

Modes of Convergence of Random Variables

Lecture Notes

September 28, 2026

Contents

1	Setup: why there are several notions of convergence	1
2	Definitions	2
2.1	Almost sure convergence	2
2.2	Convergence in probability	3
2.3	Mean-square convergence	3
2.4	Convergence in distribution	4
2.5	Uniqueness of limits	4
3	Examples and counterexamples	4
3.1	Sequences that converge in every sense	5
3.2	Convergence in distribution but not in probability	5
3.3	In probability (and m.s.) but not almost surely	6
3.4	Almost surely but not in mean square	7
3.5	In probability but neither a.s. nor in mean square	7
3.6	Summary of the examples	7
4	Almost sure convergence implies convergence in probability	7
5	Mean-square convergence implies convergence in probability	9
6	Convergence in probability implies convergence in distribution	10
7	Summary	12

1 Setup: why there are several notions of convergence

Throughout, $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $X, X_1, X_2, \dots$ are real-valued random variables, i.e. measurable functions $\Omega \rightarrow \mathbb{R}$. A point $\omega \in \Omega$ should be thought of as one complete “run” of a random experiment; for that run we observe a *sequence of numbers*

$$X_1(\omega), X_2(\omega), X_3(\omega), \dots$$

Since random variables are functions, asking whether “ $X_n \rightarrow X$ ” is the same kind of question as asking whether a sequence of functions converges. Real analysis already offers several answers (pointwise convergence, convergence in an integral norm, $\dots$), and probability adds one more that only looks at distributions. The four modes studied here answer four different questions:

Mode	Question it answers	Looks at
Almost sure	For (almost) every run ω , does the numerical sequence $X_n(\omega)$ converge to $X(\omega)$?	whole sample paths
In probability	For a single large n , is it unlikely that X_n is far from X ?	one time n at a time
Mean square	Does the average squared error $\mathbb{E}[(X_n - X)^2]$ go to 0?	size of the error
In distribution	Do the probabilities $\mathbb{P}(X_n \leq x)$ approach $\mathbb{P}(X \leq x)$?	only the laws

The first three modes compare X_n and X as functions on the same Ω . The fourth compares only their *distributions*; in particular X_n and X need not even be defined on the same probability space.

The main results of these notes are the implications

$$X_n \xrightarrow{\text{a.s.}} X \implies X_n \xrightarrow{\mathbb{P}} X \iff X_n \xrightarrow{\text{m.s.}} X, \quad X_n \xrightarrow{\mathbb{P}} X \implies X_n \xrightarrow{d} X$$

together with examples showing that no other implication holds in general.

2 Definitions

2.1 Almost sure convergence

Definition 2.1 (Almost sure convergence). We say X_n converges to X *almost surely* (a.s., also “with probability one” or “almost everywhere”), written $X_n \xrightarrow{\text{a.s.}} X$, if

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

In words: apart from a set of runs of probability zero, every observed sequence $X_1(\omega), X_2(\omega), \dots$ converges, in the ordinary calculus sense, to the number $X(\omega)$. This is simply pointwise convergence of functions, except that we are allowed to ignore a null set of ω ’s.

Remark 2.2 (The convergence set is an event). Writing out the ε - N definition of a limit with $\varepsilon = 1/k$,

$$\{X_n \rightarrow X\} = \bigcap_{k=1}^{\infty} \bigcup_{N=1}^{\infty} \bigcap_{n \geq N} \left\{|X_n - X| < \frac{1}{k}\right\}.$$

Each set on the right is an event, and countable unions and intersections of events are events, so the probability in Definition 2.1 makes sense. Read the formula aloud: “for every tolerance $1/k$, there is a time N after which the path stays within $1/k$ of the limit.”

A very useful reformulation replaces “the path converges” with “the path leaves the ε -band only finitely often”. Recall that for events $A_1, A_2, \dots$,

$$\{A_n \text{ i.o.}\} := \bigcap_{N=1}^{\infty} \bigcup_{n \geq N} A_n = \{\omega : \omega \in A_n \text{ for infinitely many } n\}.$$

Proposition 2.3 (Equivalent forms of a.s. convergence). *The following are equivalent:*

- (a) $X_n \xrightarrow{a.s.} X$;
- (b) for every $\varepsilon > 0$, $\mathbb{P}(|X_n - X| > \varepsilon \text{ i.o.}) = 0$;
- (c) for every $\varepsilon > 0$, $\lim_{N \rightarrow \infty} \mathbb{P}\left(\sup_{n \geq N} |X_n - X| > \varepsilon\right) = 0$.

Form (c) is the one to compare with convergence in probability below: a.s. convergence controls the *worst deviation over the entire future* $\sup_{n \geq N} |X_n - X|$, while convergence in probability controls only the deviation $|X_n - X|$ at a single time n . (The equivalence (b) $\Leftrightarrow$ (c) is exactly the argument used in the proof of Theorem 4.2.)

