Modes of Convergence

EECS 126 at UC Berkeley

Spring 2022

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

The classical convergence of a sequence of real numbers $(x_n)_{n=1}^{\infty}$ to a limit x is described by "eventually $[\forall n \geq N]$, any deviation $[|x_n - x|]$ can be made arbitrarily small $[< \varepsilon]$." The classical convergence of a sequence of real-valued functions $(f_n)_{n=1}^{\infty}$ is the exact same idea — *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. We note that $X_n : \Omega \to \mathbb{R}$ are also real-valued functions, so they have a notion of pointwise convergence (to some random variable $X : \Omega \to \mathbb{R}$). The points are $\omega \in \Omega$, reflecting the fact that we can evaluate $X_1, X_2, X_3, \ldots$ at different ω to get different "realizations" of the sequence.

But rarely can we say that a statement is true for every single outcome $\omega \in \Omega$ — such is far too "deterministic," and does not involve the probability measure $\mathbb{P}$ at all. The next strongest form of convergence available to us, then, is that $X_n \to X$ with complete certainty, i.e. with probability 1, or *almost surely*. Yet, as we might expect, introducing probability into convergence results in some additional complexity.

In this note we discuss a few common **modes of convergence** of sequences of random variables: almost sure convergence, convergence in probability, and convergence in distribution. It turns out there is a chain of strict implications for modes of convergence:

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

We also give some prominent examples in the Strong Law of Large Numbers, the Weak Law of Large Numbers, the Central Limit Theorem, the Borel-Cantelli lemma, and the Poisson Limit Theorem.

1 Almost sure convergence

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

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

This is denoted $X_n \xrightarrow{\text{a.s.}} X$. An equivalent definition is if $\mathbb{P}(\lim_{n\to\infty} X_n \neq X) = 0$.

Example 1. Let $\Omega = [0,1]$, and let $\mathbb{P}$ be the uniform probability measure on Ω so that $\mathbb{P}([a,b]) = b - a$. Define $X_n \sim n \cdot \mathsf{Bernoulli}(\frac{1}{n})$, more specifically $X_n(\omega) = n \cdot \mathbb{1}_{\omega \in [0, \frac{1}{n}]}$, and $X(\omega) = 0$ for all $\omega \in \Omega$. Then $X_n \xrightarrow{\text{a.s.}} X$.

i. For every nonzero ω, there is some $N \in \mathbb{N}$ such that $\frac{1}{N} < \omega$. [This is the *Archimedean property* of $\mathbb{R}$.] In other words, for all $n \geq N$, ω falls outside of $[0, \frac{1}{n}]$, so $X_n(\omega) = 0$. Thus $X_n(\omega) \to X(\omega)$.

ii. However, at $\omega = 0$, $X_n(0) = n$ for every $n \geq 1$, but $X(0) = 0$, so $X_n(0) \nrightarrow X(0)$.

Therefore $\mathbb{P}(\{\omega \in \Omega : \lim_{n\to\infty} X_n = X\}) = \mathbb{P}(\Omega \setminus \{0\}) = \mathbb{P}((0,1]) = 1$, so we have shown a.s. convergence.

Almost sure convergence is the strongest form of convergence we usually work with. Its statement is fairly strong already — "the probability of convergence is 1." The concept of *almost surely* also exists more generally: a key philosophy in probability theory (and measure theory) is to *disregard sets with zero measure*. If two random variables only disagree on a null set, $\mathbb{P}(X \neq Y) = 0$, i.e. $\mathbb{P}(X = Y) = 1$, then they are equal almost surely. The probability of an event being 1 is more important than whether it includes every single outcome $\omega \in \Omega$.

It may be difficult to show a.s. convergence in general, so we present two main results that are commonly used to prove a.s. convergence.

Theorem 1 (Strong Law of Large Numbers)**.** *If $(X_n)_{n=1}^{\infty}$ are independent and identically distributed (i.i.d.) with finite mean $|\mathbb{E}(X_1)| < \infty$, then the sample mean $\bar{X}_n$ converges almost surely to the true mean. That is,*

$$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{\text{a.s.}} \mathbb{E}(X_1).$$

The SLLN is celebrated in part due to its very weak assumptions, and in part due to its strong conclusion (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.

Theorem 2 (Borel-Cantelli Lemma)**.** *Let $(A_n)_{n=1}^{\infty}$ be a collection of events. The event that A_n happens infinitely often is given by*

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

[If $\omega \in A_n$ i.o., then for every $n \geq 1$, there exists a greater index $k \geq n$ such that $\omega \in A_k$. Otherwise, there is a maximum index N so that $\omega \notin A_k$ for any $k \geq N$, i.e. ω only appears in finitely many A_n.]

