MASSACHUSETTS INSTITUTE OF TECHNOLOGY

6.436J/15.085J Fall 2018
Lecture 17

LAWS OF LARGE NUMBERS AND CENTRAL LIMIT THEOREM

Contents

1 USEFUL INEQUALITIES

Markov inequality: If X is a nonnegative random variable, then $\mathbb{P}(X \geq a) \leq \mathbb{E}[X]/a$.

Proof: Let I be the indicator function of the event $\{X \geq a\}$. Then, $aI \leq X$. Taking expectations of both sides, we obtain the claimed result. □

Chebyshev inequality: $\mathbb{P}\big(|X - \mathbb{E}[X]| \geq \epsilon\big) \leq \mathrm{var}(X)/\epsilon^2$.

Proof: Apply the Markov inequality, to the random variable $|X - \mathbb{E}[X]|^2$, and with $a = \epsilon^2$. □

2 CONVERGENCE IN DISTRIBUTION vs CHARACTERISTIC FUNCTIONS

We know that equality of two characteristic functions implies equality of the corresponding distributions. It is then plausible to hope that "near-equality" of characteristic functions implies "near equality" of corresponding distributions. This would be essentially a statement that the mapping from characteristic functions to distributions is a continuous one.

Theorem 1. Continuity of inverse transforms: *Let X and X_n be random variables with given CDFs and corresponding characteristic functions. We have*

$$[\phi_{X_n}(t) \to \phi_X(t),\ \forall\, t] \quad \Rightarrow \quad [X_n \xrightarrow{\mathrm{d}} X].$$

Proof. First, suppose that we are in the special situation that all $|\phi_{X_n}(t)| \leq g(t)$ where $g(t)$ is positive and integrable (on $\mathbb{R}$) function. Then, the inverse Fourier transform exists and we conclude that each X_n and X in such a case must possess a pdf (i.e. X_n's and X are all continuous random variables) given by

$$f_{X_n}(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx}\phi_{X_n}(t)dt$$

and similarly for f_X. By the DCT we conclude that

$$f_{X_n}(x) \to f_X(x)$$

for every x. It will be shown later (in the lecture on uniform integrability) that convergence of pdfs implies convergence in distribution.

Second, to reduce to a special case proven above, notice the following: If Z_ϵ is a collection of random variables (independent of X_n, X) such that $\mathbb{P}[|Z_\epsilon| \leq \epsilon] = 1$ then

$$\forall \epsilon > 0 \quad X_n + Z_\epsilon \xrightarrow{\mathrm{d}} X_n \qquad \iff \qquad X_n \xrightarrow{\mathrm{d}} X\,. \tag{1}$$

Finally, notice that if we take Z_ϵ to have triangular pdf

$$f_{Z_\epsilon}(x) = \begin{cases} \frac{1}{\epsilon^2}(x+\epsilon), & x \in (-\epsilon, 0] \\ \frac{1}{\epsilon^2}(\epsilon - x), & x \in (0, \epsilon) \\ 0, & \text{o/w} \end{cases}$$

then $\phi_{Z_\epsilon}(t) = \frac{4\sin^2(t\epsilon/2)}{t^2\epsilon^2} \leq \frac{\text{const}}{1+\epsilon^2 t^2}$ (a calculation). Since $\phi_{X_n+Z_\epsilon} = \phi_{X_n}\phi_{Z_\epsilon}$ we see that sequence of random variables $X_n + Z_\epsilon$ satisfies conditions of the special case above. Application of (1) completes the proof. □

The preceding theorem involves two separate conditions: (i) the sequence of characteristic functions ϕ_{X_n} converges (pointwise), and (ii) the limit is the characteristic function associated with some other random variable. If we are only given the first condition (pointwise convergence), how can we tell if the limit is indeed a legitimate characteristic function associated with some random variable? One way is to check for various properties that every legitimate characteristic function must possess. One such property is continuity: if $t \to t^*$, then (using dominated convergence),

$$\lim_{t \to t^*} \phi_X(t) = \lim_{t \to t^*} \mathbb{E}[e^{itX}] = \mathbb{E}[e^{it^*X}] = \phi_X(t^*).$$

It turns out that continuity at zero is all that needs to be checked.