A standard sufficient condition, which follows from (b) and the first Borel–Cantelli lemma, is:

Fact 2.4 (Borel–Cantelli criteria). *Let $A_1, A_2, \dots$ be events.*

- (i) (BC1) *If $\sum_n \mathbb{P}(A_n) < \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 0$. In particular, if $\sum_n \mathbb{P}(|X_n - X| > \varepsilon) < \infty$ for every $\varepsilon > 0$, then $X_n \xrightarrow{a.s.} X$.*
- (ii) (BC2) *If the A_n are independent and $\sum_n \mathbb{P}(A_n) = \infty$, then $\mathbb{P}(A_n \text{ i.o.}) = 1$.*

2.2 Convergence in probability

Definition 2.5 (Convergence in probability). We say X_n converges to X in probability, written $X_n \xrightarrow{\mathbb{P}} X$, if for every $\varepsilon > 0$,

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

Fix a tolerance ε . At time n , the event $\{|X_n - X| > \varepsilon\}$ is the set of runs in which X_n is “badly off”. Convergence in probability says that this set has vanishing probability. It says nothing about whether the *same* run is badly off infinitely often: the bad set may move around Ω as n changes, and every ω may visit it infinitely many times (see Example 3.6).

Remark 2.6. Using $> \varepsilon$ or $\geq \varepsilon$ in the definition makes no difference, because $\{|Y| > \varepsilon\} \subseteq \{|Y| \geq \varepsilon\} \subseteq \{|Y| > \varepsilon/2\}$.

2.3 Mean-square convergence

Definition 2.7 (Mean-square convergence). Assume $\mathbb{E}[X_n^2] < \infty$ for all n and $\mathbb{E}[X^2] < \infty$. We say X_n converges to X in mean square (m.s., or in L^2), written $X_n \xrightarrow{\text{m.s.}} X$, if

$$\lim_{n \rightarrow \infty} \mathbb{E}[(X_n - X)^2] = 0.$$

$\mathbb{E}[(X_n - X)^2]$ is the *mean-squared error* of X_n as an estimate of X ; m.s. convergence says this error goes to zero. It is the natural mode in estimation and signal processing because it is computed from second moments only, and it behaves like distance in Euclidean space: $\|Y\|_2 := \sqrt{\mathbb{E}[Y^2]}$ satisfies the triangle inequality.

Remark 2.8 (Useful decomposition). For a constant limit c ,

$$\mathbb{E}[(X_n - c)^2] = \text{Var}(X_n) + (\mathbb{E}[X_n] - c)^2 \quad (\text{variance} + \text{bias}^2).$$

Hence $X_n \xrightarrow{\text{m.s.}} c$ if and only if $\mathbb{E}[X_n] \rightarrow c$ and $\text{Var}(X_n) \rightarrow 0$.

Remark 2.9. More generally, for $p \geq 1$, $X_n \rightarrow X$ in L^p if $\mathbb{E}|X_n - X|^p \rightarrow 0$. Mean square is the case $p = 2$. All statements below about m.s. convergence hold for L^p with the obvious changes.

2.4 Convergence in distribution

Let $F_n(x) = \mathbb{P}(X_n \leq x)$ and $F(x) = \mathbb{P}(X \leq x)$ denote the cumulative distribution functions (CDFs). Recall that every CDF is nondecreasing and right-continuous, and it has at most countably many jumps.

Definition 2.10 (Convergence in distribution). We say X_n converges to X *in distribution* (also *weakly* or *in law*), written $X_n \xrightarrow{d} X$, if

$$\lim_{n \rightarrow \infty} F_n(x) = F(x) \quad \text{for every } x \text{ at which } F \text{ is continuous.}$$

This definition only involves the numbers $F_n(x)$ and $F(x)$, so it is a statement about the laws of X_n and X , not about the random variables themselves. For example, if $X_n \xrightarrow{d} X$ and Y has the same distribution as X , then also $X_n \xrightarrow{d} Y$.

Why only continuity points? Consider the deterministic sequence $X_n = 1/n$ and $X = 0$. Any reasonable definition must declare $X_n \rightarrow X$. Yet

$$F_n(0) = \mathbb{P}(1/n \leq 0) = 0 \quad \text{for all } n, \quad F(0) = \mathbb{P}(0 \leq 0) = 1,$$

so $F_n(0) \not\rightarrow F(0)$. The failure happens exactly at $x = 0$, the jump of F . Since X_n can approach a jump from the right, F_n may “miss” the value of F at the jump. Excluding discontinuity points removes this artificial problem, and since there are at most countably many of them, the limit F is still determined uniquely (a right-continuous function is determined by its values on any dense set).

Fact 2.11 (Portmanteau, partial statement). $X_n \xrightarrow{d} X$ if and only if $\mathbb{E}[g(X_n)] \rightarrow \mathbb{E}[g(X)]$ for every bounded continuous function $g : \mathbb{R} \rightarrow \mathbb{R}$.

2.5 Uniqueness of limits

- If $X_n \xrightarrow{\text{a.s.}} X$ and $X_n \xrightarrow{\text{a.s.}} Y$, then $X = Y$ a.s.
- If $X_n \xrightarrow{\mathbb{P}} X$ and $X_n \xrightarrow{\mathbb{P}} Y$, then $X = Y$ a.s., because $\mathbb{P}(|X - Y| > 2\varepsilon) \leq \mathbb{P}(|X_n - X| > \varepsilon) + \mathbb{P}(|X_n - Y| > \varepsilon) \rightarrow 0$ for every $\varepsilon > 0$.
- If $X_n \xrightarrow{d} X$ and $X_n \xrightarrow{d} Y$, then X and Y have the *same distribution*, but they need not be equal. Limits in distribution are only unique in law.

