36-700: Probability and Mathematical Statistics **Spring 2019**

Lecture 6: Convergence of Random Variables, Large Sample Theory

Lecturer: Jing Lei

6.1 Review and Outline

Last class we saw:

- Markov's inequality and Chebyshev's inequality
- Gaussian tail inequality
- Exponential concentration

This lecture we will introduce concepts to facilitate the study of limiting behaviors of random variables.

6.2 Motivation: The Weak Law of Large Numbers

The weak law of large numbers essentially assures us that the average of independent and identically distributed random variables "converges" to the expectation.

We will assume that $X_1, \ldots, X_n$ are i.i.d with $\mathrm{Var}(X) < \infty$. For any $\epsilon > 0$, by Chebyshev's inequality,

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i - \mathbb{E}(X)\right| \geq \epsilon\right) \leq \frac{\mathrm{Var}(X_1)}{n\epsilon^2} \to 0,$$

as $n \to \infty$.

This is a version of the famous **Weak Law of Large Numbers**. This type of convergence is quite common and has a name "convergence in probability". So the weak LLN says that the sample average converges to the population average in probability. It is called the "weak law" because there is a stronger version that uses a different type of convergence. We will discuss convergence of random variables more systematically.

6.3 Introduction to Stochastic Convergence

The weak LLN is an example where we tried to reason about the limiting behaviour of a sequence of random variables:

$$Y_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

In statistics we commonly estimate parameters using data, and in this case our estimate forms a sequence of random variables (as we collect more data). We would like to reason about the limiting behaviour of our estimates (do they "converge" to the truth, what is their distribution around the truth and so on). This is what we refer to as **large sample theory.**

Suppose we have a sequence of random variables $X_1, \ldots, X_n$, and another random variable X. Let F_n denote the CDF of X_n, and let F be the CDF of X.

The two most basic forms of stochastic convergence are:

1. Convergence in Probability: We say the sequence converges to X if for every $\epsilon > 0$,

   $$\mathbb{P}(|X_n - X| \geq \epsilon) \to 0,$$

   as $n \to \infty$, denoted as

   $$X_n \xrightarrow{P} X\,.$$

   An important example is the weak law.

2. Convergence in Distribution: The sequence converges in distribution to X if

   $$\lim_{n\to\infty} F_n(t) = F(t),$$

   at all points t where F is continuous.

   Convergence in distribution is indeed convergence of the distribution functions rather than random variables themselves. It is commonly written as

   $$F_n \rightsquigarrow F, \quad \text{or} \quad X_n \rightsquigarrow X\,.$$

   An important example of this type of convergence is the central limit theorem. We will see this in much more detail but roughly the central limit theorem says that the average of i.i.d. RVs, rescaled appropriately converges in distribution to a standard normal distribution, i.e. that

   $$\frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma} \rightsquigarrow N(0,1).$$

   If you think about it, this is a really stunning result. You do not assume anything about the RVs except their first two moments need to exist, and you conclude that the average approaches a Gaussian distribution.

Example 1: Let $X_n \sim N(0, n^{-1})$ for $n = 1, 2, \ldots$. Then the distribution of X_n concentrates more around 0 as n increases. In this case X_n converges to 0 both in probability and in distribution, but $\mathbb{P}(X_n = 0) = 0$ for all n.

Exercise: Prove the above claim.

Example 2: Suppose X_n for $n = 1, 2, \ldots$, are i.i.d $N(0, 1)$ random variables and $X \sim N(0, 1)$. In this case, it is easy to see that there is no convergence in probability and the right notion is convergence in distribution.

While convergence in distribution only makes a statement about the distribution, convergence in probability is a statement about the value of the random variables. It is not too difficult to show that convergence in probability $\implies$ convergence in distribution (see Wasserman's book).

6.4 Other notions of convergence

There are stronger notions than convergence in probability. For instance:

1. Convergence in quadratic mean (convergence in ℓ_2): We say that a sequence X_n converges to X in quadratic mean (or in ℓ_2) if:

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

as $n \to \infty$, denoted as

$$X_n \xrightarrow{\ell_2} X, \quad \text{or} \quad X_n \xrightarrow{\text{qm}} X\,.$$

This is once again a convergence of values of a sequence of random variables. In fact, convergence in quadratic mean $\implies$ convergence in probability since by Markov's inequality we know that:

$$\mathbb{P}(|X_n - X| \geq \epsilon) \leq \frac{\mathbb{E}(X_n - X)^2}{\epsilon^2} \to 0,$$

as $n \to \infty$.

