1 Lecture 9 — Convergence in Probability and Measure

We now study what it means for a sequence of probability measures, or of random variables, to converge. There are several inequivalent notions; this lecture introduces the four standard modes for random variables (almost sure, in probability, in L^p, in distribution), together with their parent notion at the level of measures (*weak* convergence). The Portmanteau theorem packages the equivalent characterisations of weak convergence, and a small Hasse diagram records the implications between the modes.

1.1 Weak convergence of probability measures

Let $(\Omega, \mathcal{F})$ be a measurable space and let $\{\mathbb{P}_i\}_{i=1}^{\infty}$ be a sequence of probability measures on $(\Omega, \mathcal{F})$. What should "$\mathbb{P}_i \to \mathbb{P}$" mean? The naive choice "$\mathbb{P}_i(A) \to \mathbb{P}(A)$ for every $A \in \mathcal{F}$" (*setwise* convergence) is too strong to be useful in practice; the standard notion fixes a topology on Ω and tests against continuous functions.

Definition 1.1: Weak convergence of measures

Let S be a metric space with Borel σ-field $\mathcal{S} = \mathcal{B}(S)$, and let $\mathbb{P}, \mathbb{P}_1, \mathbb{P}_2, \dots$ be probability measures on $(S, \mathcal{S})$. We say $\mathbb{P}_i$ *converges weakly* to $\mathbb{P}$, written $\mathbb{P}_i \Rightarrow \mathbb{P}$, if

$$\int_S f\, d\mathbb{P}_i \longrightarrow \int_S f\, d\mathbb{P} \qquad \text{for every } f \in C_B(S),$$

where $C_B(S)$ denotes the bounded continuous real-valued functions on S.

Remark 1.1. Weak convergence is the topology of *closeness* between measures generated by the metric on S. Concretely, an ε-neighbourhood of $\mathbb{P}$ is determined by a finite collection $f_1, \dots, f_n \in C_B(S)$: the neighbourhood is the set of all probability measures Q with $\left|\int f_i\, d\mathbb{P} - \int f_i\, dQ\right| < \varepsilon$ for each i. The weakest test against $f \in C_B(S)$ gives the weakest of several useful metrics on the space of probability measures.

The next theorem lists the equivalent characterisations of weak convergence; it is the standard packaging that one cites whenever weak convergence appears in the wild.

Theorem 1.2: Portmanteau theorem

Let $\mathbb{P}$ and $\{\mathbb{P}_i\}_{i=1}^{\infty}$ be probability measures on a metric space $(S, \mathcal{S})$. The following are equivalent.

1. $\mathbb{P}_i \Rightarrow \mathbb{P}$, i.e. $\int f\, d\mathbb{P}_i \to \int f\, d\mathbb{P}$ for every $f \in C_B(S)$.
2. $\int f\, d\mathbb{P}_i \to \int f\, d\mathbb{P}$ for every bounded *uniformly* continuous f.
3. $\limsup_i \mathbb{P}_i(C) \le \mathbb{P}(C)$ for every closed $C \subseteq S$.
4. $\liminf_i \mathbb{P}_i(U) \ge \mathbb{P}(U)$ for every open $U \subseteq S$.
5. $\lim_i \mathbb{P}_i(A) = \mathbb{P}(A)$ for every $A \in \mathcal{S}$ with $\mathbb{P}(\partial A) = 0$, where $\partial A = \bar{A} \cap \overline{A^c}$ is the topological boundary.

Remark 1.2. The implications $(1) \Rightarrow (2)$ and (2)⇒(3) are the workhorse direction. For (2)⇒(3) the trick is to take $C_\delta = \{x \in S : d(x, C) < \delta\}$ and choose a uniformly continuous f

with $f = 1$ on C and $f = 0$ off C_δ (Urysohn's lemma); since $C_\delta \downarrow C$ as $\delta \to 0^+$, one has $\mathbb{P}_i(C) \le \int f\,d\mathbb{P}_i \to \int f\,d\mathbb{P} \le \mathbb{P}(C_\delta) < \mathbb{P}(C) + \varepsilon$, and $\varepsilon \downarrow 0$ finishes the argument.

Remark 1.3. Changing the test class for f tightens the convergence:

- $f \in C_B(S)$ is weak convergence;
- $\sup_f \left|\int f\,d\mathbb{P}_i - \int f\,d\mathbb{P}\right| \to 0$ over all continuous $f\colon S \to [-1,1]$ is the *Radon* metric;
- the same supremum over all *measurable* $f\colon S \to [-1,1]$ gives the *total variation* distance;
- the supremum restricted to 1-Lipschitz $f\colon S \to [-1,1]$ is the 1-Wasserstein distance, central to optimal transport.

1.2 Random variables and their distributions

We now lift the picture from measures to random variables. Fix a probability space $(\Omega, \mathcal{F}, \mu)$ and a metric space $(S, \mathcal{S})$ (with $\mathcal{S} = \mathcal{B}(S)$). A random variable is a measurable map $X\colon \Omega \to S$; its *distribution* is the pushforward

$$\mathbb{P}(A) \;=\; \mu(X^{-1}(A)), \qquad A \in \mathcal{S}.$$

With this in place, expectations have the change-of-variables form

$$\mathbb{E}[X] \;=\; \int_\Omega X(\omega)\,d\mu(\omega) \;=\; \int_S x\,d\mathbb{P}(x).$$

For a sequence $\{X_i\}_{i=1}^\infty$ we write $\mathbb{P}_i$ for the distribution of X_i, and when no confusion arises we abuse notation and write $\mathbb{P}_i(A)$ for $\mathbb{P}(X_i \in A)$.

