a maximum index N so that $\omega \notin A_k$ for any $k \geq N$ — that is, if ω appears in A_n only *finitely often*.

Theorem 2 (Borel–Cantelli Lemma).

If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$.

If $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and $(A_n)_{n=1}^{\infty}$ are independent, then $\mathbb{P}(A_n \text{ i.o.}) = 1$.

Example 2 (Common application of Borel–Cantelli).

Suppose that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, where

$$A_n := \{\omega : |X_n(\omega) - X(\omega)| \geq \varepsilon\}.$$

Then A_n i.o. is precisely the event that $X_n(\omega)$ diverges. By the first Borel–Cantelli lemma,

$$\mathbb{P}(A_n \text{ i.o.}) = \mathbb{P}\Big(\lim_{n\to\infty} X_n \neq X\Big) = 0,$$

which implies that $X_n \xrightarrow{\text{a.s.}} X$.

Example 3 (More examples of almost sure convergence).

- In a discrete-time Markov chain, the proportion of time spent in a state converges a.s. to the inverse of the expected time it takes to revisit said state (given a few assumptions).
- The *asymptotic equipartition property*: if X_n are i.i.d. over a finite alphabet, then the average surprise $-\frac{1}{n}\log_2 p(X_1, \dots, X_n)$ converges a.s. to the entropy $H(X)$.
- In machine learning, we can ask if the iterates of the *stochastic gradient descent* algorithm converge a.s. to the true minimizer of the given function.

2 Convergence in probability

Definition 4 (Convergence in probability).

$(X_n)_{n=1}^{\infty}$ converges **in probability** (i.p.) to X if for every $\varepsilon > 0$,

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0.$$

This is denoted $X_n \xrightarrow{\mathbb{P}} X$.

Example 4 (Convergence in probability but not almost surely).

Let $X_n \sim n \cdot \text{Bernoulli}(\frac{1}{n})$, $n \geq 1$ be independent, and let $X = 0$. Then X_n converges to X in probability but not almost surely.

Let $0 < \varepsilon \leq 1$ without loss of generality. Then the *probability of deviation* $\mathbb{P}(|X_n - X| \geq \varepsilon) = \frac{1}{n} \to 0$, which shows convergence i.p.

However, the events $A_n = \{|X_n - X| \geq \varepsilon\}$ are independent, and $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, so by the second Borel–Cantelli lemma, X_n in fact *diverges almost surely*!

Example 1 and Example 4 should tell us to carefully distinguish convergence a.s. and i.p.: the assumption of independence is enough to change convergence with probability 1 to *divergence* with probability 1. In the other direction, however,

Proposition 1.

Almost sure convergence implies convergence in probability.

The idea behind the proof: An entry is filled if $|X_n(\omega) - X(\omega)| \geq \varepsilon$. As $n \to \infty$, the total filled area $\mathbb{P}(A_n)$ tends to 0 by convergence a.s., so the area of its leftmost column $\mathbb{P}(B_n)$ tends to 0 as well — which is precisely convergence i.p.

$$
\begin{array}{c|c|c|c|c}
 & X_n & X_{n+1} & X_{n+2} & \cdots \\
\hline
 & * & & * & \cdots \\
 & * & * & & \cdots \\
\Omega & * & & & \cdots \\
 & & & & \cdots \\
 & & * & * & \cdots \\
\multicolumn{5}{c}{\vdots}
\end{array}
$$

Proof. Let $X_n \overset{\text{a.s.}}{\to} X$. For $n \geq 1$, we define the events

$$
\begin{aligned}
A_n &:= \{\omega : \text{for some } m \geq n,\ |X_m(\omega) - X(\omega)| \geq \varepsilon\}, \\
B_n &:= \{\omega : |X_n(\omega) - X(\omega)| \geq \varepsilon\}.
\end{aligned}
$$

To show that $\mathbb{P}(B_n) \to 0$, it suffices to show that $\mathbb{P}(A_n) \to 0$, because $B_n \subseteq A_n = \bigcup_{m \geq n} B_m$. We observe that $\lim_{n\to\infty} A_n$ is defined to be the event that X_n diverges, so by convergence a.s., $\mathbb{P}(\lim_{n\to\infty} A_n) = 0$. Then, we are finished by

$$
A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots \implies \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\Big(\lim_{n\to\infty} A_n\Big) = 0.
$$

□

Convergence i.p. is easier to prove directly from its definition than a.s. convergence. You may come across several examples in your homework; we will give a more famous example below.

