STA 711: Probability & Measure Theory
Robert L. Wolpert

7 Convergence in $\mathbb{R}^d$ and in Metric Spaces

A sequence of elements a_n of $\mathbb{R}^d$ converges to a limit a if and only if, for each $\epsilon > 0$, the sequence a_n eventually lies within a ball of radius ϵ centered at a. It's okay if the first few (or few million) terms lie outside that ball— and the number of terms that do lie outside the ball may depend on how big ϵ is (if ϵ is small enough it typically *will* take millions of terms before the remaining sequence lies inside the ball). This can be made mathematically precise by introducing a letter (say, N_ϵ) for how many initial terms we have to throw away, so that $a_n \to a$ if and only if there is an $N_\epsilon < \infty$ so that, for each $n \geq N_\epsilon$, $|a_n - a| < \epsilon$: only finitely many a_n can be farther than ϵ from a.

The same notion of convergence really works in any *metric space*, where we require that some measure of the *distance* $\mathrm{d}(a_n, a)$ from a_n to a tend to zero in the sense that it exceeds each number $\epsilon > 0$ for at most some finite number N_ϵ of terms.

Points a_n in d-dimensional Euclidean space will converge to a limit $a \in \mathbb{R}^d$ if and only if each of their coordinates converges in $\mathbb{R}$; and, since there are only finitely many coordinates, if they all converge then they do so *uniformly* (*i.e.*, for each ϵ we can take the same N_ϵ for all d of the coordinate sequences), so all notions of convergence in $\mathbb{R}^d$ are equivalent. For example,

$$\max_{1\leq i\leq d} |x_i - y_i| \leq \Big[\sum (x_i - y_i)^2\Big]^{\frac{1}{2}} \leq \sum_{1\leq i\leq d} |x_i - y_i| \leq d \max_{1\leq i\leq d} |x_i - y_i|$$

so convergence is identical for all three of these metrics. Convergence is much more complex and interesting for random variables.

7.1 Convergence of Random Variables

For *random variables* X_n the idea of convergence to a limiting random variable X is more delicate, since each X_n is a *function* of $\omega \in \Omega$ and usually there are infinitely many points $\omega \in \Omega$. What should we mean in saying that a sequence X_n converges to a limit X? That $X_n(\omega)$ converges to $X(\omega)$ for each fixed ω? Or that $X_n(\omega)$ converges *uniformly* in $\omega \in \Omega$? Or that some notion of the distance $\mathrm{d}(X_n, X)$ between X_n and the limit X decreases to zero? Should the probability measure P be involved in some way?

Here are a few different choices of what we *might* mean by the statement that "X_n converges to X," for a sequence of random variables X_n and a random variable X, all defined on the same probability space $(\Omega, \mathcal{F}, \mathsf{P})$:

pw: The real numbers $X_n(\omega) \to X(\omega)$ for every $\omega \in \Omega$ (pointwise cgce):

$$(\forall \epsilon > 0)\,(\forall \omega \in \Omega)\,(\exists N_{\epsilon,\omega} < \infty)\,(\forall n \geq N_{\epsilon,\omega}) \quad |X_n(\omega) - X(\omega)| < \epsilon.$$

uni: The sequences of real numbers $X_n(\omega) \to X(\omega)$ *uniformly* for $\omega \in \Omega$:

$$(\forall \epsilon > 0)\,(\exists N_{\epsilon} < \infty)\,(\forall \omega \in \Omega)\,(\forall n \geq N_{\epsilon}) \quad |X_n(\omega) - X(\omega)| < \epsilon.$$

a.s.: Outside some null event $\mathcal{N} \in \mathcal{F}$, each sequence of real numbers $X_n(\omega) \to X(\omega)$ (Almost-Sure convergence, or convergence "almost everywhere" (*a.e.*)): for some $\mathcal{N} \in \mathcal{F}$ with $\mathsf{P}[\mathcal{N}] = 0$,

$$(\forall \epsilon > 0)\,(\forall \omega \notin \mathcal{N})\,(\exists N_{\epsilon,\omega} < \infty)\,(\forall n \geq N_{\epsilon,\omega}) \quad |X_n(\omega) - X(\omega)| < \epsilon,$$

$$\textit{i.e.,} \qquad \mathsf{P}\Big\{ \cup_{\epsilon>0} \cap_{N<\infty} \cup_{n\geq N} \quad |X_n(\omega) - X(\omega)| \geq \epsilon \Big\} = 0.$$

L_∞: Outside some null event $\mathcal{N} \in \mathcal{F}$, the sequences of real numbers $X_n(\omega) \to X(\omega)$ converge *uniformly* ("almost-uniform" or "L_∞" convergence): for some $\mathcal{N} \in \mathcal{F}$ with $\mathsf{P}[\mathcal{N}] = 0$,

$$(\forall \epsilon > 0)\,(\exists N_{\epsilon} < \infty)\,(\forall \omega \notin \mathcal{N})\,(\forall n \geq N_{\epsilon}) \quad |X_n(\omega) - X(\omega)| < \epsilon.$$

pr.: For each $\epsilon > 0$, the probabilities $\mathsf{P}[|X_n - X| > \epsilon] \to 0$ (convergence "in probability", or "in measure"):

$$(\forall \epsilon > 0)\,(\forall \eta > 0)\,(\exists N_{\epsilon,\eta} < \infty)\,(\forall n \geq N_{\epsilon,\eta}) \quad \mathsf{P}[|X_n - X| > \epsilon] < \eta.$$

L_1: The expectation $\mathsf{E}[|X_n - X|]$ converges to zero (convergence "in L_1"):

$$(\forall \epsilon > 0)\,(\exists N_{\epsilon} < \infty)\,(\forall n \geq N_{\epsilon}) \quad \mathsf{E}[|X_n - X|] < \epsilon.$$

