Lecture 7: Law of Large Numbers and Central Limit Theorem

Yi, Yung (이융)

EE210: Probability and Introductory Random Processes
KAIST EE

MONTH DAY, 2021

Roadmap

- Most remarkable two results in probability theory history

- Weak Law of Large Numbers: Result and Meaning
- Central Limit Theorem: Result and Meaning
- Weak Law of Large Numbers: Proof
  - Inequalities: Markov and Chebyshev
- Central Limit Theorem: Proof
  - Moment Generating Function (MGF)
- Strong Law of Large Numbers

Roadmap

- Most remarkable two results in probability theory history

- Weak Law of Large Numbers: Result and Meaning
- Central Limit Theorem: Result and Meaning
- Weak Law of Large Numbers: Proof
  - Inequalities: Markov and Chebyshev
- Central Limit Theorem: Proof
  - Moment Generating Function (MGF)
- Strong Law of Large Numbers

Our interest: Sum of Random Variables

- Example 1. n students who decides their presence, depending on their feeling. Each student is happy or sad at random. How many students will show their presence?
- Example 2. I am hearing some sound. There are n noisy sources from outside.

- $X_1, X_2, \ldots, X_n$: i.i.d (independent and identically distributed) random variables
- $\mathbb{E}[X_i] = \mu$, $\text{var}[X_i] = \sigma^2$
- Our interest is to understand how the following sum behaves:

$$S_n = X_1 + X_2 + \ldots + X_n$$

Our First Strategy

$$S_n = X_1 + X_2 + \ldots + X_n$$

- Challenging if we intend to approach directly. Even just for $Z = X + Y$, finding the distribution, for example, requires the complex convolution.

  $$p_Z(z) = \mathbb{P}(X + Y = z) = \sum_x p_X(x) p_Y(z - x)$$

- Take a certain scaling with respect to n that corresponds to a new glass, and investigate the system for large n
- First, consider the sample mean, and try to understand how it behaves:

  $$M_n = \frac{X_1 + X_2 + \ldots X_n}{n}$$

Sample Mean

$$M_n = \frac{S_n}{n} = \frac{X_1 + X_2 + \ldots X_n}{n}$$

- Example. n coin tossing. $X_i = 1$ if head, and 0 otherwise. S_n: total number of heads.
- $\mathbb{E}(M_n) = \mu$, $\mathrm{var}(M_n) = \sigma^2/n$
- For large n, the variance decays. We expect that, for large n, M_n looses its randomness and concentrates around μ.
- Why important? If we take the scaling of S_n by $1/n$, it behaves like a deterministic number. This significantly simplifies how we understand the world.
- We call this law of large numbers.

Let's Establish Mathematically

$$M_n = \frac{S_n}{n} = \frac{X_1 + X_2 + \ldots X_n}{n}$$

- What about this? What's wrong?

$$M_n \xrightarrow{n \to \infty} \mu$$

- Ordinary convergence for the sequence of real numbers: $a_n \to a$
  - For every $\epsilon > 0$, there exists n_0, such that for every $n \geq n_0$, $|a_n - a| \leq \epsilon$.
- M_n is a random variable, which is a function from Ω to $\mathbb{R}$.
- Need to mathematically build up the concept of convergence for the sequence of random variables.

- Consider the sequence of rvs $(Y_n)_{n=1,2,\ldots}$, and I want to say they "converge" to a number a.
- Play the game with my friend Lin.
  - Lin, give me any $\epsilon > 0$.
  - OK. Then, let me consider the event $\{|Y_n - a| \geq \epsilon\}$, and compute its probability $a_n = \mathbb{P}(|Y_n - a| \geq \epsilon)$.
  - Now, a_n is just the real number, and I will show that $a_n \to a$ as $n \to \infty$.

Convergence in probability

For any $\epsilon > 0$, $\mathbb{P}\Big(|Y_n - a| \geq \epsilon\Big) \xrightarrow{n\to\infty} 0.$

Weak Law of Large Numbers

$$M_n = \frac{S_n}{n} = \frac{X_1 + X_2 + \ldots X_n}{n}$$

Weak law of large numbers

M_n converges to μ in probability.

- Why "Weak"? There exists a stronger stronger version, which we call "strong" law of large numbers.
- Proof requires some knowledge about useful inequalities, which we cover later.