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

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

If we define $A_n := \{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geq \varepsilon\}$, then we can verify that A_n i.o. is the event the sequence diverges, $\{\lim_{n\to\infty} X_n(\omega) \neq X(\omega)\}$! So, if $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$, i.e. $X_n \xrightarrow{\text{a.s.}} X$. Thus the first Borel-Cantelli lemma is a common way of proving almost sure convergence.

As examples, we may also see almost sure convergence in the following contexts:

- 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).
- If $(X_n)_{n=1}^{\infty}$ are i.i.d. over a finite alphabet, then the average surprise $-\frac{1}{n}\log_2 p(X_1,\ldots,X_n)$ converges a.s. to the entropy $H(X)$. This is called the *asymptotic equipartition property.*
- 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 2. $(X_n)_{n=1}^{\infty}$ 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 $X_n \xrightarrow{\mathbb{P}} X$.

Limits and probabilities commute when a sequence is monotone [see note #1 on probability], but not in general. So we cannot always exchange lim and $\mathbb{P}$, and convergence i.p. is a different condition from convergence a.s. Particularly, convergence i.p. is a statement about an "*eventual probability*" of deviation, while convergence a.s. is about a probability of an "*eventual event.*" We will clarify the difference in the following example.

Example 2. Let $X_n \sim \mathsf{Bernoulli}(\frac{1}{n})$, $n \geq 1$ be independent, and let $X \sim 0$. Then X_n converges to X in probability, but not almost surely.

Proof. [If $\varepsilon > 1$, then $\mathbb{P}(|X_n - X| \geq \varepsilon) = 0$ for every $n \geq 1$.] If $0 < \varepsilon \leq 1$, then the *probability of deviation* $\mathbb{P}(|X_n - X| \geq \varepsilon) = \frac{1}{n}$ tends to 0 as $n \to \infty$, so we have convergence i.p.

To show that the X_n do not converge almost surely, we apply the second Borel-Cantelli lemma, noting that the harmonic series $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$, and the events are independent. In fact, by our remarks above, this shows that X_n *diverges* almost surely!

What is the difference between this example and the previous example of a.s. convergence? The assumption of independence. In the previous example, we could find the set of outcomes on which X_n deviates explicitly as the event $[0, \frac{1}{n}] \to \{0\}$. Here, however, we could not. The outcomes with deviation are more "randomly dispersed," and it turns out that this set of outcomes approaches an event with probability 1!

More generally, we summarize the distinction that "convergence i.p. only describes the convergence of the probability values of events, but a.s. convergence describes the convergence of the underlying events." ■

Almost sure convergence **does** imply convergence in probability. We can visualize the following proof if we draw a table with ω as rows, X_n as columns, and entries filled if $|X_n(\omega) - X(\omega)| \geq \varepsilon$. Then A_n concerns a region extending infinitely to the right, while B_n covers the leftmost column within A_n. As we increase n, the probability of a filled entry being inside A_n approaches zero by a.s. convergence, so it does the same for $B_n \subseteq A_n$.

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

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

We wish to show that $\mathbb{P}(B_n) \to 0$. We note that $A_n \supseteq B_n$, so by monotonicity, it is enough to show $\mathbb{P}(A_n) \to 0$.

- $X_m(\omega)$ converges if there is some N so that for no $m \geq N$ does $|X_m(\omega) - X(\omega)| \geq \varepsilon$. Thus we observe that $\lim_{n\to\infty} A_n$ is the event that X_n diverges. By a.s. convergence, $\mathbb{P}(\lim_{n\to\infty} A_n) = 0$.
- $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$, because if $|X_m(\omega) - X(\omega)| \geq \varepsilon$, then ω belongs to every A_n, $n \leq m$, but not necessarily $n > m$. Probability preserves decreasing limits, so $\lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}(\lim_{n\to\infty} A_n)$.

Therefore we have convergence in probability by

$$\lim_{n\to\infty} \mathbb{P}(B_n) \leq \lim_{n\to\infty} \mathbb{P}(A_n) = \mathbb{P}\Big(\lim_{n\to\infty} A_n\Big) = 0.$$

■

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:*

$$\bar{X}_n \xrightarrow{\mathbb{P}} \mu.$$

The WLLN is implied by the SLLN with the given assumptions, but the weak law actually holds in more general cases than the strong law, making the WLLN its own law and not just a corollary.

3 Convergence in distribution

Definition 3. $(X_n)_{n=1}^{\infty}$ 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 $X_n \xrightarrow{\text{d}} X$.