L_p: For some fixed number $p > 0$, the expectation of the p^{th} absolute power $\mathsf{E}[|X_n - X|^p]$ converges to zero (convergence "in L_p," sometimes called "in the p^{th} mean"):

$$(\forall \epsilon > 0)\,(\exists N_{\epsilon} < \infty)\,(\forall n \geq N_{\epsilon}) \quad \mathsf{E}[|X_n - X|^p] < \epsilon.$$

dist.: The *distributions* of X_n converge to the distribution of X, *i.e.*, the measures $\mathsf{P} \circ X_n^{-1}$ converge in some way to $\mathsf{P} \circ X^{-1}$ ("vague" or "weak" convergence, or "convergence in distribution", sometimes written $X_n \Rightarrow X$):

$$(\forall \epsilon > 0)\,\big(\forall \phi \in \mathcal{C}_b(\mathbb{R})\big)\,(\exists N_{\epsilon,\phi} < \infty)\,(\forall n \geq N_{\epsilon,\phi}) \quad \mathsf{E}\big[|\phi(X_n) - \phi(X)|\big] < \epsilon.$$

Which of these eight notions of convergence is right for random variables? The answer is that *all* of them are useful in probability theory for one purpose or another. You will want to know which ones imply which other ones, under what conditions. All but the first two (pointwise, uniform) notions depend upon the measure P; it is possible for a sequence X_n to converge to X in any of these senses for one probability measure P, but to fail to converge for another P'. Most of them can be phrased as metric convergence for some notion of distance between random variables:

pr.: $X_n \to X$ in probability if and only if $\mathrm{d}_0(X, X_n) \to 0$ as real numbers, where:

$$\mathrm{d}_0(X,Y) := \mathsf{E}\left(\frac{|X-Y|}{1+|X-Y|}\right)$$

L_1: $X_n \to X$ in L_1 if and only if $\mathrm{d}_1(X, X_n) := \|X - X_n\|_1 \to 0$ as real numbers, where:

$$\|Z\|_1 := \mathsf{E}|Z|$$

L_p: $X_n \to X$ in L_p if and only if $\mathrm{d}_p(X, X_n) \to 0$ as real numbers, where:

$$\mathrm{d}_p(X,Y) := \begin{cases} (\mathsf{E}|X-Y|^p)^{1/p} & p \ge 1 \\ \mathsf{E}|X-Y|^p & 0 < p < 1. \end{cases}$$

L_∞: $X_n \to X$ almost uniformly if and only if $\mathrm{d}_\infty(X, X_n) := \|X - X_n\|_\infty \to 0$ as real numbers, where:

$$\|Z\|_\infty := \sup\{r \ge 0 : \ \mathsf{P}[|Z| > r] > 0\}$$

As the notation suggests, convergence in probability and in L_∞ are in some sense limits of convergence in L_p as $p \to 0$ and $p \to \infty$, respectively. Almost-sure convergence is an exception: there is no metric notion of distance $\mathrm{d}(X,Y)$ for which $X_n \to X$ almost surely if and only if $\mathrm{d}(X, X_n) \to 0$ (unless Ω is countable or P is atomic).

7.1.1 Almost-Sure Convergence

Let $\{X_n\}$ and X be a collection of RVs on some $(\Omega, \mathcal{F}, \mathsf{P})$. The set of points ω for which $X_n(\omega)$ does converge to $X(\omega)$ is just

$$\bigcap_{\epsilon>0} \bigcup_{m=1}^{\infty} \bigcap_{n=m}^{\infty} [\omega : |X_n(\omega) - X(\omega)| \le \epsilon],$$

the points which, for all $\epsilon > 0$, have $|X_n(\omega) - X(\omega)|$ less than ϵ for all but finitely-many n. The sequence X_n is said to converge "almost everywhere" (*a.e.*) to X, or to converge to X "almost surely" (*a.s.*), if this set of ω has probability one, or (conversely) if its complement is a null set:

$$\mathsf{P}\Big[\bigcup_{\epsilon>0} \bigcap_{m=1}^{\infty} \bigcup_{n=m}^{\infty} [\omega : |X_n(\omega) - X(\omega)| > \epsilon]\Big] = 0.$$

Despite their appearance these intersections and unions over $\epsilon > 0$ are only countable, since we need include only rational ϵ (or, for that matter, any sequence ϵ_k tending to zero, such as $\epsilon_k = 1/k$). Thus $X_n \to X$ *a.e.* if and only if, for each $\epsilon > 0$,

$$\mathsf{P}\Big[\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty}[\omega : |X_n(\omega) - X(\omega)| > \epsilon]\Big] = 0. \qquad (a.e.)$$

This combination of intersection and union occurs frequently in probability, and has a name; for any sequence E_n of events, $[\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty}E_n]$ is called the *lim sup* of the $\{E_n\}$, and is sometimes described more colorfully as $[E_n \ i.o.]$, the set of points in E_n "infinitely often." Its complement is the *lim inf* of the sets $F_n = E_n^c$, $[\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty}F_n]$: the set of points in all but finitely many of the F_n. Since P is countably additive, and since the intersection in the definition of *lim sup* is *decreasing* and the union in the definition of *lim inf* is *increasing*, always we have
$\mathsf{P}[\bigcup_{n=m}^{\infty}E_n] \searrow \mathsf{P}[\bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty}E_n]$ and $\mathsf{P}[\bigcap_{n=m}^{\infty}F_n] \nearrow \mathsf{P}[\bigcup_{m=1}^{\infty}\bigcap_{n=m}^{\infty}F_n]$ as $m \to \infty$. Thus,

Theorem 1 *$X_n \to X$ P-a.s. if and only if for every $\epsilon > 0$,*

$$\lim_{m\to\infty}\mathsf{P}[|X_n - X| > \epsilon \text{ for some } n \ge m] = 0.$$

In particular, $X_n \to X$ P-a.s. if $\sum_{n<\infty}\mathsf{P}[|X_n - X| > \epsilon] < \infty$ for each $\epsilon > 0$ (why?).

7.1.2 Convergence In Probability

The sequence X_n is said to converge to X "in probability" $(pr.)$ if, for each $\epsilon > 0$,

$$\mathsf{P}[\omega : |X_n(\omega) - X(\omega)| > \epsilon] \to 0. \qquad (pr.)$$

If we denote by E_n the event $[\omega : |X_n(\omega) - X(\omega)| > \epsilon]$ we see that convergence *almost surely* requires that $\mathsf{P}[\bigcup_{m\ge n}E_m] \to 0$ as $n \to \infty$, while convergence *in probability* requires only that $\mathsf{P}[E_n] \to 0$. For another proof, if $X_n \to X$ $(a.s.)$ then the DCT implies that $\mathsf{P}\big[|X_n - X| > \epsilon\big] = \mathsf{E}\mathbf{1}_{\{|X_n - X|>\epsilon\}} \to 0$ for each $\epsilon > 0$. Thus:

Theorem 2 *If $X_n \to X$* a.s. *then $X_n \to X$* pr.