Roadmap

- Most remarkable two results in probability theory history

- Weak Law of Large Numbers: Result and Meaning
- Central Limit Theorem: Result and Meaning
- Weak Law of Large Numbers: Proof
  - Inequalities: Markov and Chebyshev
- Central Limit Theorem: Proof
  - Moment Generating Function (MGF)
- Strong Law of Large Numbers

Central Limit Theorem: Start with Scaling (1)

- Loosely speaking, WLLG says:

$$(M_n - \mu) \xrightarrow{n\to\infty} 0$$

- However, we don't know how $M_n - \mu$ converges to 0. For example, what's the speed of convergence?
- Question. What should be "something"? Something should what blows up.

$$\boxed{\text{(something)}} \times (M_n - \mu) \xrightarrow{n\to\infty} \text{meaningful thing}$$

$$\boxed{n^{\alpha}} \times (M_n - \mu) \xrightarrow{n\to\infty} \text{meaningful thing}$$

- What's α for our magic?
- The answer is $\frac{1}{2}$

Central Limit Theorem: Start with Scaling (2)

- Reshaping the equation:

$$\sqrt{n} \times (M_n - \mu) = \sqrt{n}\left(\frac{S_n - n\mu}{n}\right) = \frac{S_n - n\mu}{\sqrt{n}}. \quad \text{Let } Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}.$$

- $\mathbb{E}[Z_n] = 0$ and $\mathrm{var}(Z_n) = 1$.
  - Z_n is well-centered with a constant variance irrespective of n.
- We expect that Z_n converges to something meaningful, but what?
- Some deterministic number just like WLLG?
- Interestingly, it converges to some random variable Z that we know very well.

Central Limit Theorem: Formalism

- $Z_n \xrightarrow{n\to\infty} Z$, where $Z \sim N(0,1)$.
- Wait! What kind of convergence? Convergence in probability as in WLLN? No.
- Convergence in distribution (another type of convergence of rvs)

Central Limit Theorem

For every z,

$$\mathbb{P}(Z_n \le z) \xrightarrow{n\to\infty} \mathbb{P}(Z \le z),$$

where $Z \sim N(0,1)$.

- Meaning from scaling perspective.
  - LLN: Scaling S_n by $1/n$, you go to a deterministic world.
  - CLT: Scaling S_n by $1/\sqrt{n}$, you still stay at the random world, but not an arbitrary random world. That's the normal random world, not depending on the distribution of each X_i. Very interesting!

Practical Use of CLT

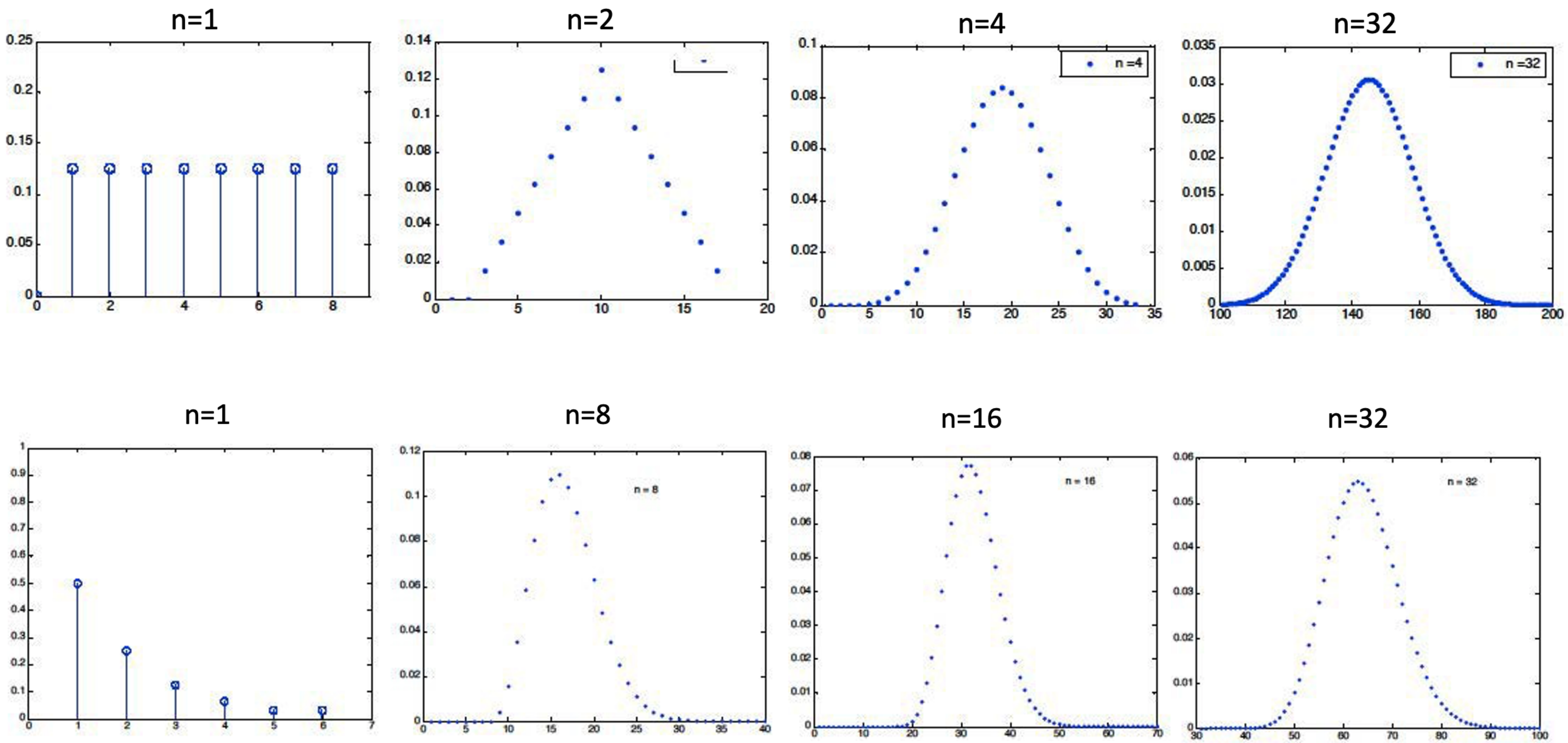
$Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}$