3 Examples and counterexamples

In several examples we use the “unit interval” probability space: $\Omega = [0, 1]$, $\mathcal{F}$ the Borel sets, and $\mathbb{P}$ the length (Lebesgue measure). Thus $\mathbb{P}([a, b]) = b - a$, and a random variable is just a function of $\omega \in [0, 1]$, where ω is a uniformly chosen point.

3.1 Sequences that converge in every sense

Example 3.1 (Vanishing noise). Let X be any random variable and let Z be a random variable with $\mathbb{E}[Z^2] < \infty$. Set $X_n = X + Z/n$. Then:

- *a.s.:* $|X_n(\omega) - X(\omega)| = |Z(\omega)|/n \rightarrow 0$ for every ω (since $Z(\omega)$ is a finite number).
- *m.s.:* $\mathbb{E}[(X_n - X)^2] = \mathbb{E}[Z^2]/n^2 \rightarrow 0$.
- *in probability and in distribution:* these follow from the theorems in Sections 4–6. Directly, $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|Z| > n\varepsilon) \rightarrow 0$.

Example 3.2 (Sample means: laws of large numbers). Let $Y_1, Y_2, \dots$ be i.i.d. with mean μ and variance $\sigma^2 < \infty$, and let $\bar{Y}_n = \frac{1}{n} \sum_{i=1}^n Y_i$. Then $\mathbb{E}[\bar{Y}_n] = \mu$ and $\text{Var}(\bar{Y}_n) = \sigma^2/n$, so

$$\mathbb{E}[(\bar{Y}_n - \mu)^2] = \frac{\sigma^2}{n} \rightarrow 0, \quad \text{i.e.} \quad \bar{Y}_n \xrightarrow{\text{m.s.}} \mu.$$

By Theorem 5.2 this gives $\bar{Y}_n \xrightarrow{\mathbb{P}} \mu$: the *weak law of large numbers*. The (harder) *strong law of large numbers* states that also $\bar{Y}_n \xrightarrow{\text{a.s.}} \mu$ (it even requires only $\mathbb{E}|Y_1| < \infty$). The strong law is the mathematical statement that the running average of a single, infinitely long experiment settles down to μ .

3.2 Convergence in distribution but not in probability

Example 3.3 (Same law, unrelated values). Let $X \sim \text{Bernoulli}(1/2)$ and define $X_n = 1 - X$ for every n . Since $1 - X$ is also Bernoulli(1/2), every X_n has exactly the same distribution as X , so trivially $F_n = F$ and $X_n \xrightarrow{d} X$. But $|X_n - X| = |1 - 2X| = 1$ always, so

$$\mathbb{P}(|X_n - X| > \tfrac{1}{2}) = 1 \quad \text{for all } n,$$

and X_n does not converge to X in probability (hence not a.s. or in m.s. either). The same works with $X \sim N(0, 1)$ and $X_n = -X$: $\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X| > \varepsilon/2)$, which does not depend on n and is not 0.

The lesson is that convergence in distribution says nothing about the values of X_n and X on the same ω ; it only compares histograms.

Example 3.4 (Central limit theorem). With Y_i as in Example 3.2 and $\sigma > 0$, let $Z_n = \sqrt{n}(\bar{Y}_n - \mu)/\sigma$. The central limit theorem says $Z_n \xrightarrow{d} Z \sim N(0, 1)$. This convergence cannot be improved to convergence in probability. Indeed, write $Z_{2n} = \frac{1}{\sqrt{2}}(Z_n + Z'_n)$, where Z'_n is the standardized sum of $Y_{n+1}, \dots, Y_{2n}$, which is independent of Z_n . Then

$$Z_{2n} - Z_n = \left(\frac{1}{\sqrt{2}} - 1\right)Z_n + \frac{1}{\sqrt{2}}Z'_n \xrightarrow{d} N(0, 2 - \sqrt{2}),$$

a nondegenerate limit. If Z_n converged in probability to some Z , then $|Z_{2n} - Z_n| \leq |Z_{2n} - Z| + |Z - Z_n|$ would force $Z_{2n} - Z_n \xrightarrow{\mathbb{P}} 0$, hence $Z_{2n} - Z_n \xrightarrow{d} 0$, a contradiction. Intuitively, Z_n keeps being refreshed by new independent randomness, so its distribution stabilizes while its values never settle down.

Example 3.5 (Examples of convergence in distribution). (a) *Discrete uniform $\rightarrow$ continuous uniform.* Let X_n be uniform on $\{1/n, 2/n, \dots, n/n\}$. For $x \in [0, 1]$, $F_n(x) = \lfloor nx \rfloor / n$, and since $nx - 1 < \lfloor nx \rfloor \leq nx$ we get $F_n(x) \rightarrow x$. Also $F_n(x) = 0$ for $x < 0$ and $F_n(x) = 1$ for $x \geq 1$. Hence $X_n \xrightarrow{d} U \sim \text{Uniform}[0, 1]$, although each X_n is discrete and U is continuous.