Here is a partial converse:

Theorem 3 *If $X_n \to X$* pr.*, then there is a subsequence n_k such that $X_{n_k} \to X$* a.s.

Proof. Set $n_0 := 0$ and, for each integer $k \ge 1$, set

$$n_k := \inf\left\{n > n_{k-1} : \mathsf{P}\left[\omega : |X_n(\omega) - X(\omega)| > \frac{1}{k}\right] \le 2^{-k}\right\}.$$

For any $\epsilon > 0$ we have $\frac{1}{k} \le \epsilon$ eventually (namely, for $k \ge k_0 := \lceil \frac{1}{\epsilon} \rceil$) and for each $m \ge k_0$,

$$\begin{aligned}
\mathsf{P}\Big[\bigcup_{k=m}^{\infty} \{\omega : |X_{n_k}(\omega) - X(\omega)| > \epsilon\}\Big] &\le \mathsf{P}\Big[\bigcup_{k=m}^{\infty} \Big\{\omega : |X_{n_k}(\omega) - X(\omega)| > \frac{1}{k}\Big\}\Big] \\
&\le \sum_{k=m}^{\infty} \mathsf{P}\Big\{\omega : |X_{n_k}(\omega) - X(\omega)| > \frac{1}{k}\Big\} \\
&\le \sum_{k=m}^{\infty} 2^{-k} = 2^{1-m} \to 0 \text{ as } m \to \infty.
\end{aligned}$$

□

In fact, this characterizes convergence in probability:

Theorem 4 *Let $\{X_n\}$, X be random variables on $(\Omega, \mathcal{F}, \mathsf{P})$. Then $X_n \to X$* pr. *if and only if every sequence $\mathbb{N} \ni n_k \nearrow \infty$ has a subsequence n_{k_i} such that $X_{n_{k_i}} \to X$* a.s *as $i \to \infty$.*

Proof. The "only if" ($\Rightarrow$) direction is just Theorem 3. Suppose (for contradiction) that $X_n \not\to X$ *pr.*; then for some $\epsilon > 0$ and $\delta > 0$ there are infinitely-many n for which $\mathsf{P}[|X_n - X| > \epsilon] > \delta$. Let n_k be an increasing sequence satisfying this bound. By hypothesis, there is a subsequence along which $X_{n_{k_i}} \to X$ *a.s.*; but by Theorem 2, also $X_{n_{k_i}} \to X$ *pr*, so $\mathsf{P}[|X_n - X| > \epsilon] \to 0$, a contradiction. □

Theorem 5 *Let $\{X_n\}$, X be real-valued random variables on $(\Omega, \mathcal{F}, \mathsf{P})$ with $X_n \to X$* pr. *and let $\phi : \mathbb{R} \to \mathbb{R}$ be continuous. Then $Y_n := \phi(X_n) \to Y := \phi(X)$* pr.

Proof. For an easy but indirect proof, simply apply Theorem 4. For a more direct approach, begin by selecting any $\epsilon > 0$ and $\delta > 0$. Find a compact set $K_\epsilon \subset \mathbb{R}$ with $\mathsf{P}[X \in K_\epsilon] \ge 1 - \epsilon/2$; since ϕ is *uniformly* continuous on K_ϵ, find $\eta > 0$ such that

$$(\forall x \in K_\epsilon)(\forall y \in \mathbb{R}) \quad |x - y| \le \eta \quad \Rightarrow \quad |\phi(x) - \phi(y)| < \delta.$$

Now, since $X_n \to X$ *pr.*, find $N \in \mathbb{N}$ such that

$$(\forall n \ge N)\ \mathsf{P}[|X_n - X| > \eta] \le \epsilon/2.$$

Then for $n \ge N$,

$$\begin{aligned}
\mathsf{P}[|Y_n - Y| > \delta] &\le \mathsf{P}[X \notin K_\epsilon] + \mathsf{P}[X \in K_\epsilon,\ |\phi(X_n) - \phi(X)| > \delta] \\
&\le \mathsf{P}[X \notin K_\epsilon] + \mathsf{P}[X \in K_\epsilon,\ |X_n - X| > \eta] \\
&\le \epsilon/2 + \epsilon/2 = \epsilon.
\end{aligned}$$

□

The same result (with the same proof) is true for random variables X_n, X taking values in any σ-compact complete separable metric space $\mathcal{X}$ and, in particular, for $\mathcal{X} = \mathbb{R}^d$.

7.1.3 A Counter-Example

If $X_n \to X$ *a.s.* implies $X_n \to X$ *pr.*, and if the converse holds at least along subsequences, are the two notions really identical? Or is it possible for RVs X_n to converge to X *pr.*, but not *a.s.*? The answer is that the two notions *are* different, and that *a.s.* convergence is strictly stronger than convergence *pr.* Here's an example:

First notice that every integer $n \in \mathbb{N}$ can be written uniquely in the form $n = i + 2^j$ for integers $j \geq 0$ and $0 \leq i < 2^j$ (set $j := \lfloor \log_2 n \rfloor$ and $i := n - 2^j$). Let $(\Omega, \mathcal{F}, \mathsf{P})$ be the unit interval with Borel sets and Lebesgue measure. Define a sequence of random variables $X_n : \Omega \to \mathbb{R}$ by

$$X_n(\omega) = \begin{cases} 1 & \text{if } \frac{i}{2^j} < \omega \leq \frac{i+1}{2^j} \\ 0 & \text{otherwise} \end{cases} \qquad \text{where } n = i + 2^j,\ 0 \leq i < 2^j.$$