$\mathbb{P}(Z_n \leq z) \xrightarrow{n\to\infty} \mathbb{P}(Z \leq z),\ Z \sim N(0,1)$

- Can approximate Z_n with a standard normal rv
- Can approximate S_n with a normal rv $\sim (n\mu, n\sigma^2)$
  - $S_n = n\mu + Z_n\sigma\sqrt{n}$
- How large should n be?
  - A moderate n (20 or 30) usually works, which the power of CLT.
  - If X_i resembles a normal rv more, smaller n works: symmetry and unimodality[1]

[1] Only unique mode. A single maximum or minimum.

CLT: Examples of n



Roadmap

- Most remarkable two results in probability theory history

- Weak Law of Large Numbers: Result and Meaning
- Central Limit Theorem: Result and Meaning
- Weak Law of Large Numbers: Proof
  - Inequalities: Markov and Chebyshev
- Central Limit Theorem: Proof
  - Moment Generating Function (MGF)
- Strong Law of Large Numbers

KAIST EE

- (Q) Knowing $\mathbb{E}(X)$, can we say something about the distribution of X?
- Intuition: small $\mathbb{E}(X) \implies$ small $\mathbb{P}(X \geq a)$

Markov Inequality

If $X \geq 0$ and $a > 0$, then $\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}(X)}{a}$.

Proof. For any $a > 0$, define Y_a as:

$$Y_a \triangleq \begin{cases} 0, & \text{if } X < a, \\ a, & \text{if } X \geq a \end{cases}$$

Then, using non-negativity of X, $Y_a \leq X$, which leads to $\mathbb{E}[Y_a] \leq \mathbb{E}[X]$.

Note that we have:

$$\mathbb{E}[Y_a] = a\mathbb{P}(Y_a = a) = a\mathbb{P}(X \geq a).$$

Thus, $a \cdot \mathbb{P}(X \geq a) \leq \mathbb{E}[X]$. $\square$

- (Q) Knowing both $\mathbb{E}(X)$ and $\mathrm{var}(X)$, can we say something about the distribution of X?
- Intuition: small $\mathrm{var}(X) \Longrightarrow X$ is unlikely to be too far away from its mean.
- $\mathbb{E}(X) = \mu$, $\mathrm{var}(X) = \sigma^2$.

Chebyshev Inequality

$$\mathbb{P}\Big(|X-\mu| \geq c\Big) \leq \frac{\sigma^2}{c^2}$$

- Proof.

$$\mathbb{P}\Big(|X-\mu| \geq c\Big) = \mathbb{P}\Big((X-\mu)^2 \geq c^2\Big) \leq \frac{\mathbb{E}\Big[(X-\mu)^2\Big]}{c^2} = \frac{\mathrm{var}(X)}{c^2}$$

Example

- $X \sim \exp(1)$. Then, $\mathbb{E}[X] = 1$ and $\mathrm{var}[X] = 1$.

- $\mathbb{E}(X \geq a) = e^{-a}$

- Markov inequality

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[X]}{a} = \frac{1}{a}$$

- Chebyshev inequality

$$\begin{aligned}\mathbb{P}(X \geq a) &= \mathbb{P}(X - 1 \geq a - 1) \\ &\leq \mathbb{P}(|X - 1| \geq a - 1) \leq \frac{1}{(a-1)^2}\end{aligned}$$

- For reasonably large a, CI provides much better bound.

- knowing the variance helps

- Both bounds are the ones that bound the probability of rare events.

$$M_n = \frac{S_n}{n} = \frac{X_1 + X_2 + \dots X_n}{n}$$

Weak law of large numbers

M_n converges to μ in probability.