(b) *Maximum of uniforms.* Let $U_1, U_2, \dots$ be i.i.d. $\text{Uniform}[0, 1]$, $M_n = \max_{i \leq n} U_i$ and $W_n = n(1 - M_n)$. For $x \in [0, n]$,

$$\mathbb{P}(W_n > x) = \mathbb{P}\left(M_n < 1 - \frac{x}{n}\right) = \left(1 - \frac{x}{n}\right)^n \rightarrow e^{-x},$$

so $W_n \xrightarrow{d} W \sim \text{Exp}(1)$.

(c) *Shrinking Gaussians.* If $X_n \sim N(0, 1/n)$, then $F_n(x) = \Phi(\sqrt{n}x) \rightarrow 0$ for $x < 0$ and $\rightarrow 1$ for $x > 0$, which is the CDF of the constant 0 at all its continuity points. So $X_n \xrightarrow{d} 0$, even though $F_n(0) = \frac{1}{2}$ for every n . (In fact $X_n \xrightarrow{\text{m.s.}} 0$ as well, if all X_n are defined on one space, since $\mathbb{E}[X_n^2] = 1/n$.)

3.3 In probability (and m.s.) but not almost surely

Example 3.6 (The “typewriter” sequence). On the unit interval, sweep an indicator of shrinking width across $[0, 1]$ again and again. Precisely, write each $n \geq 1$ uniquely as $n = 2^k + j$ with $k \geq 0$ and $0 \leq j < 2^k$, and set

$$X_n = \mathbf{1}_{I_n}, \quad I_n = \left[\frac{j}{2^k}, \frac{j+1}{2^k}\right].$$

So $X_1 = \mathbf{1}_{[0,1]}$, then X_2, X_3 are the indicators of the two halves, $X_4, \dots, X_7$ of the four quarters, and so on.

- *In probability:* for $0 < \varepsilon < 1$, $\mathbb{P}(|X_n| > \varepsilon) = \mathbb{P}(I_n) = 2^{-k} \rightarrow 0$, since $k \rightarrow \infty$ as $n \rightarrow \infty$. So $X_n \xrightarrow{\mathbb{P}} 0$.
- *Mean square:* $\mathbb{E}[X_n^2] = \mathbb{P}(I_n) = 2^{-k} \rightarrow 0$, so $X_n \xrightarrow{\text{m.s.}} 0$.
- *Not a.s.:* fix any $\omega \in [0, 1]$. At each “level” k the intervals I_n cover $[0, 1]$, so $X_n(\omega) = 1$ for at least one n in each level; for $k \geq 1$ some other I_n in that level misses ω , so $X_n(\omega) = 0$ for at least one n in each level. Therefore the sequence $X_n(\omega)$ takes both values 0 and 1 infinitely often and does not converge, for *every* ω . So $\mathbb{P}(X_n \rightarrow 0) = 0$.

The bad set $\{X_n = 1\}$ has shrinking probability, which is all that convergence in probability asks for. But it keeps moving, and every ω is swept over once per pass.

Example 3.7 (Independent version). Let $X_1, X_2, \dots$ be independent with $X_n \sim \text{Bernoulli}(1/n)$. Then $\mathbb{P}(|X_n| > \varepsilon) = 1/n \rightarrow 0$ and $\mathbb{E}[X_n^2] = 1/n \rightarrow 0$, so $X_n \xrightarrow{\mathbb{P}} 0$ and $X_n \xrightarrow{\text{m.s.}} 0$. But the events $A_n = \{X_n = 1\}$ are independent with $\sum_n \mathbb{P}(A_n) = \sum_n 1/n = \infty$, so by BC2 (Fact 2.4), $\mathbb{P}(X_n = 1 \text{ i.o.}) = 1$. Almost every path contains infinitely many 1’s, so $X_n \not\xrightarrow{\text{a.s.}} 0$. If instead $X_n \sim \text{Bernoulli}(1/n^2)$, then $\sum 1/n^2 < \infty$ and BC1 gives $X_n \xrightarrow{\text{a.s.}} 0$. The borderline between the two behaviours is exactly the summability of $\mathbb{P}(A_n)$.

3.4 Almost surely but not in mean square

Example 3.8 (A tall thin spike). On the unit interval let $X_n = n \mathbf{1}_{(0,1/n)}$.

- *a.s. (in fact everywhere):* $X_n(0) = 0$ for all n ; and for $\omega > 0$, $X_n(\omega) = 0$ as soon as $1/n \leq \omega$. So $X_n(\omega) \rightarrow 0$ for every ω , and hence also $X_n \xrightarrow{\mathbb{P}} 0$.
- *Not m.s.:* $\mathbb{E}[X_n^2] = n^2 \cdot \mathbb{P}((0, 1/n)) = n^2 \cdot \frac{1}{n} = n \rightarrow \infty$. Even $\mathbb{E}|X_n| = 1$ for all n , so X_n does not converge to 0 in L^1 either.

The spike is confined to a set of vanishing probability (good for a.s. and in-probability convergence), but its height grows fast enough that its contribution to the mean-square error blows up.