SO,

X_1	is	one on	1	interval	of length	1,
X_2, X_3	are	one on	2	intervals	each of length	1/2,
$X_4, \ldots, X_7$	are	one on	4	intervals	each of length	1/4,
$X_8, \ldots, X_{15}$	are	one on	8	intervals	each of length	1/8,

and, in general, each X_n is one on an interval of length 2^{-j}. Since $\frac{1}{n} \leq \frac{1}{2^j} < \frac{2}{n}$,

$$\mathsf{P}[|X_n| > \epsilon] = 2^{-j} < \frac{2}{n} \to 0$$

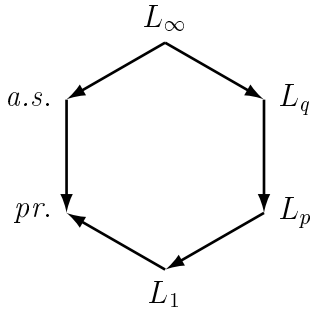
for each $0 < \epsilon < 1$ and $X_n \to 0$ *pr.* On the other hand, for every $j > 0$ we have each $\omega \in \Omega$ in one of the 2^j intervals of length 2^{-j} where some X_n is one,

$$(\forall j \in \mathbb{N}) \quad \Omega = \bigcup_{i=0}^{2^j-1} \left(\frac{i}{2^j}, \frac{i+1}{2^j} \right] = \bigcup_{n=2^j}^{2^{j+1}-1} \left[\omega : X_n(\omega) = 1 \right]$$

For every $\omega \in \Omega$ and $j \in \mathbb{N}$ there is some $n \geq 2^j$ with $X_n(\omega) = 1$ (and so there are infinitely-many such X_n); thus $[\omega : X_n(\omega) \to 0]$ is *empty*, not a set of probability one! Obviously X_n does not converge *a.s*, although it does converge *pr.*

This example is a building-block for several examples to come, so getting to know it well is worthwhile. Try to verify that $X_n \to 0$ in probability (how large must n be to ensure $\mathsf{P}[|X_n| > \delta] < \epsilon$?) and in L_p (how large must n be to ensure $\|X_n\|_p < \epsilon$?) but *not* almost surely (what *is* $\{\omega : \ X_n \to 0\}$? Why?). Find $\|X_n\|_p$ explicitly. Why *doesn't* $X_n \to 0$ *a.s.*? What would happen if we multiplied X_n by n? By n^2? By $j := \lfloor \log_2 n \rfloor$? What about the subsequence $Y_n := X_{2^n}$? Does X_n converge in L_∞?

In summary, for $1 < p < q < \infty$ the convergence implications are:



with *partial* converses ($pr. \to a.s.$ along subsequences, $pr. \to L_1, L_p, L_q$ under UI— see the next section). All of these imply convergence in distribution, which we'll consider later.

7.2 Uniform Integrability

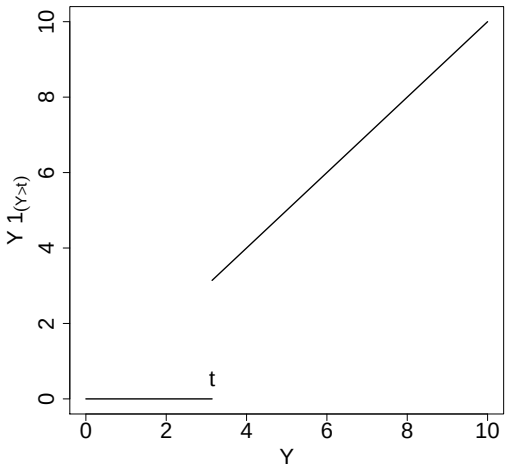
Let $Y \geq 0$ be integrable on some probability space $(\Omega, \mathcal{F}, \mathsf{P})$,

$$\mathsf{E}[Y] = \int_\Omega Y \, d\mathsf{P} < \infty.$$

By Lebesgue's DCT it follows that

$$\lim_{t\to\infty} \mathsf{E}[Y \, \mathbf{1}_{\{Y>t\}}] = \lim_{t\to\infty} \int_{[\omega: Y(\omega)>t]} Y \, d\mathsf{P} = 0$$

since $\left\{Y \, \mathbf{1}_{\{Y>t\}}\right\}$ is dominated by $Y \in L_1$ and converges to zero $a.s.$ as $t \to \infty$.



Consequently, for any sequence of random variables X_n that are dominated by Y in the sense that $|X_n| \leq Y$ $a.s.$,

$$\mathsf{E}[|X_n| \, \mathbf{1}_{\{|X_n|>t\}}] \leq \mathsf{E}[Y \, \mathbf{1}_{\{Y>t\}}] \to 0, \text{ uniformly in } n.$$

Call the sequence X_n *uniformly integrable* (or simply UI) if $\mathsf{E}[|X_n|\,\mathbf{1}_{\{|X_n|>t\}}] \to 0$ uniformly in n as $t \to \infty$, even if it isn't dominated by some integrable $Y \in L_1$. The big result is:

Theorem 6 *If $X_n \to X$ pr. and if X_n is UI then $X_n \to X$ in L_1.*

Proof. Without loss of generality take $X \equiv 0$. Fix any $\epsilon > 0$; find (by UI) $t_\epsilon > 0$ such that $\mathsf{E}[|X_n|\,\mathbf{1}_{\{X_n>t_\epsilon\}}] \le \epsilon$ for all n. Now (by $X_n \to X$ *pr.*) find $N_\epsilon \in \mathbb{N}$ such that, for $n \ge N_\epsilon$, $\mathsf{P}[|X_n| > \epsilon] < \epsilon/t_\epsilon$. Then:

$$\begin{aligned}
\mathsf{E}\big[|X_n|\big] &= \int_{[|X_n|\le\epsilon]} |X_n|\,d\mathsf{P} + \int_{[\epsilon<|X_n|\le t_\epsilon]} |X_n|\,d\mathsf{P} + \int_{[t_\epsilon<|X_n\|]} |X_n|\,d\mathsf{P} \\
&\le \int_{[|X_n|\le\epsilon]} \epsilon\,d\mathsf{P} \quad + \int_{[\epsilon<|X_n|\le t_\epsilon]} t_\epsilon\,d\mathsf{P} \quad + \int_{[t_\epsilon<|X_n\|]} |X_n|\,d\mathsf{P} \\
&\le \epsilon \qquad\qquad + t_\epsilon \times \mathsf{P}\big[|X_n| > \epsilon\big] \quad + \epsilon \\
&\le 3\epsilon. \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad \square
\end{aligned}$$