1.3 Modes of convergence

We collect the four standard modes; throughout, $X, X_1, X_2, \ldots$ are random variables on a common probability space taking values in a metric space (S, d).

Definition 1.3: Convergence in distribution

X_i converges to X *in distribution*, written $X_i \xrightarrow{d} X$, if the laws $\mathbb{P}_i$ of X_i converge weakly to the law $\mathbb{P}$ of X: $\mathbb{P}_i \Rightarrow \mathbb{P}$ (Result 1.1).

Definition 1.4: Convergence in probability

X_i converges to X *in probability*, written $X_i \xrightarrow{\mathbb{P}} X$, if for every $\varepsilon > 0$,

$$\mu(\{\omega \in \Omega : d(X_i(\omega), X(\omega)) > \varepsilon\}) \;\longrightarrow\; 0 \qquad \text{as } i \to \infty.$$

In shorthand, $\mathbb{P}(d(X_i, X) > \varepsilon) \to 0$ for every $\varepsilon > 0$.

Definition 1.5: Almost sure convergence

X_i converges to X *almost surely* (or μ-*almost everywhere*), written $X_i \xrightarrow{\text{a.s.}} X$, if

$$\mu(\{\omega \in \Omega : X_i(\omega) \to X(\omega)\}) \; = \; 1,$$

or equivalently $\mu(\{\omega : X_i(\omega) \not\to X(\omega)\}) = 0$. This is pointwise convergence except on a μ-null set; the metric d does not enter the statement.

Definition 1.6: Convergence in L^p

For $p \in [1, \infty)$, X_i converges to X *in* L^p, written $X_i \xrightarrow{L^p} X$, if

$$\mathbb{E}[d(X_i, X)^p] \; = \; \int_\Omega d(X_i(\omega), X(\omega))^p \, d\mu(\omega) \; \longrightarrow \; 0.$$

When $S = \mathbb{R}$ this reduces to $\int |X_i - X|^p \, d\mu \to 0$.

1.4 Hierarchy of convergence

The four modes are not equivalent; they are linked by a small lattice of implications.

Proposition 1.7: Hierarchy of modes

For random variables on a probability space $(\Omega, \mathcal{F}, \mu)$:

1. $X_i \xrightarrow{\text{a.s.}} X \implies X_i \xrightarrow{\mathbb{P}} X$.
2. $X_i \xrightarrow{\mathbb{P}} X \implies X_i \xrightarrow{d} X$.
3. For any $p \in [1, \infty]$, $X_i \xrightarrow{L^p} X \implies X_i \xrightarrow{\mathbb{P}} X$.
4. For $1 \le q < p \le \infty$, $X_i \xrightarrow{L^p} X \implies X_i \xrightarrow{L^q} X$ (on a probability space; this uses Jensen).

None of the converses hold without extra hypotheses; in particular, a.s. convergence and L^p convergence are incomparable.

Remark 1.4. The implication $\xrightarrow{L^p} \Rightarrow \xrightarrow{\mathbb{P}}$ is a one-line consequence of **??**: for any $\varepsilon > 0$,

$$\mu(d(X_i, X) > \varepsilon) \; \le \; \frac{\mathbb{E}[d(X_i, X)^p]}{\varepsilon^p} \; \longrightarrow \; 0.$$

The L^p-monotonicity uses Jensen applied to the convex map $t \mapsto t^{p/q}$ on a probability space: $\mathbb{E}|Y|^q \le (\mathbb{E}|Y|^p)^{q/p}$. The direction $\xrightarrow{\mathbb{P}} \Rightarrow \xrightarrow{d}$ is a corollary of the Portmanteau theorem (Result 1.2); the direction $\xrightarrow{\text{a.s.}} \Rightarrow \xrightarrow{\mathbb{P}}$ is dominated convergence applied to the indicator $\mathbf{1}_{\{d(X_i, X) > \varepsilon\}}$.

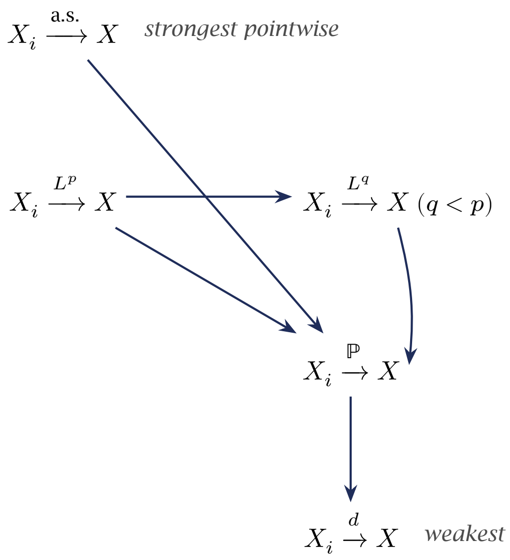

Figure 1. Hasse diagram of the four modes of convergence on a probability space. Arrows point from stronger to weaker; the horizontal arrow is the L^p-monotonicity from Jensen's inequality. There is no arrow between $\xrightarrow{\text{a.s.}}$ and $\xrightarrow{L^p}$ without extra integrability.

> ■ **Example 1.1 (Why a.s. and L^p are incomparable).** On $([0,1], \mathcal{B}, \lambda)$, set $X_n = n\,\mathbf{1}_{(0,1/n)}$. Then $X_n \to 0$ pointwise (so $X_n \xrightarrow{\text{a.s.}} 0$) but $\mathbb{E}[X_n] = 1$ for every n, so X_n does not converge to 0 in L^1. Conversely, the "typewriter" sequence of indicators of dyadic sub-intervals of $[0,1]$ satisfies $X_n \xrightarrow{L^p} 0$ for every p yet fails to converge at any single ω.