3.5 In probability but neither a.s. nor in mean square

Example 3.9. Let X_n be independent with $\mathbb{P}(X_n = n) = 1/n$ and $\mathbb{P}(X_n = 0) = 1 - 1/n$.

- *In probability:* $\mathbb{P}(|X_n| > \varepsilon) = 1/n \rightarrow 0$.
- *Not m.s.:* $\mathbb{E}[X_n^2] = n^2 \cdot \frac{1}{n} = n \rightarrow \infty$.
- *Not a.s.:* $\sum \mathbb{P}(X_n = n) = \infty$ and the events are independent, so by BC2 $X_n = n$ infinitely often with probability one.

Replacing $1/n$ by $1/n^2$ gives $\mathbb{P}(X_n = n) = 1/n^2$: now $X_n \xrightarrow{\text{a.s.}} 0$ by BC1, but $\mathbb{E}[X_n^2] = 1 \not\rightarrow 0$. This is another sequence that converges a.s. but not in mean square.

3.6 Summary of the examples

Example	a.s.	m.s.	in prob.	in dist.
3.1 $X + Z/n$	yes	yes	yes	yes
3.2 sample mean $\rightarrow \mu$	yes (SLLN)	yes	yes	yes
3.3 $X_n = 1 - X$	no	no	no	yes
3.4 CLT, $Z_n \rightarrow N(0, 1)$	no	no	no	yes
3.6 typewriter	no	yes	yes	yes
3.7 indep. Bernoulli($1/n$)	no	yes	yes	yes
3.8 spike $n\mathbf{1}_{(0,1/n)}$	yes	no	yes	yes
3.9 $\mathbb{P}(X_n = n) = 1/n$	no	no	yes	yes
3.9 $\mathbb{P}(X_n = n) = 1/n^2$	yes	no	yes	yes

Notice that no row has “yes” in the a.s. or m.s. column and “no” in the probability column, and no row has “yes” in the probability column and “no” in the distribution column. That is exactly what the three theorems below guarantee.

4 Almost sure convergence implies convergence in probability

Intuition. If almost every path eventually enters the ε -band around X and stays there forever, then at a large time n the only runs that can be outside the band are those whose paths have *not yet*

settled. The set of runs that have not settled by time n shrinks as n grows, and in the limit it consists only of runs that never settle, which have probability zero. Hence the probability of being outside the band at time n goes to zero.

The proof makes this precise by comparing the “bad at time n ” event A_n with the larger “bad at some time $\geq n$ ” event B_n .

We need one basic property of probability measures.

Lemma 4.1 (Continuity from above). *If $B_1 \supseteq B_2 \supseteq \dots$ are events, then $\lim_{n \rightarrow \infty} \mathbb{P}(B_n) = \mathbb{P}\left(\bigcap_n B_n\right)$.*

Proof. Let $B = \bigcap_n B_n$ and write $B_1 \setminus B$ as the disjoint union of the “layers” $B_m \setminus B_{m+1}$, $m \geq 1$. By countable additivity,

$$\mathbb{P}(B_1) - \mathbb{P}(B) = \sum_{m=1}^{\infty} (\mathbb{P}(B_m) - \mathbb{P}(B_{m+1})) = \lim_{n \rightarrow \infty} (\mathbb{P}(B_1) - \mathbb{P}(B_n)),$$

the last sum being telescoping. Since $\mathbb{P}(B_1) \leq 1 < \infty$ we may cancel it, which gives $\mathbb{P}(B_n) \rightarrow \mathbb{P}(B)$. $\square$

Theorem 4.2. *If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{\mathbb{P}} X$.*

Proof. Fix $\varepsilon > 0$. Define the events

$$\begin{aligned} A_n &= \{|X_n - X| > \varepsilon\} && \text{(bad at time } n), \\ B_n &= \bigcup_{m \geq n} A_m = \left\{ \sup_{m \geq n} |X_m - X| > \varepsilon \right\} && \text{(bad at some time } m \geq n). \end{aligned}$$

Step 1: $A_n \subseteq B_n$. This is immediate, since A_n is one of the sets in the union defining B_n . Hence

$$\mathbb{P}(A_n) \leq \mathbb{P}(B_n) \quad \text{for every } n. \quad (1)$$

Step 2: the B_n decrease. B_{n+1} is a union over fewer sets than B_n , so $B_1 \supseteq B_2 \supseteq \dots$. By Lemma 4.1,

$$\lim_{n \rightarrow \infty} \mathbb{P}(B_n) = \mathbb{P}(B), \quad B := \bigcap_{n \geq 1} B_n = \{X_n \text{ i.o.}\}. \quad (2)$$

Step 3: B is contained in the non-convergence set. Let $\omega \in B$. Then $\omega \in B_n$ for every n , so for every n there is some $m \geq n$ with $|X_m(\omega) - X(\omega)| > \varepsilon$. That is, the sequence $X_m(\omega)$ is more than ε away from $X(\omega)$ for infinitely many m , and so $X_m(\omega) \not\rightarrow X(\omega)$. We have shown

$$B \subseteq \{X_n \not\rightarrow X\}, \quad \text{hence} \quad \mathbb{P}(B) \leq \mathbb{P}(X_n \not\rightarrow X) = 1 - \mathbb{P}(X_n \rightarrow X) = 0.$$