Theorem 2. Continuity of inverse transforms: *Let X_n be random variables with characteristic functions ϕ_{X_n}, and suppose that the limit $\phi(t) = \lim_{n\to\infty} \phi_{X_n}(t)$ exists for every t. Then, either*

(i) The function ϕ is discontinuous at zero (in this case X_n does not converge in distribution); or

(ii) The function ϕ is continuous at zero, there exists a random variable X whose characteristic function is ϕ, and $X_n \xrightarrow{\mathrm{d}} X$.

To illustrate the two possibilities in Theorem 2, consider a sequence $\{X_n\}$, and assume that X_n is exponential with parameter λ_n, so that $\phi_{X_n}(t) = \lambda_n/(\lambda_n - it)$.

(a) Suppose that λ_n converges to a positive number λ. Then, the sequence of characteristic functions ϕ_{X_n} converges to the function ϕ defined by $\phi(t) = \lambda/(\lambda - it)$. We recognize this as the characteristic function of an exponential distribution with parameter λ. In particular, we conclude that X_n converges in distribution to an exponential random variable with parameter λ.

(b) Suppose now that λ_n converges to zero. Then,

$$\lim_{n\to\infty} \phi_{X_n}(t) = \lim_{n\to\infty} \frac{\lambda_n}{\lambda_n - it} = \lim_{\lambda \downarrow 0} \frac{\lambda}{\lambda - it} = \begin{cases} 1, & \text{if } t = 0, \\ 0, & \text{if } t \neq 0. \end{cases}$$

Thus, the limit of the characteristic functions is discontinuous at $t = 0$, and X_n does not converge in distribution. Intuitively, this is because the distribution of X_n keeps spreading in a manner that does not yield a limiting distribution.

Proof. We only need to show (ii). The main step is to show that if ϕ is continuous at zero, then collection of measures $\{\mathbb{P}_{X_n}, n = 1, 2, \ldots\}$ is tight. Indeed, from tightness and Prokhorov's criterion we conclude that there exists a convergent subsequence $\mathbb{P}_{X_{n_k}} \to \mathbb{P}_X$ and since $\phi_{n_k} \to \phi$ the characteristic function of $\mathbb{P}_X$ is precisely ϕ, and thus $\mathbb{P}_X$ is identified uniquely. A short argument (Exercise!) shows that then we must have $\mathbb{P}_{X_n} \to \mathbb{P}_X$.

Showing that continuity of ϕ implies tightness requires the following (Fourier-analytic) trick: Tails of the distribution can be read off the small-neighborhood averages of ϕ around 0. Formally, we have

Lemma 1. *Let Y have characteristic function ϕ_Y then for all $a > 0$:*

$$\mathbb{P}\left[|Y| \geq \frac{1}{a}\right] \leq \frac{7}{a}\int_0^a [1 - \operatorname{Re}\phi_Y(t)]dt$$

Lemma indeed implies tightness: From continuity of ϕ for every $\epsilon > 0$ there exists small enough $a > 0$ such that

$$\frac{1}{a}\int_0^a (1 - \operatorname{Re}\phi(t))dt < \frac{\epsilon}{2}$$

and from the DCT there is also an n_0 such that for all $n \geq n_0$ we have

$$\frac{1}{a}\int_0^a (1 - \operatorname{Re}\phi_n(t))dt \leq \frac{1}{a}\int_0^a (1 - \operatorname{Re}\phi(t)) + \frac{\epsilon}{2} \leq \epsilon\,.$$

Finally, we may take $A \geq a$ such that

$$\sup_{n \leq n_0} \mathbb{P}[|X_n| \geq A] \leq \epsilon$$

to conclude the tightness of the whole of $\{\mathbb{P}_{X_n}\}$.

