What's the craic?

Today we are going to speak about convergences, their kinds and applications.

1 For those unprepared

First of all, we have to figure out the framework of the discussion. Our starting point is a space, the space X is just a set with any kinds of elements (elementary outcomes, animals, functions, anything else could represent the space). $\mathcal{F}$ - is some family of subsets. (Usually, $\mathcal{F}$ is a semi-ring, meaning that
1) $\varnothing \in \mathcal{F}$,
2) $\forall A, B \subset \mathcal{F}, A \cap B \subset \mathcal{F}$,
3) $\forall A, A_1 \subset A : \exists A_2, \ldots, A_n \, s.t. \, A_1 \sqcup A_2 \sqcup \cdots \sqcup A_n = A$)

By measure we will further imply the function which allows us to "measure"those subsets, meaning that it maps those with some real numbers, formally $\mu : \mathcal{F} \to [0, +\infty]$. Measure properties:

1. $\mu(\varnothing) = 0$
2. $\forall A, B \, s.t. \, A \cap B = \varnothing, \mu(A \cup B) = \mu(A) + \mu(B)$
3. People usually speak about the countably additive measure (for the purpose of adding the measures).

The measure is called a probability measure, when $\mathcal{F}$ is not just any kind of subset family, but a σ-algebra. Meaning that:

1. The space itself $X \subset \mathcal{F}$
2. If $A \subset \mathcal{F}$, то и $X \backslash A \subset \mathcal{F}$
3. Countable union or intersection of sets from $\mathcal{F}$ lies in $\mathcal{F}$

Moreover, μ now yields values from 0 to 1, in other words,
μ: $\mathcal{F} \to [0, 1]$ and $\mu(X) = 1$. Now, hopefully, $(X, \mathcal{F}, \mu)$ looks not that scary and denotes the probability space.

2 Convergences

Today, all convergences are related to the probability theory, hence we look at the functional sequences in probability spaces
This time we will consider

- Convergence in L^p (when $p = 1$ it is called the convergence in mean, when $p = 2$ - mean-squared convergence)
- Convergence almost surely
- Convergence in probability (The special case of convergence in measure)
- Convergence in distribution

For the most interested people: you can also look up the unitary and pointwise convergence(latter even yields the almost sure convergence!).

2.1 Convergence in distribution

The weakest of all convergences, follows from any other convergence, the converse is generally false.

Rigid definition:
The sequence of random variables X_n converges in distribution to the random variable X when

$$\lim_{n\to\infty} E(f(X_n)) = E(f(X))$$

For any continuous bounded function $f : \mathbb{R} \to \mathbb{R}$
Example of application of the convergence in distribution:

Theorem (Central Limit Theorem):
Let $X_1, X_2, \ldots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with finite mean μ and finite variance σ^2. Define the standardized sum:

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu)$$

or equivalently, the standardized sample mean:

$$Z_n = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$$

where $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$ is the sample mean. Then, as $n \to \infty$, we have:

$$Z_n \xrightarrow{D} \mathcal{N}(0,1)$$

That is, the random variable Z_n converges in distribution to a standard normal distribution $\mathcal{N}(0,1)$.

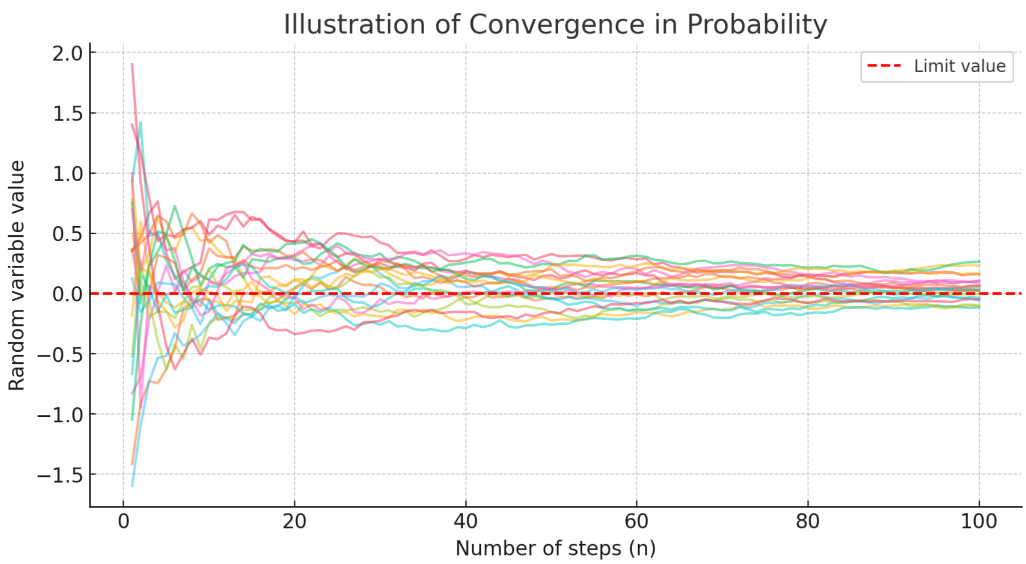
2.2 Convergence in Probability

The second weakest form of convergence, from which convergence in distribution follows. It can be derived both from almost sure convergence and from convergence in L^p.