Proof.

$$\mathbb{P}\Big(|M_n - \mu| \geq \epsilon\Big) \leq \frac{\mathrm{var}(M_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2} \xrightarrow{n\to\infty} 0$$

Roadmap

- Most remarkable two results in probability theory history

- Weak Law of Large Numbers: Result and Meaning
- Central Limit Theorem: Result and Meaning
- Weak Law of Large Numbers: Proof
  - Inequalities: Markov and Chebyshev
- Central Limit Theorem: Proof
  - Moment Generating Function (MGF)
- Strong Law of Large Numbers

Moment Generating Function

- For a rv X, we introduce a kind of transform, called moment generating function (MGF).
- A function of a scalar parameter s, defined by
$$M_X(s) = \mathbb{E}[e^{sX}]$$
- If clear, we omit X and use $M(s)$.

$$M(s) = \sum_x e^{sx} p_X(x) \quad \text{(discrete)}$$

$$M(s) = \int e^{sx} f_X(x) dx \quad \text{(continuous)}$$

Ex1) $X \sim \exp(\lambda)$, $f_X(x) = \lambda e^{-\lambda x}, x \geq 0$

$$\begin{aligned} M(s) &= \lambda \int_0^\infty e^{sx} e^{-\lambda x} dx \\ &= \lambda \left. \frac{e^{(s-\lambda)x}}{s-\lambda} \right|_0^\infty \quad \text{(if } s < \lambda) \\ &= \frac{\lambda}{\lambda - s} \end{aligned}$$

Ex2) $X \sim N(0,1)$ (homework problem)

$$M(s) = e^{s^2/2}$$

Useful Properties of MGF

1. $M'(0) = \mathbb{E}[X]$

$$\frac{d}{ds}M(s) = \frac{d}{ds}\int_{-\infty}^{\infty} e^{sx} f_X(x)dx = \int_{-\infty}^{\infty} \frac{d}{ds}e^{sx} f_X(x)dx = \int_{-\infty}^{\infty} xe^{sx} f_X(x)dx$$

$$= \frac{d}{ds}M(s)\bigg|_{s=0} = \mathbb{E}[X]$$

2. Similarly, $M''(0) = \mathbb{E}[X^2]$

3. $\frac{d^n}{ds^n}M(s)\bigg|_{s=0} = \mathbb{E}[X^n]$

4. MGF provides a convenient way of generating moments. That's why it is called moment generating function.

Inversion Property

The transform $M_X(s)$ associated with a random variable X uniquely determines the CDF of X, assuming that $M_X(s)$ is finite for all s in some interval $[-a, a]$, where a is a positive number.

- In easy words, we can recover the distribution if we know the MGF.
- Thus, each rv has its own MGF.

Back to CLT

- Without loss of generality, assume $\mathbb{E}(X_i) = 0$ and $\text{var}(X_i) = 1$
- $Z_n = \frac{S_n}{\sqrt{n}} = \frac{X_1 + X_2 + \ldots X_n}{\sqrt{n}}$
- Show: MGF of Z_n converges to MFG of $N(0,1)$ (using inversion property)

Proof.

$$\mathbb{E}\left[e^{sS_n/\sqrt{n}}\right] = \mathbb{E}\left[e^{sX_1/\sqrt{n}}\right] \times \cdots \times \mathbb{E}\left[e^{sX_n/\sqrt{n}}\right]$$
$$= \left(\mathbb{E}\left[e^{sX_1/\sqrt{n}}\right]\right)^n = \left(M_{X_1}\left(\frac{s}{\sqrt{n}}\right)\right)^n$$

- For simplicity, let $M(\cdot) = M_{X_1}(\cdot)$
- Facts: $M(0) = 1,\ M'(0) = 0,\ M''(0) = 1$
- $\left(M\left(\frac{s}{\sqrt{n}}\right)\right)^n \to$ what???
- Taking log, $n \log M\left(\frac{s}{\sqrt{n}}\right) \to$ what???

For convenience, do the change of variable $y = \frac{1}{\sqrt{n}}$. Then, we have

$$\lim_{y \to 0} \frac{\log M(ys)}{y^2}$$

- If we apply l'hopital's rule twice (please check), we get

$$\lim_{y \to 0} \frac{\log M(ys)}{y^2} = \frac{s^2}{2}$$

$\square$

Roadmap

- Most remarkable two results in probability theory history

- Weak Law of Large Numbers: Result and Meaning
- Central Limit Theorem: Result and Meaning
- Weak Law of Large Numbers: Proof
  - Inequalities: Markov and Chebyshev
- Central Limit Theorem: Proof
  - Moment Generating Function (MGF)
- Strong Law of Large Numbers (Optional)

Questions?

Review Questions

1) What's the practical value of LLN and CLT?
2) Explain LLN and CLT from the scaling perspective.
3) Why are LLN and CLT great?
4) Why do we need different concepts of convergence for random variables?
5) Explain what is convergence in probability.
6) Explain what is convergence in distribution.
7) Why is MGF useful?