It remains to prove the Lemma. Roughly, the idea is the following. Let Y have PDF f_Y with mass $\delta > 0$ outside $[-A, A]$. Then ϕ_Y is a Fourier transform of f_Y. It is well-known that multiplication of functions corresponds to convolution of Fourier transforms, and vice-versa. Thus, we conclude that $\frac{1}{2\epsilon}\phi_Y * 1_{(-\epsilon,\epsilon)}$ is a Fourier transform of $f_Y(x) \cdot \frac{\sin \epsilon x}{\epsilon x}$. However, note that $\frac{\sin \epsilon x}{\epsilon x}$ kills the tails of f_Y and hence the Fourier transform of the product evaluated at zero should be around $1 - \frac{\delta}{\epsilon A}$.

Rigorously, from

$$1 - \operatorname{Re}\phi_Y(t) = \mathbb{E}[1 - \cos(tY)]$$

by Fubini we have

$$\begin{aligned}
\frac{1}{a}\int_0^a [1 - \operatorname{Re}\phi_Y(t)]dt &= \mathbb{E}\frac{1}{a}\int_0^a [1 - costY]dt & (2)\\
&= \mathbb{E}\left[1 - \frac{\sin aY}{aY}\right] & (3)\\
&\geq (1 - \sin 1)\mathbb{P}\left[|Y| \geq \frac{1}{a}\right]\,, & (4)
\end{aligned}$$

where in the last step we used the fact that $1 - \frac{\sin u}{u}$ is a non-negative function, exceeding $(1 - \sin 1)$ for $|u| > 1$. From (4) lemma follows by noting $(1 - \sin 1) > \frac{1}{7}$. This concludes the proof of Lemma and Theorem. $\square$

3 THE WEAK LAW OF LARGE NUMBERS

Intuitively, an expectation can be thought of as the average of the outcomes over an infinite repetition of the same experiment. If so, the observed average in a finite number of repetitions (which is called the **sample mean**) should approach the expectation, as the number of repetitions increases. This is a vague statement, which is made more precise by so-called laws of large numbers.

Theorem 3. (Weak law of large numbers) *Let X_n be a sequence of i.i.d. random variables, and assume that $\mathbb{E}[|X_1|] < \infty$. Let $S_n = X_1 + \cdots + X_n$. Then,*

$$\frac{S_n}{n} \stackrel{\text{i.p.}}{\to} \mathbb{E}[X_1].$$

This is called the "weak law" in order to distinguish it from the "strong law" of large numbers, which asserts, under the same assumptions, that $X_n \stackrel{\text{a.s.}}{\to} \mathbb{E}[X_1]$. Of course, since almost sure convergence implies convergence in probability, the strong law implies the weak law. On the other hand, the weak law can be easier to prove, especially in the presence of additional assumptions. Indeed, in the special case where the X_i have mean μ and **finite variance**, Chebyshev's inequality yields, for every $\epsilon > 0$,

$$\mathbb{P}\big(|(S_n/n) - \mu| \geq \epsilon\big) \leq \frac{\text{var}(S_n/n)}{\epsilon^2} = \frac{\text{var}(X_1)}{n\epsilon^2}, \tag{5}$$

which converges to zero, as $n \to \infty$, thus establishing convergence in probability.

Historical note: WLLN has been one of the focal points of the development of the probability theory. Reader is welcome to muse upon the mathematical progress made since 1713, when J. Bernoulli proved WLLN for iid $X_j \sim \text{Bern}(p)$. It took him 20 years (his own account) and he referred to it as his "Golden Theorem". The simple proof (5) under finite variance only appeared in Chebyshev's work in 1867 (who used an inequality due to Bienaymé, which we now call Chebyshev's). In 1913 A. Markov organized a big celebration on the occasion of 200'th anniversary of LLN. The final form of the WLLN as given in Theorem 3 was obtained by Khintchine in 1929. For more history see [3].

Before we proceed to the proof for the general case, we note two important facts that we will use.