Strict Definition:
A sequence of random variables X_n converges in probability to a random variable X if

$$\forall \varepsilon > 0 : P(|X_n - X| > \varepsilon) \to 0, \text{as } n \to \infty.$$

Intuitively, the larger n becomes, the less frequently the sequence X_n deviates from X.



Now let us show why convergence in probability implies convergence in distribution.

Proof:
We need to show that if $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{d} X$.

1. **Use the definition of convergence in probability**
   From the definition of convergence in probability, it follows that for any $\varepsilon > 0$, the probability $P(|X_n - X| > \varepsilon)$ tends to zero. This means that the values of X_n become arbitrarily close to X with high probability as $n \to \infty$.

2. **Show that this implies convergence in distribution**
   Consider the cumulative distribution function $F_{X_n}(x) = P(X_n \leq x)$. For any x in the continuity set of $F_X(x)$, take a small $\varepsilon > 0$. Then:
   $$P(X_n \leq x) = P(X_n \leq x, |X_n - X| < \varepsilon) + P(X_n \leq x, |X_n - X| \geq \varepsilon).$$
   - The first term $P(X_n \leq x, |X_n - X| < \varepsilon)$ is approximately equal to $P(X \leq x + \varepsilon)$ for large n, since X_n is close to X ($X \leq x$ and $|X_n - X| < \varepsilon \Rightarrow X - \varepsilon \leq X_n \leq X + \varepsilon, \Rightarrow X - \varepsilon \leq X_n \leq x, \Rightarrow X \leq x + \varepsilon$).
   - The second term $P(X_n \leq x, |X_n - X| \geq \varepsilon)$ tends to 0, since $P(|X_n - X| > \varepsilon) \to 0$ by the definition of convergence in probability.

   Then:
   $$P(X_n \leq x) \approx P(X \leq x + \varepsilon) \to P(X \leq x) \text{ as } \varepsilon \to 0.$$
   In other words $F_{X_n}(x) \to F_X(x)$, which means the convergence in distribution.

An example of practical application:

Weak Law of Large Numbers (Chebyshev's form):
Let $X_1, X_2, \ldots, X_n$ be a sequence of uncorrelated random variables with bounded variance. Denoting
$$S_n = X_1 + \cdots + X_n,$$
we obtain:
$$\frac{S_n - E(S_n)}{n} \xrightarrow{P} 0.$$

Proof (you won't believe it, using Chebyshev's inequality)
Recall the boundedness of variances:
$$Var\left(\sum_{i=1}^{n} X_i\right) = \langle\text{uncorrelatedness}\rangle = \sum_{i=1}^{n} Var(X_i) \leq \max\{Var(X_i)\} \cdot n =$$
$$= cn \text{ for some } c.$$

Then

$$Var\left(\frac{S_n}{n}\right) = \frac{1}{n^2}Var(S_n) \leq \frac{c}{n}.$$

Now

$$P\left(\left|\frac{S_n - E(S_n)}{n}\right| \geq \varepsilon\right) = P\left(\left|\frac{S_n}{n} - \frac{E(S_n)}{n}\right| \geq \varepsilon\right) =$$
$$= \langle\text{Chebyshev's inequality}\rangle \leq \frac{Var\left(\frac{S_n}{n}\right)}{\varepsilon^2} \leq \frac{c}{n\varepsilon^2} \to 0 \text{ as } n \to \infty.$$

The Law of Large Numbers (yes, weak, but so fundamental!) is proven.

2.3 Convergence in L_p

Difficult to understand, even if you sit and cry for a couple of hours over functional analysis textbooks, I can't guarantee it will help.
Let me try to explain what L_p has to do with it. Assuming the reader has been diligently inoculated in the preface, I will say that L_p is the space of measurable functions from a measure space $(X, \mathcal{F}, \mu)$, between which we can define a "distance"in the following way:

$$d_p(f, g) = \left(\int |f(x) - g(x)|^p \, dx\right)^{\frac{1}{p}}.$$

Here, f and g are functions from the L_p space, and p is the power to which these functions are integrable. The same p appears in the name: L_1, L_2, L_p. The L in the name is in honor of Lebesgue, whose integral is used in the metric.
In particular, if the space is probabilistic, the integral is taken with respect to the measure $\mathbb{P}$:

$$d_p(f, g) = \left(\int_\Omega |f(\omega) - g(\omega)|^p \, d\mathbb{P}(\omega)\right)^{\frac{1}{p}}.$$

In a kinder and more familiar form for most readers:

$$d_p(f_n, f) = E(|X_n - X|^p).$$

Strict Definition:
Convergence of a sequence of random variables $X_1, \ldots, X_n$ to a variable X occurs when

$$E(|X_n - X|^p) \to 0, \text{as } n \to \infty.$$

In the case where $p = 1$, the convergence is called convergence in mean, and when $p = 2$, it is called convergence in mean square.