Theorem 6 has a partial converse— if $\{X_n\} \subset L_1$ and $X_n \to X$ *pr.* and if $\mathsf{E}|X_n| \to \mathsf{E}|X| < \infty$, then $\{X_n\}$ is UI (see Theorem 9). For another proof, first note $\mathsf{E}[|X_n|\mathbf{1}_{\{|X_n|>t\}}] = \mathsf{E}[|X_n|] - \mathsf{E}[|X_n|\mathbf{1}_{\{|X_n|\le t\}}]$. For $\epsilon > 0$ first pick $t < \infty$ s.t. $\mathsf{E}[|X|\mathbf{1}_{\{|X|>t\}}] < \epsilon$ (possible, since $X \in L_1$); then (by DCT) find N s.t. for $n \ge N$, $\|X_n\mathbf{1}_{\{|X_n|\le t\}} - X\mathbf{1}_{\{|X|\le t\}}\|_1 < \epsilon$ and (by assumption) $\big|\mathsf{E}|X_n| - \mathsf{E}|X|\big| < \epsilon$. Then $\mathsf{E}[|X_n|\mathbf{1}_{\{|X_n|>t\}}] \le 3\epsilon$ for $n \ge N$; since $X_j \in L_1$ for $j < N$, the result follows (with $t^* = \max(t, t_1, ..., t_{N-1})$).

Similarly, for any $p > 0$, $X_n \to X$ (*pr.*) and $|X_n|^p$ UI (for example, $|X_n| \le Y \in L_p$, or $\|X_n\|_q \le B < \infty$ for some $q > p$) gives $X_n \to X$ (L_p). In the special case of $|X_n| \le Y \in L_1$ this is just Lebesgue's **Dominated Convergence Theorem** (DCT), in a little stronger version than we proved in Week 4 because here we require only $X_n \to X$ (*pr.*), while then we required $X_n \to X$ (*a.s.*).

We have seen that $\{X_n\}$ is UI whenever $|X_n| \le Y \in L_1$, but UI is more general than that. Here are two more criteria. The first is called the "crystal ball condition":

Theorem 7 *If $\{X_n\}$ is uniformly bounded in L_p for some $p > 1$ then $\{X_n\}$ is UI.*

Proof. Let $B \in \mathbb{R}_+$ be an upper bound for $\|X_n\|_p$ and set $q := \frac{p}{p-1}$. By Hölder's inequality

$$\begin{aligned}
\mathsf{E}\left[|X_n|\,\mathbf{1}_{\{|X_n|>t\}}\right] &\le \|X_n\|_p \ \|\mathbf{1}_{\{|X_n|>t\}}\|_q \\
&\le B\ \left\{\mathsf{P}[|X_n| > t]\right\}^{1/q} \\
&\le B\ \left\{\mathsf{E}\left[|X_n|^p/t^p\right]\right\}^{1/q} \quad \text{by Markov's inequality} \\
&= B\ \left\{\|X_n\|_p^p/t^p\right\}^{1/q} \\
&= B\ \|X_n\|_p^{p-1}/t^{p-1} \qquad \text{since } p/q = p-1 \\
&\le B^p\ t^{1-p} \to 0
\end{aligned}$$

uniformly in $\{n\}$. $\square$

Theorem 8 *The random variables $\{X_\alpha\}$ are UI if and only if they are uniformly bounded in L_1 and*

$$(\forall \epsilon > 0)(\exists \delta > 0)(\forall A \in \mathcal{F}) \quad \mathsf{P}(A) < \delta \Rightarrow \mathsf{E}[|X_\alpha|\mathbf{1}_A] < \epsilon.$$

If $(\Omega, \mathcal{F}, \mathsf{P})$ is non-atomic, the condition "$\{X_\alpha\}$ is uniformly bounded in L_1" is unnecessary.

Proof. Straightforward. If $(\Omega, \mathcal{F}, \mathsf{P})$ is non-atomic, for each $\epsilon > 0$ find $\delta > 0$ by hypothesis and then cover Ω with $N := \lfloor 1 + 1/\delta \rfloor$ sets $A_i \in \mathcal{F}$ with $\mathsf{P}(A_i) < \delta$ to see $\mathsf{E}|X_\alpha| < N\epsilon$ uniformly. Try to offer an example to illustrate what can go wrong if P has an atom $\omega^* \in \Omega$ with $c = \mathsf{P}[\{\omega_*\}] > 0$.
In fact, $X_n \to X$ in L_1 **if and only if** $\{X_i\}$ are UI:

Theorem 9 *Let $\{X_n\}, X \in L_1(\Omega, \mathcal{F}, \mathsf{P})$ with $X_n \to X$ pr. Then the following are equivalent:*

1. *$\{X_n\}$ is UI*
2. *$X_n \to X$ in L_1*
3. *$\mathsf{E}|X_n| \to \mathsf{E}|X|$*

Proof. Since $1 \Rightarrow 2$ by Theorem 6 and $2 \Rightarrow 3$ by the triangle inequality, it remains only to show that $3 \Rightarrow 1$. Let $X_n \to X$ *pr.* and $\mathsf{E}|X_n| \to \mathsf{E}|X|$. Fix $\epsilon > 0$ and, for each $t > 0$, set

$$f_t(x) := \begin{cases} |x| & 0 \le |x| \le t \\ t(t+1-|x|) & t < |x| \le t+1 \\ 0 & t+1 < |x| < \infty \end{cases}$$

and note that $0 \le |x|\mathbf{1}_{\{|x| \le t\}} \le f_t(x) \le |x|\mathbf{1}_{\{|x| \le t+1\}}$ and $0 \le f_t(x) \le t$.
Since $X \in L_1$, $\mathsf{E}|X|\mathbf{1}_{\{|X|>t\}} \to 0$ as $t \to \infty$ by DCT so we can find t sufficiently large that

$$\mathsf{E}|X|\mathbf{1}_{\{|X|>t\}} < \epsilon. \tag{1}$$

Since f_t is continuous and bounded, $f_t(X_n) \to f_t(X)$ as $n \to \infty$ in probability by Theorem 5 and also in L_1 by DCT. For large enough n (say, $n \ge N'$), $\|f_t(X_n) - f_t(X)\|_1 \le \epsilon$ and so

$$\mathsf{E}|X_n|\mathbf{1}_{\{|X_n \le t+1|\}} \ge \mathsf{E}f_t(X_n) \ge \mathsf{E}f_t(X) - \epsilon \ge \mathsf{E}|X|\mathbf{1}_{\{|X| \le t\}} - \epsilon. \tag{2}$$

Since $\mathsf{E}|X_n| \to \mathsf{E}|X|$, for sufficiently large n (say, $n \ge N \ge N'$)

$$\mathsf{E}|X_n| \le \mathsf{E}|X| + \epsilon. \tag{3}$$

Upon subtracting (2) from (3) we have

$$\mathsf{E}|X_n|\mathbf{1}_{\{|X_n|>t+1\}} \le \mathsf{E}|X|\mathbf{1}_{\{|X|>t\}} + 2\epsilon.$$

Applying (1) we get

$$\mathsf{E}|X_n|\mathbf{1}_{\{|X_n|>t+1\}} < 3\epsilon$$

completing the proof that $\{X_n\}$ is UI.