(a) **First-order Taylor series expansion.** Let $g : \mathbb{R} \to \mathbb{R}$ be a function that has a derivative at zero, denoted by d. Let h be a function that represents the error in a first order Taylor series approximation:

$$g(\epsilon) = g(0) + d\epsilon + h(\epsilon).$$

By the definition of the derivative, we have

$$d = \lim_{\epsilon \to 0} \frac{g(\epsilon) - g(0)}{\epsilon} = \lim_{\epsilon \to 0} \frac{d\epsilon + h(\epsilon)}{\epsilon} = d + \lim_{\epsilon \to 0} \frac{h(\epsilon)}{\epsilon}.$$

Thus, $h(\epsilon)/\epsilon$ converges to zero, as $\epsilon \to 0$. A function h with this property is often written as $o(\epsilon)$. This discussion also applies to complex-valued functions, by considering separately the real and imaginary parts.

(b) **A classical sequence.** Recall the well known fact

$$\lim_{n \to \infty} \left(1 + \frac{a}{n}\right)^n = e^a, \qquad a \in \mathbb{R}. \tag{6}$$

We note (without proof) that this fact remains true even when a is a complex number. Furthermore, with little additional work, it can be shown that if $\{a_n\}$ is a sequence of complex numbers that converges to a, then,

$$\lim_{n \to \infty} \left(1 + \frac{a_n}{n}\right)^n = e^a.$$

Proof of Theorem 3: Let $\mu = \mathbb{E}[X_1]$. Fix some $t \in \mathbb{R}$. Using the assumption that the X_i are independent, and the fact that the derivative of ϕ_{X_1} at $t = 0$ equals $i\mu$, the characteristic function of S_n/n is of the form

$$\phi_n(t) = \big(\mathbb{E}[e^{itX_1/n}]\big)^n = \big(\phi_{X_1}(t/n)\big)^n = \Big(1 + \frac{\mu i t}{n} + o(t/n)\Big)^n,$$

where the function o satisfies $\lim_{\epsilon \to 0} o(\epsilon)/\epsilon = 0$. Therefore,

$$\lim_{n \to \infty} \phi_{X_n}(t) = e^{i\mu t}, \qquad \forall\, t.$$

We recognize $e^{i\mu t}$ as the characteristic function associated with a random variable which is equal to μ, with probability one.

Applying Theorem 1 from the previous lecture (continuity of inverse transforms), we conclude that S_n/n converges to μ, in distribution. Furthermore, as mentioned in the previous lecture, convergence in distribution to a constant implies convergence in probability. $\square$

Remark: It turns out that the assumption $\mathbb{E}[|X_1|] < \infty$ can be relaxed, although not by much. Suppose that the distribution of X_1 is symmetric around zero. It is known that $S_n/n \to 0$, in probability, if and only if $\lim_{n\to\infty} n\mathbb{P}(|X_1| > n) = 0$. There exist distributions that satisfy this condition, while $\mathbb{E}[|X_1|] = \infty$. On the other hand, it can be shown that any such distribution satisfies $\mathbb{E}[|X_1|^{1-\epsilon}] < \infty$, for every $\epsilon > 0$, so the condition $\lim_{n\to\infty} n\mathbb{P}(|X_1| > n) = 0$ is not much weaker than the assumption of a finite mean.

4 THE CENTRAL LIMIT THEOREM

Suppose that $X_1, X_2, \ldots$ are i.i.d. with common (and finite) mean μ and variance σ^2. Let $S_n = X_1 + \cdots + X_n$. The central limit theorem (CLT) asserts that

$$\frac{S_n - n\mu}{\sigma\sqrt{n}}$$

converges in distribution to a standard normal random variable. For a discussion of the uses of the central limit theorem, see the handout from [BT] (pages 388-394).