Let us now show that convergence in mean implies convergence in probability: Using Markov's inequality:

$$P(|X_n - X|^p \geq \varepsilon^p) \leq \frac{E(|X_n - X|^p)}{\varepsilon^p} \to 0 \text{ as } n \to \infty.$$

Convergence in probability is proven. (The fact that convergence in L_p implies convergence in $L_{p'}$, where $p > p'$, is left for the reader to explore. If the reader is not interested in the full proof, they can look up something about Hölder's inequality.)

As for applications, convergence in L_p is a beast from functional analysis, so its applications are not limited by probability theory (for example, Empirical Risk Minimization in reinforcement learning checks the convergence of empirical and real risk to confirm the stability of the model). However, L_p-convergence is a strong form of convergence and can be used, for instance, in the Law of Large Numbers:

Law of Large Numbers (Strong) in L_p-formulation
If a sequence of independent and identically distributed random variables $X_1, \ldots, X_n$ satisfies the condition $E(|X_i|^p) < \infty$, then the sample mean $S_n = \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow{L_p} E(X)$.

2.4 Almost Sure Convergence

Alongside L_p, it is considered a strong form of convergence. Intuitively, if we take convergence in probability, it states that the probability of a single deviation of X_n from X by more than ε tends to zero, whereas almost sure convergence states that we can find an index i such that the maximum deviation among all subsequent $X_{j>i}$ exceeds ε with probability tending to zero.

Formal Definition:
A sequence of random variables X_n converges to a random variable X almost surely if

$$P\left(\underset{k>n}{sup} |X_k - X| > \varepsilon\right) \to 0 \text{ as } n \to \infty.$$

We write $X_n \xrightarrow{a.s.} X$.
From this definition, it is clear that almost sure convergence implies convergence

in probability (because $sup_{k>n}|X_k - X| \geq |X_n - X|$).

In practice, almost sure convergence is used, for example, in the Strong Laws of Large Numbers and Lebesgue's Dominated Convergence Theorem.

Why do L_p and almost sure convergence not imply each other? Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, P)$, where

$$X_n = \begin{cases} n^{-1/p}, & \text{with probability } n^{-1}, \\ 0, & \text{otherwise.} \end{cases}$$

Step 1: Checking Convergence in L^p
Consider convergence in L^p to the zero function. Compute the L^p-norm of X_n:

$$\mathbb{E}[|X_n - 0|^p] = \mathbb{E}[|X_n|^p] = (n^{-1/p})^p \cdot n^{-1} + 0 = n^{-1}n^{-1} = n^{-2}.$$

Since the series $\sum_{n=1}^{\infty} n^{-2}$ converges, it follows that

$$\mathbb{E}[|X_n - 0|^p] \to 0 \quad \text{as } n \to \infty.$$

Thus, $X_n \to 0$ in L^p.

Step 2: Checking Almost Sure Convergence
For almost sure convergence, consider:

$$\sum_{n=1}^{\infty} \mathbb{P}(X_n \neq 0) = \sum_{n=1}^{\infty} n^{-1}.$$

This series is harmonic and diverges. By the Borel-Cantelli lemma, infinitely many X_n are nonzero with probability 1, so X_n does *not* converge to zero almost surely.

Conversely
Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables defined as:

$$X_n = n^{1/p}\mathbf{1}_{\{U\leq 1/n\}},$$

where $U \sim \text{Uniform}(0, 1)$ is a random variable uniformly distributed on $[0, 1]$.

Step 1: Checking Almost Sure Convergence

Since the indicator function $\mathbf{1}_{\{U \leq 1/n\}}$ tends to zero almost surely as $n \to \infty$, it follows that

$$X_n \to 0 \quad \text{almost surely.}$$

Step 2: Checking Convergence in L^p

Compute the L^p-norm:

$$\mathbb{E}[|X_n|^p] = \mathbb{E}[(n^{1/p}\mathbf{1}_{\{U \leq 1/n\}})^p] = n \cdot n^{-p/p} = n \cdot n^{-1} = 1.$$

Since $\mathbb{E}[|X_n|^p] \not\to 0$, the sequence does not converge in L^p.

3 Myths and Legends

There exists Riesz's theorem: From every sequence of random variables converging in probability, one can extract a subsequence that converges almost surely.

Under certain conditions, convergence in L_p can imply almost sure convergence, and vice versa.

If $f_n \xrightarrow{L_p} f$ and $f_n \xrightarrow{L_p} g$, then $f = g$ with probability 1.