□

7.3 Cauchy Convergence

Every form of convergence we have considered for random variables *except* a.s. can be represented as convergence in a metric space. A sequence X_n converges to a limit X in a metric space if every ball centered at X contains all but finitely-many of $\{X_n\}$. Sometimes we wish to consider a sequence X_n that converges to *some* limit, without knowing the limit in advance. The concept of *Cauchy Convergence* is ideal for this— we insist that for each $\epsilon > 0$, all but finitely-many of the ϵ-balls centered *at points X_m of the sequence* contain all but finitely-many of the points. For any of distance measures d_p of Section (7.1), with $0 \le p \le \infty$, say "X_n is a Cauchy sequence in L_p" if

$$(\forall \epsilon > 0)(\exists N_\epsilon < \infty)(\forall m, n \ge N_\epsilon) \qquad \mathrm{d}_p(X_m, X_n) < \epsilon.$$

The spaces L_p for $0 \le p \le \infty$ are all *complete* in the sense that *if* X_n is Cauchy for d_p *then* there exists $X \in L_p$ for which $\mathrm{d}_p(X_n, X) \to 0$. To see this, take an increasing subsequence N_k along which $\mathrm{d}_p(X_m, X_n) < 2^{-k}$ for $n \ge m \ge N_k$, and set $X_0 := 0$ and $N_0 = 0$; for $k \in \mathbb{N}$ set $Y_k := X_{N_k} - X_{N_{k-1}}$. Check to confirm that $X := \sum_{k=1}^{\infty} Y_k$ is well-defined and $X \in L_p$ *a.s.*, and that $\mathrm{d}_p(X_n, X) \to 0$.

7.4 Convergence in Distribution

Some of the most famous results in probability theory concern the limiting *distribution* of sequences of random variables. For example,

- **Central Limit Theorem**: For iid $\{X_n\} \subset L_2$ and large n, $(S_n - n\mu)/\sqrt{n\sigma^2}$ has approximately the Normal $\mathsf{No}(0,1)$ distribution;
- **Law of Small Numbers**: For independent $\mathbb{N}_0$-valued random variables X_i with small means μ_i, S_n has approximately the Poisson $\mathsf{Po}(\lambda_n)$ distribution with mean $\lambda_n = \sum \mu_i$;
- **Extreme Value** or **Three Types Theorem**: For iid $\{X_n\}$ and suitable sequences $\{a_n\}$, $\{b_n\}$ of constants, $\max\{(X_j - a_n)/b_n : \ j \le n\}$ has approximately the Gumbel, Fréchet, or reversed Weibull distribution.

But what does it mean to have a distribution "approximately"???

If X_n has the discrete uniform distribution on the points $\{i/n\}$ for $1 \le i \le n$, and if n is huge (say, $2^{51} \approx 10^{17}$) then most of us would regard X_n as "approximately" a standard uniform $\mathsf{Un}(0,1)$ random variable— in fact, this is the best one can hope for in a 64-bit double-precision floating-point representation. Yet each X_n has a *discrete* distribution while the limit is a *continuous* distribution, so whatever we mean by "converging in distribution" it can't involve either density functions or probability mass functions, and it is too much to ask for the distributions $\mu_n(B) = \mathsf{P}[X_n \in B]$ to converge for every Borel set $B \subset \mathbb{R}$.

In this example the distribution functions $F_n(x) = \mathsf{P}[X_n \le x]$ do in fact converge at every $x \in \mathbb{R}$, but that too is a bit too much to ask— consider constant random variables $Y_n :=$

$-1/n$, for example, which converge almost-surely to $Y := 0$, but the distribution functions do not converge at the point $y = 0$. This is closer to the right idea, however:

Definition 1 *Let $\{\mu_n(dx)\}$ and $\mu(dx)$ be distributions on a measurable space $(\mathcal{X}, \mathcal{E})$. Say "$\mu_n$ converges in distribution to μ" or write "$\mu_n \Rightarrow \mu$" if*

$$\int_{\mathcal{X}} \phi(x)\, \mu_n(dx) \to \int_{\mathcal{X}} \phi(x)\, \mu(dx)$$

as $n \to \infty$ for every continuous bounded function $\phi : \mathcal{X} \to \mathbb{R}$.

If $\{X_n\}$ are random variables with distributions $\{\mu_n\}$, we also say that "X_n converges in distribution to μ" and write "$X_n \Rightarrow \mu$." This definition works not only for real-valued random variables where $\mathcal{X} = \mathbb{R}$ and $\mathcal{E} = \mathcal{B}$ is the Borel sets, but also in $\mathbb{R}^d$ or any other complete separable metric space $\mathcal{X}$ with Borel sets $\mathcal{E}$. The connection with distribution functions is given by:

Proposition 1 *Let $\{\mu_n(dx)\}$ and $\mu(dx)$ be distributions on the real line $(\mathbb{R}, \mathcal{B})$, and let $F_n(x) := \mu_n\big((-\infty, x]\big)$ and $F(x) := \mu\big((-\infty, x]\big)$ be the associated DFs. Then the following are equivalent:*

- *$\mu_n \Rightarrow \mu$;*
- *$F_n(x) \to F(x)$ on any dense set of $\{x\}$ in $\mathbb{R}$;*
- *$F_n(x) \to F(x)$ at every $x \in \mathbb{R}$ where $F(x-) = F(x)$ is continuous.*

A similar but more general result for complete separable metric spaces $\mathcal{X}$ (like $\mathbb{R}^n$) with their Borel sets is that $\mu_n \Rightarrow \mu$ if and only if $\mu_n(B) \to \mu(B)$ for every Borel set B whose boundary $\partial B := \overline{B} \cap \overline{B^c}$ has zero measure for the limiting distribution, $\mu(\partial B) = 0$. Convergence in distribution is weaker than any of the other forms of convergence we have considered:

Proposition 2 *Let $\{X_n\}$, X be random variables on some probability space $(\Omega, \mathcal{F}, \mathsf{P})$ with distributions $\{\mu_n\}$, μ, and suppose $X_n \to X$ pr. Then $X_n \Rightarrow \mu$.*

The proof is a simple application of the UI Convergence Theorem to the (uniformly bounded and hence UI) random variables $\phi(X_n)$.