Proof of the CLT: For simplicity, suppose that the random variables X_i have zero mean and unit variance. Finiteness of the first two moments of X_1 implies that $\phi_{X_1}(t)$ is twice differentiable at zero. The first derivative is the mean (assumed zero), and the second derivative is $-\mathbb{E}[X^2]$ (assumed equal to one), and we can write

$$\phi_X(t) = 1 - t^2/2 + o(t^2),$$

where $o(t^2)$ indicates a function such that $o(t^2)/t^2 \to 0$, as $t \to 0$. The characteristic function of $S_n/\sqrt{n}$ is of the form

$$\big(\phi_X(t/\sqrt{n})\big)^n = \Big(1 - \frac{t^2}{2n} + o(t^2/n)\Big)^n.$$

For any fixed t, the limit as $n \to \infty$ is $e^{-t^2/2}$, which is the characteristic function ϕ_Z of a standard normal random variable Z. Since $\phi_{S_n/\sqrt{n}}(t) \to \phi_Z(t)$ for every t, we conclude that $S_n/\sqrt{n}$ converges to Z, in distribution. $\square$

The central limit theorem, as stated above, does not give any information on the PDF or PMF of S_n. However, some further refinements are possible, under some additional assumptions. We state, without proof, two such results.

(a) Suppose that $\int |\phi_{X_1}(t)|^r \, dt < \infty$, for some positive integer r. Then, S_n is a continuous random variable for every $n \geq r$, and the PDF f_n of $(S_n -$

$\mu_n)/(\sigma\sqrt{n})$ converges pointwise to the standard normal PDF:

$$\lim_{n\to\infty} f_n(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}, \qquad \forall\, z.$$

In fact, convergence is uniform over all z:

$$\lim_{n\to\infty} \sup_z \quad f_n(z) - \frac{1}{\sqrt{2\pi}} e^{-z^2/2} \quad = 0.$$

(b) Suppose that X_i is a discrete random variable that takes values of the form $a+kh$, where a and h are constants, and k ranges over the integers. Suppose furthermore that X has zero mean and unit variance. Then, for any z of the form $z = (na+kh)/\sqrt{n}$ (these are the possible values of $S_n/\sqrt{n}$), we have

$$\lim_{n\to\infty} \frac{\sqrt{n}}{h} \mathbb{P}(S_n = z) = \frac{1}{2\pi} e^{-z^2/2}.$$

4.1 Berry-Esseen theorem

It turns out that CDF of normalized sums approaches the CDF of standard normal *uniformly* on all of $\mathbb{R}$ with speed $\frac{1}{\sqrt{n}}$:

$$\mathbb{P}\left[\frac{S_n - n\mu}{\sigma\sqrt{n}} \le \lambda\right] = \Phi(\lambda) \pm \frac{\mathrm{const}}{\sqrt{n}} \qquad \forall \lambda\,.$$

The following is a precise version. Just like for the CLT there are great many refinements and extensions. For proof see e.g. Theorem 2, Chapter XVI.5 in [1].

Theorem 4 (Berry-Esseen)**.** *Let X_k, $k = 1, \dots, n$ be independent (possibly not identically distributed) with*

$$\mu_k = \mathbb{E}[X_k]\,, \tag{7}$$
$$\sigma_k^2 = \mathrm{var}[X_k]\,, \tag{8}$$
$$t_k = \mathbb{E}[|X_k - \mu_k|^3]\,, \tag{9}$$
$$\sigma^2 = \sum_{k=1}^{n} \sigma_k^2\,, \tag{10}$$
$$T = \sum_{k=1}^{n} t_k\,. \tag{11}$$

Then for any [1] $-\infty < \lambda < \infty$

$$\mathbb{P}\left[\sum_{k=1}^{n}(X_k - \mu_k) \leq \lambda\sigma\right] - \Phi(\lambda) \ \leq \frac{6T}{\sigma^3}\,, \tag{12}$$

whree Φ *is the CDF of* $\mathcal{N}(0,1)$.

References

[1] W. Feller, *An Introduction to Probability Theory and Its Applications, Volume II*, Second edition, John Wiley & Sons, Inc., New York, 1971.

[2] P. Van Beeck, “An application of Fourier methods to the problem of sharpening the Berry-Esseen inequality,” *Z. Wahrscheinlichkeitstheorie und Verw. Geb.*, vol. 23, 187-196, 1972.

[3] E. Seneta, “A Tricentenary history of the Law of Large Numbers,” *Bernoulli*, vol. 19, no. 4, pp.1088–1121, 2013.

[1] Note that for i.i.d. X_k it is known [2] that the factor of 6 in (12) can be replaced by 0.7975.