Step 4: conclude. Combining (1), (2) and Step 3,

$$0 \leq \limsup_{n \rightarrow \infty} \mathbb{P}(|X_n - X| > \varepsilon) \leq \lim_{n \rightarrow \infty} \mathbb{P}(B_n) = \mathbb{P}(B) = 0.$$

Since $\varepsilon > 0$ was arbitrary, $X_n \xrightarrow{\mathbb{P}} X$. $\square$

Remark 4.3 (Where the converse breaks). The proof shows that a.s. convergence is equivalent to $\mathbb{P}(B_n) \rightarrow 0$ (“the whole future is good”), while convergence in probability is $\mathbb{P}(A_n) \rightarrow 0$ (“the present is good”). The inequality $\mathbb{P}(A_n) \leq \mathbb{P}(B_n)$ can be very far from equality. In the typewriter sequence (Example 3.6) with $\varepsilon = \frac{1}{2}$, $\mathbb{P}(A_n) = 2^{-k} \rightarrow 0$, but $B_n = [0, 1]$ for every n because the remaining passes cover the whole interval, so $\mathbb{P}(B_n) = 1$.

Fact 4.4 (Partial converse). *If $X_n \xrightarrow{\mathbb{P}} X$, then there is a subsequence $n_1 < n_2 < \dots$ with $X_{n_k} \xrightarrow{a.s.} X$. (Choose n_k with $\mathbb{P}(|X_{n_k} - X| > 2^{-k}) \leq 2^{-k}$ and apply BC1.) In the typewriter example, the subsequence $X_{2^k} = \mathbf{1}_{[0, 2^{-k}]}$ converges to 0 for every $\omega > 0$.*

5 Mean-square convergence implies convergence in probability

Intuition. The mean-square error $\mathbb{E}[(X_n - X)^2]$ is an average of the squared error over all runs. Suppose the error exceeds ε on a set of runs of probability p . On that set the squared error is at least ε^2 , so that set *alone* contributes at least $\varepsilon^2 p$ to the average. If the average tends to zero, then $\varepsilon^2 p$ must tend to zero, and so must p . In short: a small average cannot hide a large error on a set of substantial probability. This is exactly Markov's (or Chebyshev's) inequality.

Lemma 5.1 (Markov's inequality). *If $Y \geq 0$ and $a > 0$, then $\mathbb{P}(Y \geq a) \leq \frac{\mathbb{E}[Y]}{a}$.*

Proof. Compare two random variables pointwise. For every ω ,

$$a \mathbf{1}_{\{Y \geq a\}}(\omega) \leq Y(\omega).$$

Indeed, if $Y(\omega) \geq a$ the left side equals $a \leq Y(\omega)$, and otherwise the left side is $0 \leq Y(\omega)$. Taking expectations, which preserves inequalities, gives $a \mathbb{P}(Y \geq a) \leq \mathbb{E}[Y]$. $\square$

Theorem 5.2. *If $X_n \xrightarrow{m.s.} X$, then $X_n \xrightarrow{\mathbb{P}} X$.*

Proof. Fix $\varepsilon > 0$. Since $t \mapsto t^2$ is increasing on $[0, \infty)$, we have the equality of events

$$\{|X_n - X| \geq \varepsilon\} = \{(X_n - X)^2 \geq \varepsilon^2\}.$$

Apply Markov's inequality with $Y = (X_n - X)^2 \geq 0$ and $a = \varepsilon^2$:

$$\mathbb{P}(|X_n - X| > \varepsilon) \leq \mathbb{P}(|X_n - X| \geq \varepsilon) = \mathbb{P}((X_n - X)^2 \geq \varepsilon^2) \leq \frac{\mathbb{E}[(X_n - X)^2]}{\varepsilon^2}.$$

The right-hand side tends to 0 as $n \rightarrow \infty$ by assumption, with ε fixed. Hence $\mathbb{P}(|X_n - X| > \varepsilon) \rightarrow 0$ for every $\varepsilon > 0$, i.e. $X_n \xrightarrow{\mathbb{P}} X$. $\square$

Remark 5.3. (a) The proof gives more than convergence: it gives a *rate*. For instance, in Example 3.2, $\mathbb{P}(|\bar{Y}_n - \mu| > \varepsilon) \leq \sigma^2/(n\varepsilon^2)$ (Chebyshev's inequality), which is the classical proof of the weak law of large numbers.

(b) The same proof with $Y = |X_n - X|^p$ and $a = \varepsilon^p$ shows that convergence in L^p implies convergence in probability, for any $p > 0$.

(c) *Where the converse breaks.* Markov's inequality bounds the probability of a large error by the average error, but not conversely: an error that is rare but huge makes the probability small and the average large. Examples 3.8 and 3.9 are built exactly this way.

(d) *Partial converse.* If $X_n \xrightarrow{\mathbb{P}} X$ and $|X_n| \leq Y$ for all n with $\mathbb{E}[Y^2] < \infty$ (a uniform bound that rules out rare huge values), then $X_n \xrightarrow{m.s.} X$. This is a form of the dominated convergence theorem.