Convergence in distribution is *strictly* weaker than any other notion, because it even applies if the $\{X_n\}$ are defined on different probability spaces $(\Omega_n, \mathcal{F}_n, \mathsf{P}_n)$, where none of the other forms of convergence even makes sense. As usual there is a *partial* converse, however:

Proposition 3 *Let $\mu_n \Rightarrow \mu$ on the real line $(\mathbb{R}, \mathcal{B})$. Then there exists a probability space $(\Omega, \mathcal{F}, \mathsf{P})$ and random variables $\{X_n\}$, X with distributions $\{\mu_n\}$, μ for which $X_n \to X$ a.s.*

It is enough to take $\Omega = (0, 1]$, $\mathcal{F} = \mathcal{B}$, $\mathsf{P} = \lambda$, and for $\omega \in \Omega$ set

$$X_n(\omega) := \inf\big\{x \in \bar{\mathbb{R}}_+ :\ F_n(x) \ge \omega\big\} \qquad X(\omega) := \inf\big\{x \in \bar{\mathbb{R}}_+ :\ F(x) \ge \omega\big\}$$

and verify that $X_n(\omega) \to X(\omega)$ for all but countably-many $\omega \in \Omega$.

7.5 Metrics on Distributions [Optional]

Convergence in distribution can be metrized by the *Lévy-Prokhorov* metric. For Borel probability measures μ, ν on a complete separable metric space $(\mathcal{X}, d)$, set

$$\begin{aligned} \mathrm{d}_{LP}(\mu,\nu) := \inf\Big\{\epsilon>0:\ \mu(A) \le \nu(A^\epsilon)+\epsilon \\ \text{and} \quad \nu(A) \le \mu(A^\epsilon)+\epsilon \text{ for all } A\in\mathcal{B}(\mathcal{X})\Big\} \end{aligned} \tag{4a}$$

where $A^\epsilon := \{y\in\mathcal{X}:\ (\exists x\in A)\ \mathrm{d}(x,y)<\epsilon\} = \cup_{x\in A} B_\epsilon(x)$, the union of open ϵ-balls around all points in A. In $d=1$ dimension this reduces to the *Lévy metric* for the DFs F and G of μ and ν,

$$\mathrm{d}_{LP}(F,G) := \inf\{\epsilon>0 \mid F(x-\epsilon)\le G(x)\le F(x+\epsilon) \text{ for all } x\in\mathbb{R}\}$$

While convergence in probability (and metric (4a)) is the most important notion of closeness and convergence of probability distributions, several others arise. Many of these are described and compared in the Central Limit Theorem notes for Week 9 at

`http://www.stat.duke.edu/courses/Fall16/sta711/lec/topics/dstn.pdf`.

Some of the most important are:

Total Variation:

$$\begin{aligned} \mathrm{d}_{TV}(\mu,\nu) :=& \sup\{|\mu(A)-\nu(A)|:\quad A\in\mathcal{B}(\mathcal{X})\} \\ =& \inf\{\mathsf{P}[X\ne Y:\quad X\sim\mu,\ Y\sim\nu]\}\,. \end{aligned} \tag{4b}$$

If μ and ν have density functions with respect to some σ-finite reference measure λ (like Lebesgue measure, for continuous distributions, or counting measure, for discrete ones), this can be evaluated as half the L_1 distance between their densities,

$$= \tfrac{1}{2}\int_{\mathcal{X}} |f(x)-g(x)|\ \lambda(dx),$$

Hellinger:

Again if $\mu(dx)=f(x)\lambda(dx)$ and $\nu(dx)=g(x)\lambda(dx)$ have densities, the *Hellinger Distance* is

$$\begin{aligned} \mathrm{d}_H(\mu,\nu) :=& \left\{\frac{1}{2}\int_{\mathcal{X}}\left(\sqrt{f(x)}-\sqrt{g(x)}\right)^2\ \lambda(dx)\right\}^{1/2} \\ =& \left\{1-\int_{\mathcal{X}}\sqrt{f(x)g(x)}\,\lambda(dx)\right\}^{1/2} \end{aligned}$$

Hellinger and Total Variation determine the same topology, *i.e.*, so a sequence converges in one if and only if it does in the other.

Kolmogorov-Smirnov:

The *Kolmogorov-Smirnov* distance between two probability measures on the real line $\mathcal{X} = \mathbb{R}$ is the L_∞ distance between their DFs:

$$\mathrm{d}_{KS}(\mu,\nu) := \sup_{x\in\mathbb{R}} \left\{ \left| \mu\big((-\infty,x]\big) - \nu\big((-\infty,x]\big) \right| \right\} \tag{4c}$$

Wasserstein W_1:

The *Wasserstein* W_1 distance between two distributions on a complete separable metric space $(\mathcal{X},\mathrm{d})$ is

$$\begin{aligned} \mathrm{d}_W(\mu,\nu) &:= \inf_{\gamma\in\Gamma(\mu,\nu)} \iint_{\mathcal{X}\times\mathcal{X}} \mathrm{d}(x,y)\,\gamma(dx\,dy) \\ &= \inf\left\{ \|X-Y\|_1 :\ X\sim\mu,\ Y\sim\nu \right\} \end{aligned} \tag{4d}$$

where $\Gamma(\mu,\nu)$ is the space of probability measures on $\mathcal{X}^2$ with marginals μ and ν. It can also be evaluated in terms of unit Lipschitz functions $f(x)$ as:

$$= \sup_{f:\mathcal{X}\to\mathbb{R}} \left\{ \left| \int_{\mathcal{X}} f(x)\,\mu(dx) - \int_{\mathcal{X}} f(x)\,\nu(dx) \right| :\ \mathrm{Lip}(f)\le 1 \right\}.$$

Sometimes called the "transportation metric," d_W can be interpreted as the minimum cost of moving mass distributed on $\mathcal{X}$ according to μ to mass distributed according to ν if moving cost is proportional to the product of mass times distance. There are also L_p versions of this metric, the Wasserstein W_p distance, but they are encountered less frequently.