If the X_n are discrete, convergence i.d. is also equivalent to the convergence of the pmf, $p_{X_n} \to p_X$. If the X_n are continuous, convergence i.d. is (strictly) implied by the convergence of the pdf, $f_{X_n} \to f_X$.

You may notice that this is a weaker form of convergence. Its statement is "eventually, the values of X_n resemble values drawn from the distribution of X." However, it says nothing about the actual values drawn from $\lim_{n\to\infty} X_n$ and X, in particular about their deviations. We illustrate this idea in the following example.

Example 3. Convergence in distribution does not imply convergence in probability. Let $\Omega = [0, 1]$ with the uniform probability measure as before. Let $X_n \sim \text{Uniform}([0, 1])$, more specifically $X_{2k}(\omega) = \omega$ and $X_{2k+1}(\omega) = 1 - \omega$, and

let $X(\omega) = \omega$. Then $X_n \xrightarrow{\mathsf{d}} X$ trivially, as they share the same distribution, but the probability of deviation $\mathbb{P}(|X_n - X| \geq \varepsilon)$ oscillates between zero and a nonzero value, so it cannot tend to 0.

However, convergence in probability does imply convergence in distribution. The key idea is that the probability of deviation $\mathbb{P}(|X_n - X| \geq \varepsilon)$ tending to zero allows us approximate the event $\{X_n \leq x\}$ by $\{X \leq x\}$.

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

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

$\varepsilon > 0$ can be made arbitrarily small, so we have shown convergence i.d. ■

We will state the most common example of convergence i.d., one version of one of the most ubiquitous theorem in statistics, which is often used in confidence intervals and to justify 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,*

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

Without a careful look, the following two results of convergence i.d. may appear to contradict the Central Limit Theorem, but notice that their random variables are *not* identically distributed. We leave their proofs as discussion problems or exercises; you may find the identity $\lim_{n\to\infty}(1 - \frac{\lambda}{n})^n = e^{-\lambda}$ helpful.

Theorem 5 (Poisson limit theorem)**.** *Let $X_n \sim \mathsf{Binomial}(n, p_n)$, where $\lim_{n\to\infty} np_n$ is equal to the constant $\lambda > 0$, and let $X \sim \mathsf{Poisson}(\lambda)$. Then $X_n \xrightarrow{\mathsf{d}} X$.*

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, such as popular random graph models or balls and bins especially, also have Poisson limits.

Theorem 6 (Limit of geometric distribution)**.** *Let $X_n \sim \mathsf{Geom}(p_n)$, where $\lim_{n\to\infty} \frac{p_n}{n}$ is equal to the constant $\lambda > 0$, and let $X \sim \mathsf{Exponential}(\lambda)$. Then $X_n \xrightarrow{\mathsf{d}} X$.*

The following section is fairly interesting but entirely supplemental.

4 Convergence in expectation

Definition 4. $(X_n)_{n=1}^{\infty}$ converges **in expectation** (or 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 convergence 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 4. Consider $X_n \xrightarrow{\text{a.s.}} X$ as in example 1. However, $\mathbb{E}(X_n) = \frac{1}{n} \cdot n = 1$ is constant for every $n \geq 1$, but $\mathbb{E}(X) = 0$, so $\mathbb{E}(X_n) \nrightarrow \mathbb{E}(X)$.

This counterexample may seem troubling, because expectation is equivalent up to almost sure equivalence, yet $\lim_{n\to\infty} X_n \overset{\text{a.s.}}{=} X$ does not imply $\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)$ — we conclude that $\mathbb{E}$ and lim do not always commute. The key to the divergence in expectation is that even as the probability of eventual deviation tends to zero, the *value* of said deviation can grow unboundedly. If instead $X_n \sim 2^n \cdot \mathsf{Bernoulli}(\frac{1}{n})$, then $\mathbb{E}(X_n) = \frac{2^n}{n} \to \infty$!

If we suppose that $X_n \xrightarrow{\text{a.s.}} X$, then there turns out to be two quite strong conditions that imply convergence in expectation: if the X_n form a nondecreasing sequence, or if the X_n are bounded.

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

ii. **Dominated convergece theorem**. If there exists a nonnegative random variable $Y \geq 0$ 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.

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

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

c. Convergence in expectation implies convergence in probability.

d. Convergence in distribution implies that $\mathbb{E}(g(X_n)) \to \mathbb{E}(g(X))$ if g is bounded and continuous. [Note that the identity function is not bounded, so convergence in expectation is not included.]

e. Convergence in L^2-norm $[\mathbb{E}(|X_n - X|^2) \to 0]$ implies convergence in L^1-norm. In general, convergence in L^p-norm implies convergence in L^q-norm for all $1 \leq q \leq p < \infty$.

■