Theorem 3 (Weak Law of Large Numbers).

Let $(X_n)_{n=1}^{\infty}$ be as described in the SLLN. Then the sample mean $\bar{X}_n$ converges in probability to the true mean:

$$
\bar{X}_n \overset{\mathbb{P}}{\to} \mu.
$$

The WLLN does hold with greater generality and weaker assumptions than the SLLN, so it can be more than a corollary of the SLLN and convergence almost surely $\implies$ in probability.

3 Convergence in distribution

Definition 5 (Convergence in distribution).

$(X_n)_{n=1}^{\infty}$ converges **in distribution** (i.d.) to X if for every $x \in \mathbb{R}$ with $\mathbb{P}(X = x) = 0$,

$$\lim_{n\to\infty} \mathbb{P}(X_n \leq x) = \mathbb{P}(X \leq x).$$

This is denoted $X_n \xrightarrow{\mathsf{d}} X$. Equivalently,

- $X_n \xrightarrow{\mathsf{d}} X$ for integer-valued X_n iff the pmfs p_{X_n} converge pointwise to p_X.
- $X_n \xrightarrow{\mathsf{d}} X$ for continuous X_n if (but not only if) the pdfs f_{X_n} converge pointwise to f_X.

Convergence i.d. is a weaker form of convergence: "eventually, the values of X_n resemble values drawn from the distribution of X." But it says nothing about the actual values drawn from X_n and X, and in particular their deviations. We illustrate this idea in the following example.

Example 5 (Convergence in distribution but not in probability).

Let X be distributed as any continuous symmetric probability distribution, and let

$$X_{2k} = X \text{ and } X_{2k+1} = -X.$$

Then $X_n \xrightarrow{\mathsf{d}} X$ trivially, but the probability of deviation $\mathbb{P}(|X_n - X| \geq \varepsilon)$ oscillates between 0 and some nonzero value, so it cannot converge to 0.

In the other direction, however,

Proposition 2.

Convergence in probability implies convergence in distribution.

The key idea behind the proof: $\mathbb{P}(|X_n - X| \geq \varepsilon) \to 0$ allows us to approximate the event $\{X_n \leq x\}$ by $\{X \leq x - \varepsilon\}$ and $\{X \leq x + \varepsilon\}$.

Proof. Suppose that $X_n \xrightarrow{\mathbb{P}} X$. We can observe graphically that for all $n \geq 1$,

$$\mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x + \varepsilon) + \mathbb{P}(|X_n - X| \geq \varepsilon)$$
$$\mathbb{P}(X \leq x - \varepsilon) \leq \mathbb{P}(X_n \leq x) + \mathbb{P}(|X_n - X| \geq \varepsilon).$$

As n tends to infinity, convergence i.p. gives us the inequality

$$\mathbb{P}(X + \varepsilon \leq x) \leq \lim_{n\to\infty} \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X - \varepsilon \leq x).$$

$\varepsilon > 0$ can be made arbitrarily small, so we have shown that $\mathbb{P}(X_n \leq x) \to \mathbb{P}(X \leq x)$. $\square$

The most common example of convergence i.d. is one of the most ubiquitous theorems in statistics and common justification for the importance of the normal distribution. Similar results exist for other statistical distributions, such as *chi-squared* or *Student's* t.

Theorem 4 (Central Limit Theorem).

If $(X_n)_{n=1}^{\infty}$ are i.i.d. with mean μ and variance σ^2, then the standard score of the sample mean $\bar{X}_n$ converges in distribution to the standard normal distribution. That is,

$$\frac{\bar{X}_n - \mu}{\sqrt{\frac{\sigma^2}{n}}} \xrightarrow{\mathsf{d}} \mathcal{N}(0,1).$$

At first glance, the following two results of convergence i.d. may appear to contradict the CLT, but their random variables are *not* identically distributed. We leave their proofs as discussion problems or exercises; you may find the following identity helpful.

$$\lim_{n\to\infty} \left(1 - \frac{\lambda}{n}\right)^n = e^{-\lambda}.$$

Theorem 5 (Poisson limit theorem).

Suppose that $\lim_{n\to\infty} np_n = \lambda > 0$. Then

$$X_n \sim \text{Binomial}(n, p_n) \xrightarrow{\mathsf{d}} X \sim \text{Poisson}(\lambda).$$

The Poisson limit theorem is also called the *law of rare events*, and it justifies the use of Poisson distributions in modelling rare occurrences. Many other situations, including common random graph models or balls and bins models, also have Poisson limits.