Kullback-Leibler Divergence:

The "Kullback-Leibler divergence" (Kullback and Leibler, 1951), also called Relative Entropy, from distribution μ (with pdf $f(x)$) to ν (with pdf $g(x)$) on a Polish space $\mathcal{X}$ is:

$$\mathrm{KL}(\mu\|\nu) := \int_{\mathcal{X}} -\log\left[\frac{g(x)}{f(x)}\right]\ f(x)\,\lambda(dx). \tag{4e}$$

It is nonnegative, because $\log y \le y-1$ for all $y>0$ (or by Jensen's inequality), but it is not actually a distance metric because it's not symmetric in μ and ν and doesn't satisfy the triangle inequality. It does determine a topology, though, and hence a notion of convergence.

7.6 Summary: UI and Convergence Concepts

I. Uniform Integrability (UI)

A. $|X_n| \le Y \in L_r$, $r>0$, implies $\int_{[|X_n|>t]} |X_n|^r\,d\mathsf{P} \le \int_{[Y>t]} Y^r\,d\mathsf{P} \to 0$ as $t\to\infty$, uniformly in n, so $\{|X_n|^r\}$ is UI. This is the *definition* of UI for $r=1$.

B. If $(\Omega,\mathcal{F},\mathsf{P})$ nonatomic, X_n UI iff $(\forall\epsilon)(\exists\delta) \ni \mathsf{P}[\Lambda]\le\delta \Rightarrow \int_\Lambda |X_n|\,d\mathsf{P} < \epsilon$ (take $\delta = \frac{\epsilon}{2t}$). If $(\Omega,\mathcal{F},\mathsf{P})$ has atoms, also need uniform bound $\mathsf{E}|X_n| \le B$.

C. $\mathsf{E}|X_n|^q \le c^q < \infty$ implies $|X_n|^p$ UI for each $p < q$.

1. Remark: *not* for $p = q$ (counter-eg: $X_n := n\mathbf{1}_{(0,1/n]}$)

D. Main result: If $X_n \to X$ *pr.*, then

$$|X_n|^p \text{ is UI} \quad \Leftrightarrow \quad X_n \to X \text{ in } L_p \quad \Leftrightarrow \quad \mathsf{E}|X_n|^p \to \mathsf{E}|X|^p.$$

II. Convergence in Distribution (aka *Vague* Convergence)

A. $X_n \to X$ *pr.* iff $(\forall n_k \uparrow)(\exists n_{k_i} \uparrow) \ni X_{n_{k_i}} \to X$ *a.s.* (by contradiction)

B. $X_n \to X$ *a.s.* and $g(x)$ continuous implies $g(X_n) \to g(X)$ *a.s.*

C. $X_n \to X$ *pr.* and $g(x)$ continuous $\Rightarrow g(X_n) \to g(X)$ *pr.* (use A, B)

D. Definition: $X_n \Rightarrow X$ if $\big(\forall \phi \in C_b(\mathbb{R})\big)\mathsf{E}\phi(X_n) \to \mathsf{E}\phi(X)$

1. Prop: $X_n \to X$ *pr.* implies $X_n \Rightarrow X$ (use II.C)

2. Prop: $X_n \Rightarrow X$ implies $F_n(r) \to F(r)$ for all r s.t. $F(r) = F(r-)$.

a. Remark: Even if $X_n \Rightarrow X$, $F_n(r)$ may not converge where $F(r)$ jumps;

b. Remark: Even if $X_n \Rightarrow X$, $f_n(r) := F_n'(r)$ may not converge to $f(r) := F'(r)$; in fact, either may fail to exist.

III. Implications among cgce notions: *a.s.*, *pr.*, L_p, L_q, L_∞, *dist.* ($0<p<q<\infty$):

A. *a.s.* $\Longrightarrow$ *pr.* (by Easy Borel-Cantelli)

1. *pr.* $\Longrightarrow$ *a.s.* along subsequences

2. *pr.* $\not\Longrightarrow$ *a.s.* (counter-eg: $X_n(\omega) = \mathbf{1}_{(i/2^j,(i+1)/2^j]}(\omega)$, $n = i + 2^j$)

B. $L_p \Longrightarrow$ *pr.* (by Markov's inequality)

1. *pr.* $\Longrightarrow L_p$ under Uniform Integrability

2. *pr.* $\not\Longrightarrow L_p$ (counter-eg: $X_n = n^{1/p}\mathbf{1}_{(0,1/n]}$)

C. $L_q \Longrightarrow L_p$ for $p < q$ (by Jensen's inequality)

1. $L_p \not\Longrightarrow L_q$ (counter-eg: $X_n = n^{1/q}\mathbf{1}_{(0,1/n]}$)

D. $L_\infty \Longrightarrow L_p$ (simple estimate, or $\|X\|_p \nearrow$ as $p \nearrow$ by Jensen)

1. $L_p \not\Longrightarrow L_\infty$ (counter-eg: $X_n = \mathbf{1}_{(0,1/n]}$)

E. $L_\infty \Longrightarrow$ *a.s.* (almost-uniform cgce implies almost-pointwise cgce)

F. *pr.* $\Longrightarrow$ *dist.* (II.D.1 above)

1. *dist.* $\not\Longrightarrow$ *pr.* (counter-eg: X_n, X on different spaces or iid $\{X_n\}$)

2. *dist.* $\Longrightarrow$ *a.s.* ($\exists(\Omega, \mathcal{F}, \mathsf{P})$, $X_n, X \ni X_n \to X$ *a.s.*)

References

Kullback, S. and Leibler, R. A. (1951), "On Information and sufficiency," *Annals of Mathematical Statistics*, 22, 79–86.