

Lecture Notes 4

36-705

In today's lecture we discuss the convergence of random variables. At a high-level, our first few lectures focused on non-asymptotic properties of averages i.e. the tail bounds we derived applied for any fixed sample size n . For the next few lectures we focus on asymptotic properties, i.e. we ask the question: what happens to the average of n i.i.d. random variables as $n \rightarrow \infty$.

Roughly, from a theoretical perspective the idea is that many expressions will considerably simplify in the asymptotic regime. Rather than have many different tail bounds, we will derive simple "universal results" that hold under extremely weak conditions. From a slightly more practical perspective, asymptotic theory is often useful to obtain approximate confidence intervals.

1 Reminder: convergence of sequences

When we think of convergence of deterministic real numbers the corresponding notions are classical.

Formally, we say that a sequence of real numbers a_1, a_2, \dots converges to a fixed real number a if, for every positive number ϵ , there exists a natural number $N(\epsilon)$ such that for all $n \geq N(\epsilon)$, $|a_n - a| < \epsilon$. We call a the limit of the sequence and write $\lim_{n \rightarrow \infty} a_n = a$.

Our focus today will be in trying to develop analogues of this notion that apply to sequences of random variables. We will first give some definitions and then try to circle back to relate the definitions and discuss some examples.

Throughout, we will focus on the setting where we have a sequence of random variables X_1, \dots, X_n and another random variable X , and would like to define what it means for the sequence to converge to X . In each case, to simplify things you should also think about the case when X is deterministic, i.e. when $X = c$ with probability 1 (for some constant c).

Importantly, we will *not assume that the RVs X_1, \dots, X_n are independent.*

2 Almost sure convergence

We will not use almost sure convergence in this course so you should feel free to ignore this section. A natural analogue of the usual convergence would be to hope that,

$$\lim_{n \rightarrow \infty} X_n = X.$$

We write $X_n \xrightarrow{a.s.} X$. These are both however random variables so one has to at least specify on what event we are hoping for this statement to be true.

The correct analogue turns out to be to require:

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

There are measure theoretic subtleties to be aware of here. In particular, the sample space inside the probability statement here is set of infinite sequences and it requires some machinery to be precise here.

There are other equivalent (this is somewhat difficult to see) ways to define almost sure convergence. Equivalently, we say that X_n converges almost surely to X if we let Ω be a set of probability mass 1, i.e. $\mathbb{P}(\Omega) = 1$, and for every $\omega \in \Omega$, and for every $\epsilon > 0$, we have that there is some $n \geq N(\omega, \epsilon)$ such that:

$$|X_n(\omega) - X(\omega)| \leq \epsilon.$$

Roughly, the way to think about this type of convergence is to imagine that there is some set of exceptional events on which the random variables can disagree, but these exceptional events have probability 0 as $n \rightarrow \infty$. Barring, these exceptional events the sequence converges just like sequences of real numbers do. The exceptional events is where the “almost” in almost sure arises.

3 Convergence in probability

A sequence of random variables X_1, \dots, X_n converges in probability to a random variable X if for every $\epsilon > 0$ we have that,

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

We write $X_n \xrightarrow{P} X$. To build intuition it is perhaps useful to consider the case when X is deterministic, i.e. $X = c$ with probability 1. Then convergence in probability is saying that as n gets large the distribution of X_n gets more peaked around the value c . Convergence in probability can be viewed as a statement about the convergence of probabilities, while almost sure convergence is a convergence of the values of a sequence of random variables.

We will not prove this statement but convergence in probability is implied by almost sure convergence. The notes contain a counterexample to the reverse implication which we may or may not cover in the lecture.

Weak Law of Large Numbers Suppose that Y_1, Y_2, \dots are i.i.d. with $\mathbb{E}[Y_i] = \mu$ and $\text{Var}(Y_i) = \sigma^2 < \infty$. Define,

$$X_n = \frac{1}{n} \sum_{j=1}^n Y_j.$$

The WLLN says that the sequence X_1, X_2, \dots converges in probability to μ . That is $X_n \xrightarrow{P} \mu$.

Proof: The proof is simply an application of Chebyshev's inequality. We note that by Chebyshev's inequality:

$$\mathbb{P}(|X_n - \mathbb{E}[X]| \geq \epsilon) \leq \frac{\sigma^2}{n\epsilon^2}.$$

This in turn implies that,

$$\lim_{n \rightarrow \infty} \mathbb{P}(|X_n - \mathbb{E}[X]| \geq \epsilon) = 0,$$

as desired.

Notes:

1. Strictly speaking the WLLN is true even without the assumption of finite variance, as long as the first absolute moment is finite. This proof is a bit more difficult.
2. There is a statement that says that under similar assumptions the average converges almost surely to the expectation. This is known as the strong law of large numbers. This is actually quite a bit more difficult to prove.

Consistency: Convergence in probability will frequently recur in this course. Usually we will construct an estimator $\hat{\theta}_n$ for some quantity θ^* . We will then say that the estimator is *consistent* if the sequence of RVs $\hat{\theta}_n$ converges in probability to θ^* .