Theorem 6 (Limit of geometric distribution).

Suppose that $\lim_{n\to\infty} \frac{p_n}{n} = \lambda > 0$. Then

$$X_n \sim \text{Geometric}(p_n) \xrightarrow{\mathsf{d}} X \sim \text{Exponential}(\lambda).$$

4 Convergence in expectation*

Definition 6 (Convergence in expectation).

$(X_n)_{n=1}^{\infty}$ converges **in expectation**, *in mean*, or *in L^1 norm* to X if

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X), \text{ equivalently } \lim_{n\to\infty} \mathbb{E}(|X_n - X|) = 0.$$

This is denoted $X_n \xrightarrow{\mathbb{E}} X$.

This mode of convergence is somewhat anomalous for us: we might expect that convergence in distribution implies the same in expectation, because "expectation is a feature of the distribution." However, in general, none of convergence a.s., i.p., or i.d. imply convergence in expectation.

Example 6 (Convergence almost surely but not in expectation).

Consider $X_n \xrightarrow{\text{a.s.}} X$ as in Example 1, so that X_n converges i.p. and i.d. to X as well. Then $\mathbb{E}(X_n) = n \cdot \frac{1}{n} = 1$ for every $n \geq 1$, yet $\mathbb{E}(X) = 0$, so $\mathbb{E}(X_n) \not\to \mathbb{E}(X)$.

Expectation is equivalent up to almost sure equivalence, so while $X_n \xrightarrow{\text{a.s.}} X$ does imply that

$$\mathbb{E}\Big(\lim_{n\to\infty} X_n\Big) = \mathbb{E}(X),$$

we cannot always exchange the $\mathbb{E}$ operator and the $\lim$ sign. The key to the divergence above is that even as the probability of eventual deviation tends to zero, the *value* of said deviation can tend much faster to infinity. If instead $X_n \sim 2^n \cdot \text{Bernoulli}(\frac{1}{n})$, then $\mathbb{E}(X_n) = \frac{2^n}{n} \to \infty$!

If we suppose that $X_n \xrightarrow{\text{a.s.}} X$, then there does turn out to be two quite strong conditions that imply convergence in expectation:

a. **Monotone convergence theorem**. If $0 \leq X_1 \leq X_2 \leq X_3 \leq \cdots$, then $X_n \xrightarrow{\mathbb{E}} X$.

b. **Dominated convergece theorem**. If there exists $Y \geq 0$ with $\mathbb{E}(|Y|) < \infty$ such that $|X_n| \leq |Y|$ for all n and $|X| \leq |Y|$, then $X_n \xrightarrow{\mathbb{E}} X$.

The convergence of sequences of functions is in general quite complex, and is explored in depth in fields such as measure theory or functional analysis. We will only state a few relevant, interesting results without proof.

c. **Continuous mapping theorem**. If f is continuous, often $\log$ or $\exp$, then f preserves convergence a.s., i.p., and i.d.: if $X_n \to X$, then $f(X_n) \to f(X)$ in the same manner.

d. **Slutsky's theorem**. If $X_n \xrightarrow{\text{d}} X$ and $Y_n \xrightarrow{\mathbb{P}} c$ for some constant c, then $X_n + Y_n \xrightarrow{\text{d}} X + c$, $X_nY_n \xrightarrow{\text{d}} cX$, and $X_n/Y_n \xrightarrow{\text{d}} X/c$.

e. Convergence in expectation implies convergence in probability.

f. Convergence in distribution implies that if g is bounded and continuous, $\mathbb{E}(g(X_n)) \to \mathbb{E}(g(X))$. [Note: $g(x) = x$ is not bounded, so convergence in expectation is not implied.]

g. Convergence in L^2 norm implies convergence in L^1 norm. In general, convergence in L^p norm implies convergence in L^q norm for all $1 \leq q \leq p < \infty$.

Definition 7 (Convergence in L^p).

$(X_n)_{n=1}^{\infty}$ converges **in L^p norm** or *in p-norm* to X for $p \geq 1$ if

$$\lim_{n \to \infty} \mathbb{E}(|X_n - X|^p) = 0.$$

This is sometimes denoted $X_n \xrightarrow{L^p} X$.

A particular L^p space, the *Hilbert space of random variables* $\mathcal{H} = L^2(\Omega; \mathbb{R})$, will resurface and feature prominently in later notes to solve several problems in *estimation*.

■