CSE 840: Computational Foundations of Artificial Intelligence April 7, 2025

Convergence of Random Variables

Instructor: Vishnu Boddeti *Scribe: Alex Logan and Tyler Matson*

1 Motivation and Practical Context

In machine learning, we often try to estimate things we don't know, like how accurate a model or biased a model is, or properties of a distribution, by taking random samples. To do this, we use sequences of estimators (random variables):

- Do estimators converge to the true value X?
- If so, how and in what way?

Understanding convergence helps validate empirical risk minimization, approximate inference, and MLE consistency.

2 Formal Setup

Let $(\Omega, \mathcal{A}, P)$ be a probability space. Suppose we have a sequence of random variables:

$$X_n : \Omega \to \mathbb{R}$$

and we're interested in whether X_n converges to some limiting random variable X, and if so, in what sense. We can define and compare several different types of convergence.

3 Types of Convergence: Definitions and Intuition

3.1 1. Sure (Pointwise) Convergence

Definition 1 *We say that X_n converges surely (or pointwise) to X if*

$$\forall \omega \in \Omega, \quad \lim_{n \to \infty} X_n(\omega) = X(\omega)$$

This type of convergence is fully deterministic. The sequence must converge for every outcome (ω). There's no probability involved.

Example: Let $X_n(\omega) = \omega + \frac{1}{n} \to X(\omega) = \omega$.

In this case, the convergence happens individually at each $\omega \in \Omega$

3.2 2. Almost Sure (a.s.) Convergence

Definition 2

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

Note: This type of convergence can fail on a set of outcomes with probability zero (a measure-zero set).

Machine Learning Example: In the Strong Law of Large Numbers, the empirical average of a sample converges almost surely to the expected value, meaning it holds with probability 1, even though it may fail on rare edge cases.

Note: In online learning or continual learning systems, almost sure convergence ensures that model parameters converge to a stable value across data streams, except on a set of zero probability. This is crucial when proving consistency of algorithms under stochastic gradient descent with diminishing learning rates.

3.3 3. Convergence in Probability

We say that X_n converges to X in probability if:

$$\forall \varepsilon > 0, \quad \lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$$

As n increases, the probability that X_n is far from X becomes negligible.

A sequence might not converge pointwise or almost surely, but it can still converge in probability, making this a more flexible and commonly used notion.

AI Example: In supervised learning, we often use empirical risk (the average loss on training data) as a stand-in for the true expected loss. As we get more data, the empirical risk gets closer to the expected loss, not exactly every time, but with high probability. This kind of convergence in probability is what helps justify why empirical risk minimization actually works in practice.

3.4 4. Convergence in L^p

We say that X_n converges to X in L^p if:

$$X_n \to X \text{ in } L^p \quad \iff \quad \mathbb{E}[|X_n - X|^p] \to 0$$

The expected p-th power of the error between X_n and X goes to zero. This is a stronger condition than just convergence in probability.

L^p convergence is especially useful when we care about the expected size of the error, for example, $\mathbb{E}[|X_n - X|]$ is often used in risk estimation and evaluating the performance of learning algorithms.

3.5 5. Convergence in Distribution (Weak Convergence)

Definition 3 *$X_n \xrightarrow{d} X$ if for all bounded continuous $f : \mathbb{R} \to \mathbb{R}$,*

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

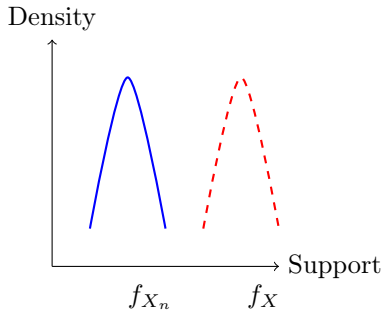
We say that X_n converges in distribution to X, written $X_n \xrightarrow{d} X$, if for every bounded and continuous function $f : \mathbb{R} \to \mathbb{R}$,

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

The distributions of X_n start to "look like" the distribution of X, even if the values of X_n don't get close to X pointwise.

Example: In the Central Limit Theorem (CLT), the normalized sum of independent random variables converges in distribution to a Gaussian, even though it might not converge almost surely or in probability.

Diagram:



4 Hierarchy of Implications

$$\text{Sure} \Rightarrow \text{Almost Sure} \Rightarrow L^p \Rightarrow \text{Probability} \Rightarrow \text{Distribution}$$

None of the reverse implications hold in general.

5 Worked Examples: In-Depth

Example 1 (a.s. + probability, not L^1)

$X_n = n \cdot \chi_{[0,1/n]}$

Area: $n \cdot \frac{1}{n} = 1$ always $\to$ no L^1 convergence.

Example 2 (prob. + L^1, not a.s.)

Sliding block support over disjoint intervals $\to$ convergence fails at every ω, yet error vanishes in expectation and probability.

Example 3 (in distribution only)

$X_n = \chi_{[0,1/2]}, X = \chi_{[1/2,1]}$ distributions are same (Bernoulli), but pointwise/probability convergence fails.

6 Measurability of Convergence Events

We write:

$$\left\{\omega \mid |X_n(\omega) - X(\omega)| < \frac{1}{k}\right\} \in \mathcal{A}$$

Because X_n, X are measurable, these sets are too.

This validates our convergence definitions via events.

7 The Borel–Cantelli Lemma

Definition 4 *An event A_n occurs **infinitely often (i.o.)** if $\omega \in A_n$ for infinitely many n.*

Theorem 5 (Borel–Cantelli)

- *If $\sum P(A_n) < \infty$ then $P(A_n \text{ i.o.}) = 0$*
- *If A_n are independent and $\sum P(A_n) = \infty$, then $P(A_n \text{ i.o.}) = 1$*

Usage: Control tail behavior of convergence events.

8 Application to ML Theory

Suppose $\varepsilon_n = 1/n$ and

$$P(|X_n - X| > \varepsilon_n) \leq \delta_n, \quad \sum \delta_n < \infty$$

Then:

$$P(|X_n - X| > \varepsilon_n \text{ i.o.}) = 0 \Rightarrow X_n \to X \text{ a.s.}$$

Conclusion: Estimator converges strongly, which is crucial in generalization theory.

(Source: `https://en.wikipedia.org/wiki/Empirical_risk_minimization`)