The WLLN/Chebyshev can already be used to prove some rudimentary consistency guarantees. For instance, if we consider the sample variance:

$$\hat{S}_n = \frac{1}{n-1} \sum_{i=1}^n (X_i - \hat{\mu}_n)^2,$$

then by Chebyshev's inequality we obtain,

$$\mathbb{P}(|\hat{S}_n - \sigma^2| \geq \epsilon) \leq \frac{\text{Var}(\hat{S}_n)}{\epsilon^2},$$

so a sufficient condition for consistency is that $\text{Var}(\hat{S}_n) \rightarrow 0$ as $n \rightarrow \infty$.

Convergence in probability does not imply almost sure convergence: Suppose we have a sample space $S = [0, 1]$, with the uniform distribution, we draw $s \sim U[0, 1]$ and define $X(s) = s$. We define the sequence as:

$$\begin{aligned} X_1(s) &= s + \mathbb{I}_{[0,1]}(s), & X_2(s) &= s + \mathbb{I}_{[0,1/2]}(s), & X_3(s) &= s + \mathbb{I}_{[1/2,1]}(s) \\ X_4(s) &= s + \mathbb{I}_{[0,1/3]}(s), & X_5(s) &= s + \mathbb{I}_{[1/3,2/3]}(s), & X_6(s) &= s + \mathbb{I}_{[2/3,1]}(s). \end{aligned}$$

Now one can check that this sequence converges in probability but not almost surely. Roughly, the “ $1 + s$ ” spike becomes less frequent down the sequence (allowing convergence in probability) but the limit is not well defined. For any s , $X_n(s)$ alternates between s and $1 + s$.

4 Convergence in quadratic mean

An often useful way to show convergence in probability is to show something stronger known as convergence in quadratic mean. We say that a sequence converges to X in quadratic mean if:

$$\mathbb{E}(X_n - X)^2 \rightarrow 0,$$

as $n \rightarrow \infty$. We write $X_n \xrightarrow{qm} X$.

5 Convergence in distribution

The other commonly encountered mode of convergence is convergence in distribution. We say that a sequence converges to X in distribution if:

$$\lim_{n \rightarrow \infty} F_{X_n}(t) = F_X(t),$$

for all points t where the CDF F_X is continuous. We will see why the exception matters in a little while but for now it is worth noting that convergence in distribution is the weakest form of convergence. We write $X_n \rightsquigarrow X$.

For instance, a sequence of i.i.d. $N(0, 1)$ RVs converge in distribution to an independent $N(0, 1)$ RV, even though the values of the random variables are not close in any meaningful sense (their distributions are however, identical). A famous result that we will discuss in the next lecture is the central limit theorem. The central limit theorem says that an average of i.i.d. random variables (appropriately normalized) converges in distribution to a $N(0, 1)$ random variable.

The picture to keep in mind to understand the relationships is the following one:

$$\begin{array}{c} \text{q.m.} \\ \downarrow \\ \text{a.s.} \rightarrow \text{prob} \rightarrow \text{distribution} \end{array}$$

We will re-visit this in the next lecture and perhaps try to prove some of the implications (or disprove some of the non-implications).

6 Examples

Example 1: Suppose we consider a sequence $X_n = N(0, 1/n)$. Intuitively, it seems like this sequence converges to 0. Let us first consider what happens in distribution.

Let X be such that $P(X = 0) = 1$. The CDF is $F_X(x) = 0$, for $x < 0$ and $F_X(x) = 1$ for $x \geq 0$. Note that $X_n \stackrel{d}{=} Z/n$ where $Z \sim N(0, 1)$. So,

$$F_{X_n}(x) = \mathbb{P}(X_n \leq x) = \mathbb{P}(Z \leq \sqrt{n}x),$$

where $Z \sim N(0, 1)$. If $x > 0$ this tends to 1, and if $x < 0$ this tends to 0. Interestingly, at $x = 0$, $F_{X_n}(x) = 1/2$, and does not converge to $F_X(0) = 1$. Remember, however, that we had an exception at points of discontinuity. So $X_n \rightsquigarrow X$.

Example 2: Let us consider the same example and consider convergence in probability.

$$\mathbb{P}(|X_n - X| \geq \epsilon) = \frac{\mathbb{E}[X_n^2]}{\epsilon^2} = \frac{1}{n\epsilon^2} \rightarrow 0,$$

so the sequence converges to 0 in probability.

Example 3: Suppose $X_1, \dots \sim U[0, 1]$. Let us define $X_{(n)} = \max_{1 \leq i \leq n} X_i$. Now, we verify two things:

1. $X_{(n)}$ converges in probability to 1. To see this observe that,

$$\begin{aligned} \mathbb{P}(|X_{(n)} - 1| \geq \epsilon) &= \mathbb{P}(X_{(n)} \leq 1 - \epsilon) \\ &= \prod_{i=1}^n \mathbb{P}(X_i \leq 1 - \epsilon) = (1 - \epsilon)^n \\ &\rightarrow 0. \end{aligned}$$

2. The random variable $n(1 - X_{(n)})$ converges in distribution to an $\text{Exp}(1)$ RV. To see this we compute:

$$\begin{aligned} F_{X_{(n)}}(t) &= \mathbb{P}(n(1 - X_{(n)}) \leq t) = 1 - \mathbb{P}(X_{(n)} \leq 1 - t/n) \\ &= 1 - (1 - t/n)^n \rightarrow 1 - \exp(-t) = F_X(t). \end{aligned}$$