6 Convergence in probability implies convergence in distribution

Intuition. If $X_n \xrightarrow{\mathbb{P}} X$, then X_n is a slightly perturbed copy of X : apart from a set of small probability, X_n differs from X by at most ε . Moving every point of a distribution by at most ε can only shift the CDF sideways by at most ε . So, up to a small error probability, the CDF F_n is trapped between the two shifted copies $F(\cdot - \varepsilon)$ and $F(\cdot + \varepsilon)$ of F (see Figure 1). Letting first $n \rightarrow \infty$ and then $\varepsilon \rightarrow 0$ squeezes $F_n(x)$ onto $F(x)$, provided F has no jump at x .

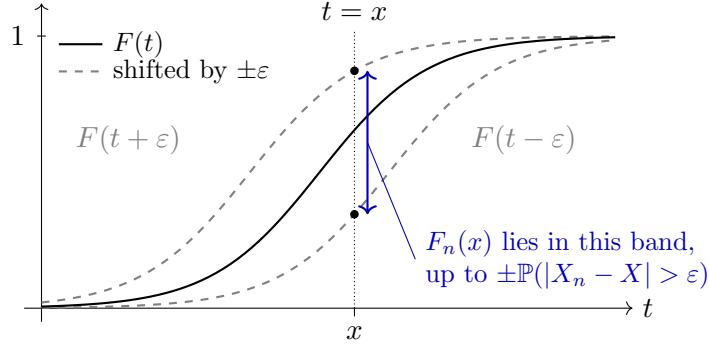


Figure 1: The sandwich in the proof of Theorem 6.2: for large n , $F(x - \varepsilon) - \delta_n \leq F_n(x) \leq F(x + \varepsilon) + \delta_n$ with $\delta_n = \mathbb{P}(|X_n - X| > \varepsilon) \rightarrow 0$. As $\varepsilon \downarrow 0$ the two dashed curves close in on F at every point where F is continuous.

The entire proof rests on one elementary inclusion of events.

Lemma 6.1 (Comparison of CDFs). *For any random variables Y, Z on the same space, any $t \in \mathbb{R}$ and any $\varepsilon > 0$,*

$$\mathbb{P}(Y \leq t) \leq \mathbb{P}(Z \leq t + \varepsilon) + \mathbb{P}(|Y - Z| > \varepsilon).$$

Proof. We show $\{Y \leq t\} \subseteq \{Z \leq t + \varepsilon\} \cup \{|Y - Z| > \varepsilon\}$ and then use monotonicity and subadditivity of $\mathbb{P}$. Take ω with $Y(\omega) \leq t$. Either $|Y(\omega) - Z(\omega)| > \varepsilon$, and then ω is in the second set; or $|Y(\omega) - Z(\omega)| \leq \varepsilon$, and then

$$Z(\omega) \leq Y(\omega) + \varepsilon \leq t + \varepsilon,$$

so ω is in the first set. In words: if Y is at most t and Z is within ε of Y , then Z is at most $t + \varepsilon$. $\square$

Theorem 6.2. *If $X_n \xrightarrow{\mathbb{P}} X$, then $X_n \xrightarrow{d} X$.*

Proof. Let F_n and F be the CDFs of X_n and X , let x be a continuity point of F , and fix $\varepsilon > 0$. Write $\delta_n = \mathbb{P}(|X_n - X| > \varepsilon)$; by assumption $\delta_n \rightarrow 0$.

Upper bound. Apply Lemma 6.1 with $Y = X_n$, $Z = X$, $t = x$:

$$F_n(x) = \mathbb{P}(X_n \leq x) \leq \mathbb{P}(X \leq x + \varepsilon) + \delta_n = F(x + \varepsilon) + \delta_n. \quad (3)$$

Lower bound. Apply Lemma 6.1 with the roles reversed, $Y = X$, $Z = X_n$, and $t = x - \varepsilon$:

$$F(x - \varepsilon) = \mathbb{P}(X \leq x - \varepsilon) \leq \mathbb{P}(X_n \leq x) + \delta_n = F_n(x) + \delta_n,$$

that is,

$$F_n(x) \geq F(x - \varepsilon) - \delta_n. \quad (4)$$

Let $n \rightarrow \infty$. From (3) and (4), using $\delta_n \rightarrow 0$,

$$F(x - \varepsilon) \leq \liminf_{n \rightarrow \infty} F_n(x) \leq \limsup_{n \rightarrow \infty} F_n(x) \leq F(x + \varepsilon). \quad (5)$$

(We cannot yet write $\lim F_n(x)$, because we do not know it exists; the $\liminf$ and $\limsup$ always exist.)

Let $\varepsilon \downarrow 0$. The middle terms of (5) do not depend on ε . As $\varepsilon \downarrow 0$:

- $F(x + \varepsilon) \rightarrow F(x)$, because every CDF is right-continuous;
- $F(x - \varepsilon) \rightarrow F(x^-)$, the left limit, which equals $F(x)$ because x is a continuity point of F .