2. Convergence in ℓ_1: We say that a sequence X_n converges to X in ℓ_1 if:

$$\mathbb{E}|X_n - X| \to 0,$$

as $n \to \infty$, denoted as

$$X_n \xrightarrow{\ell_1} X_n\,.$$

Exercise. Prove that convergence in quadratic mean $\implies$ convergence in ℓ_1.

3. Almost-sure convergence: Almost-sure convergence requires that:

   $$\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1\,,$$

   and denoted as

   $$X_n \overset{a.s.}{\to} X\,.$$

   Under the same conditions as the WLLN, we actually have that the sample mean converges to the population mean almost surely. This is actually substantially harder to prove, but it is worthwhile trying to interpret it. In the LLN setting, we define

   $$X = \mathbb{E}[Y] \quad \text{and} \quad X_n = \frac{1}{n}\sum_{i=1}^{n} Y_i.$$

   Convergence in probability roughly tells us that after a point, most of the values X_n are quite close to X. We can still have "failures" where we obtain a sudden erratic average but these are rare (and increasingly rare further down the sequence). Almost sure convergence says that after a point there are no more failures, and every random variable of the sequence X_n is close to X. At least roughly, convergence in probability allows for a few erratic sample averages, as long as they are not too likely, while almost-sure convergence does not.

Exercise: Prove that almost sure convergence implies convergence in probability.

Exercise: Find examples of X_n and X to show the following.

1. Convergence in probability does not necessarily imply almost sure convergence.
2. Convergence in probability does not necessarily imply convergence in ℓ_1.
3. Convergence in ℓ_1 and almost sure convergence do not imply each other.

Here we state the strongest version of the LLN.

Kolmogorov's Strong Law of Large Numbers. Let X_1, ..., X_n, ... be iid random variables with finite expectation $\mathbb{E}(X_1) = \mu$. Then

$$\frac{1}{n}\sum_{i=1}^{n} X_i \overset{a.s.}{\to} \mu\,.$$

6.5 Central Limit Theorem

Let us begin by stating the Central Limit Theorem:

Let $X_1, \ldots, X_n$ be i.i.d. with mean μ and variance σ^2, both finite. Define:

$$\hat{\mu}_n = \frac{1}{n}\sum_{i=1}^{n} X_i.$$

Then

$$Z_n = \frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma} \rightsquigarrow N(0,1)\,.$$

1. Perhaps the first thing to verify is that at least the mean and variance are correct. Concretely, we can see that:

   - $\mathbb{E}[Z_n] = 0$.
   - $\mathbb{E}[Z_n^2] = 1$.

   This reveals the need for the standardization: i.e. subtracting μ, multiplying by $\sqrt{n}$ and dividing by σ.

2. There are two common ways to interpret the Central Limit Theorem (and more broadly convergence in distribution):

   - The first is to suppose I repeated the experiment, i.e. I draw many sequences: $\{X_1^1, \ldots, X_n^1\}$, $\{X_1^2, \ldots, X_n^2\}$, $\ldots$, $\{X_1^k, \ldots, X_n^k\}$, and computed their centered and normalized averages $Z_n^1, \ldots, Z_n^k$.

     The central limit theorem then tells us that these normalized averages will (approximately) have a standard Gaussian distribution. For instance, you could imagine computing a histogram of the averages, and this will follow a Gaussian law.
   - The second way is to suppose that before I did the experiment I asked the question what is the probability of a certain outcome:

     $$\mathbb{P}(a \leq Z_n \leq b).$$

     Then the central limit theorem tells us that:

     $$\mathbb{P}(a \leq Z_n \leq b) \approx \Phi(b) - \Phi(a).$$

     This is quite useful in statistics as it precisely quantifies the behavior of $\hat{\mu} - \mu$.

3. Some further topics:

   - Rate of convergence: If we define the distance between the CDF of the average, and the CDF of a Gaussian appropriately, we can ask how far the two CDFs are for a $f$inite sample size n.

     These results are typically called Berry-Esseen bounds. They assure us that the convergence to normality can happen quite quickly in some important cases.

- Multivariate CLT: If we average a collection of independent random vectors then they will converge in distribution to a multivariate Gaussian.
- Delta method: Given that Y_n converges in distribution to a Gaussian, one can ask about functions of Y_n. Under some regularity conditions these also converge to a Gaussian, and the delta method tells us how to compute the mean and variance of the new Gaussian.