36-700: Probability and Mathematical Statistics I **Fall 2016**

Lecture 7: September 14

Lecturer: Siva Balakrishnan

7.1 Review and Outline

Last class we saw:

- Hoeffding's inequality (exponential concentration for bounded RVs)
- Weak LLN
- Convergence of random variables: in probability and in distribution.

This class we will first talk a bit about moments, go over a few more convergence concepts and then discuss the central limit theorem. See Chapter 5 of the Wasserman book. The material on moments is developed across many chapters.

7.2 Moments re-visited

First let us recall the definitions, the k^{th} moment of a RV is:

$$M_k = \mathbb{E}X^k,$$

the k^{th} central moment of a RV is:

$$C_k = \mathbb{E}(X - \mathbb{E}(X))^k,$$

and the MGF is:

$$M(t) = \mathbb{E}\exp(tX).$$

Moments will recur throughout the course but lets preview some ideas:

1. Moments are summaries of the distribution: Particularly, we have extensively used the mean and variance. At a very high-level the central moments give you some idea about how the probability mass is distributed around the mean. A typical way to try

to interpret higher moments is by comparing them to the corresponding moments of a Gaussian.

For instance, the third central moment of a Gaussian is zero, and in general for any symmetric distribution the third central moment is zero. So we call the third moment the *skewness*, if it is non-zero it means the distribution is skewed, and the sign tells you to which side of the mean it is skewed.

Similarly, the fourth central moment of a standard Gaussian is 3. The fourth central moment is called the kurtosis of a distribution. Distributions with fourth central moment larger than 3 are called leptokurtic (they are more peaked than a Gaussian near the mean), smaller than 3 are platykurtic (they are flatter than a Gaussian).

Broadly, the central moments are just summaries of a distribution (i.e. numbers associated with a distribution). One way to assign meaning to them is by comparing them to the corresponding summaries of a Gaussian.

2. Moments can be used to reason about convergence: We will not see much of this in our course, but a typical way to prove the central limit theorem (for instance) is by showing that all moments of

$$\frac{\sqrt{n}(\hat{\mu} - \mu)}{\sigma}$$

(which are now just sequences of numbers) converge to the corresponding moments of a standard normal. We then appeal to the fact that: convergence of moments + some regularity conditions $\implies$ convergence in distribution.

Conceptually, this "reduces" showing convergence in distribution to show that sequences of numbers (i.e. moments) converge.

3. Moments give estimators: We will again see this later on, but as a simple example:

Suppose $X_1, \ldots, X_n$ are drawn i.i.d. from a Gaussian distribution $N(\mu, \sigma^2)$, and I want to estimate the parameters of the Gaussian. A simple (and very general strategy) is to match moments. We can calculate that:

$$\mathbb{E}(X) = \mu, \quad \text{and} \quad \mathbb{E}(X^2) = \sigma^2 + \mu^2.$$

The sample moments are:

$$\hat{\mathbb{E}}(X) = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad \text{and} \quad \hat{\mathbb{E}}(X^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2,$$

so that we could try to solve the equations:

$$\hat{\mathbb{E}}(X) = \mu, \quad \text{and} \quad \hat{\mathbb{E}}(X^2) = \sigma^2 + \mu^2,$$

to obtain estimates of μ and σ. This is one of the general ways to come up with estimators for parameters, i.e. you "match" moments of the sample (which are easy to estimate) with moments of the unknown distribution.

4. Moments give tail bounds: We have already seen Chebyshev's inequality.

$$\mathbb{P}(|X - \mu| \geq t\sigma) \leq \frac{1}{t^2}.$$

To derive Chebyshev we just squared things and used Markov's inequality. We could instead have taken higher powers to arrive at:

$$\mathbb{P}(|X - \mu| \geq t) = \mathbb{P}(|X - \mu|^k \geq t^k) \leq \frac{\mathbb{E}|X - \mu|^k}{t^k}.$$

So if our distribution has higher central moments (that are small) then we will typically get much tighter control by choosing a higher value for k in the expression above.

A refinement of this technique that is much more commonly useful (in particular, to show exponential concentration) is called the Chernoff technique. Suppose we first center our random variable (by subtracting its mean). We then observe that,

$$\mathbb{P}(X \geq t) = \mathbb{P}(\exp(uX) \geq \exp(ut)) \leq \frac{\mathbb{E}\exp(uX)}{\exp(ut)}.$$

In the above expression $\mathbb{E}\exp(uX)$ is just the MGF of X and u is a free parameter that we can try to choose to make the bound small (tight).

Illustration of Chernoff technique: For a mean 0 Gaussian RV we can calculate the MGF:

$$\mathbb{E}(\exp(uX)) = \exp(\sigma^2 u^2/2),$$

so the Chernoff method gives us that:

$$\mathbb{P}(X \geq t) \leq \frac{\exp(\sigma^2 u^2/2)}{\exp(ut)} = \exp(\sigma^2 u^2/2 - ut).$$

The right hand side is minimized when I choose $u = t/\sigma^2$, and this gives us the bound:

$$\mathbb{P}(X \geq t) \leq \exp(-t^2/(2\sigma^2)).$$

This is a much simpler way to show that Gaussians have exponential tails. I will also add an important remark:

A simple conclusion of the above exercise is that if the MGF of a RV was dominated by the MGF of a Gaussian, then the exact same bound would hold, i.e., if some (not necessarily Gaussian) RV Y had an MGF that satisfied:

$$\mathbb{E}(\exp(uY)) \leq \exp(\sigma^2 u^2/2),$$

for some value σ, then it would also satisfy the tail bound:

$$\mathbb{P}(Y \geq t) \leq \exp(-t^2/(2\sigma^2)).$$

It turns out that a very large family of RVs actually satisfy the above condition: they are called sub-Gaussian RVs and include things like bounded RVs.

At a higher-level, the MGF can be used the understand the tails of a RV and importantly if the MGF is smaller then the random variable has lighter tails. So once again, moments help us understand tails of a distribution.

7.3 Stochastic convergence

Let us begin with some examples that we looked at informally last lecture:

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. We want to say that X_n converges to 0 but $\mathbb{P}(X_n = 0) = 0$ for all n.

This is an example where the right notion of convergence is that of convergence in probability. Defining, X to be the (non-)random variable that is always 0 we observe that

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

as $n \to \infty$. So the sequence converges in probability.

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.

Concretely, convergence in distribution is much weaker: it only makes a statement about the distribution, whereas convergence in probability is a statement about the value of the random variable. It is not too difficult to show that convergence in probability $\implies$ convergence in distribution (see Wasserman's book).

I will also point out that there are stronger notions than convergence in probability. For instance:

1. Convergence in quadratic mean: We say that a sequence X_n converges to X in quadratic mean if:

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

   as $n \to \infty$. 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 Chebyshev'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 $n \to \infty$. Convergence in quadratic mean $\implies$ convergence in ℓ_1. To prove this we can just use the Cauchy-Schwarz inequality:

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

as $n \to \infty$.

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

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

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.

7.4 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. Define:

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

and

$$Z_n = \frac{\sqrt{n}(\hat{\mu}_n - \mu)}{\sigma},$$

then Z_n converges in distribution to a standard Gaussian.

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 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:

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

3. I will also point out a few extensions that we will cover in detail if we need them:

   - 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 *finite* 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.