Therefore $F(x) \leq \liminf_n F_n(x) \leq \limsup_n F_n(x) \leq F(x)$. The $\liminf$ and $\limsup$ are equal, so the limit exists and $\lim_{n \rightarrow \infty} F_n(x) = F(x)$. As x was an arbitrary continuity point of F , $X_n \xrightarrow{d} X$. $\square$

Remark 6.3 (Role of the continuity point). At a jump of F the argument only yields $F(x^-) \leq \liminf F_n(x) \leq \limsup F_n(x) \leq F(x)$, which does not pin the limit down. This is not a weakness of the proof: $X_n = 1/n \xrightarrow{\mathbb{P}} 0$, but $F_n(0) = 0 \not\rightarrow 1 = F(0)$.

Remark 6.4 (Where the converse breaks). Convergence in distribution does not even require X_n and X to be defined on the same space, so in general the event $\{|X_n - X| > \varepsilon\}$ may be meaningless. Even when it makes sense, the values can be unrelated (Examples 3.3 and 3.4).

Proposition 6.5 (Partial converse: constant limits). *If $X_n \xrightarrow{d} c$ for a constant c , then $X_n \xrightarrow{\mathbb{P}} c$.*

Proof. The CDF of the constant c is $F(x) = \mathbf{1}\{x \geq c\}$, which is continuous everywhere except at c . Fix $\varepsilon > 0$; then $c - \varepsilon$ and $c + \varepsilon$ are continuity points, with $F(c - \varepsilon) = 0$ and $F(c + \varepsilon) = 1$. Hence

$$\mathbb{P}(|X_n - c| > \varepsilon) = \mathbb{P}(X_n < c - \varepsilon) + \mathbb{P}(X_n > c + \varepsilon) \leq F_n(c - \varepsilon) + (1 - F_n(c + \varepsilon)) \rightarrow 0 + (1 - 1) = 0.$$

$\square$

When the limit is a constant, there is no “randomness of the limit” whose values could disagree with X_n , so knowing the distribution is enough. Example 3.5(c) is an instance.

7 Summary

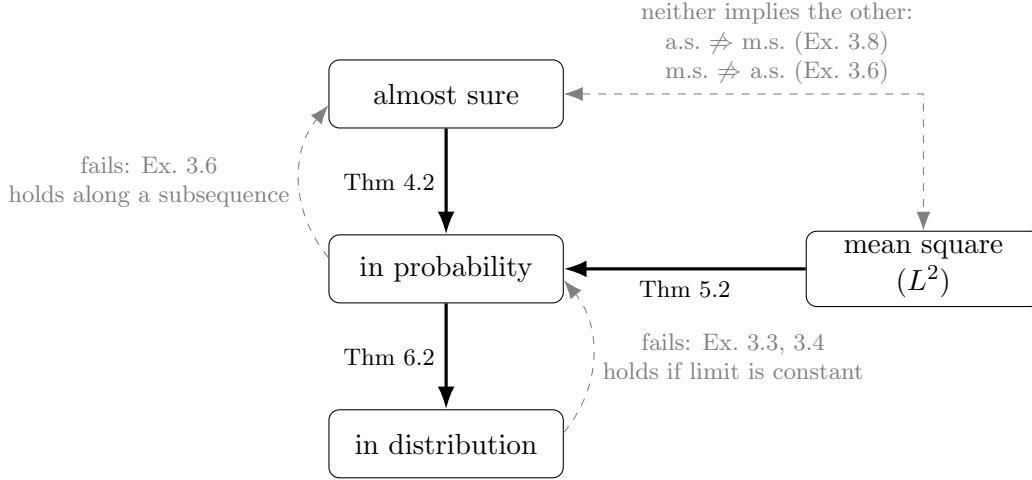


Figure 2: Solid arrows: implications proved in these notes. Dashed arrows: implications that fail in general, with a counterexample and, where available, a partial converse.

- **Proved implications.** $X_n \xrightarrow{\text{a.s.}} X \Rightarrow X_n \xrightarrow{\mathbb{P}} X$ (Theorem 4.2); $X_n \xrightarrow{\text{m.s.}} X \Rightarrow X_n \xrightarrow{\mathbb{P}} X$ (Theorem 5.2); $X_n \xrightarrow{\mathbb{P}} X \Rightarrow X_n \xrightarrow{d} X$ (Theorem 6.2).
- **Failed converses.** In probability $\not\Rightarrow$ a.s. (typewriter, Ex. 3.6); in probability $\not\Rightarrow$ m.s. (Ex. 3.9); in distribution $\not\Rightarrow$ in probability (Ex. 3.3).
- **a.s. and m.s. are not comparable.** a.s. $\not\Rightarrow$ m.s. (spike, Ex. 3.8); m.s. $\not\Rightarrow$ a.s. (typewriter, Ex. 3.6).
- **Partial converses.** In probability $\Rightarrow$ a.s. along a subsequence; in probability plus a square-integrable bound $|X_n| \leq Y \Rightarrow$ m.s.; in distribution to a *constant* $\Rightarrow$ in probability (Proposition 6.5).
- **The key ideas in one line each.** a.s. $\Rightarrow$ prob.: “bad now” $\subseteq$ “bad sometime from now on”, and the latter shrinks to a null set. m.s. $\Rightarrow$ prob.: Markov, a small average cannot hide a large error on a large set. prob. $\Rightarrow$ dist.: a small perturbation shifts the CDF sideways by at most ε , up to a small probability.