S&DS 241 Lecture 23
Law of large numbers, Moment generating function
B-H: 10.2, 6.4
Setting

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables with mean $\mu$ and variance $\sigma^2$. Let

$$S_n = X_1 + \cdots + X_n, \quad \overline{X}_n = \frac{S_n}{n}$$

Intuition:

- For fair coin flips, we expect $\overline{X}_n$ (fraction of Heads) to be close to $\frac{1}{2}$ if we flip it many times.
Setting

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables with mean $\mu$ and variance $\sigma^2$. Let

$$S_n = X_1 + \cdots + X_n, \quad \overline{X}_n = \frac{S_n}{n}$$

Intuition:

- For fair coin flips, we expect $\overline{X}_n$ (fraction of Heads) to be close to $\frac{1}{2}$ if we flip it many times.
Setting

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables with mean $\mu$ and variance $\sigma^2$. Let

$$S_n = X_1 + \cdots + X_n, \quad \overline{X}_n = \frac{S_n}{n}$$

Intuition:

- For fair coin flips, we expect $\overline{X}_n$ (fraction of Heads) to be close to $\frac{1}{2}$ if we flip it many times.
Setting

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables with mean $\mu$ and variance $\sigma^2$. Let

$$S_n = X_1 + \cdots + X_n, \quad \overline{X}_n = \frac{S_n}{n}$$

Intuition:

- For fair coin flips, we expect $\overline{X}_n$ (fraction of Heads) to be close to $\frac{1}{2}$ if we flip it many times.
Setting

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables with mean $\mu$ and variance $\sigma^2$. Let

$$S_n = X_1 + \cdots + X_n, \quad \bar{X}_n = \frac{S_n}{n}$$

Intuition:

- For fair coin flips, we expect $\bar{X}_n$ (fraction of Heads) to be close to $\frac{1}{2}$ if we flip it many times.
Setting

Let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed (iid) random variables with mean $\mu$ and variance $\sigma^2$. Let

$$S_n = X_1 + \cdots + X_n, \quad \overline{X}_n = \frac{S_n}{n}$$

Intuition:

- For fair coin flips, we expect $\overline{X}_n$ (fraction of Heads) to be close to $\frac{1}{2}$ if we flip it many times.
Law of Large Numbers (LLN)

- **Informal statement**: $\overline{X}_n$ “converges” to the expectation $\mu$ as $n \to \infty$, that is, $\overline{X}_n$ is likely to be close to $\mu$.

  sample (empirical) average $\approx$ population average (expectation)
Law of Large Numbers (LLN)

- **Informal statement**: $\overline{X}_n$ “converges” to the expectation $\mu$ as $n \to \infty$, that is, $\overline{X}_n$ is likely to be close to $\mu$.

  sample (empirical) average $\approx$ population average (expectation)

- **Precise statement**: For any $\epsilon > 0$,

  $$P(|\overline{X}_n - \mu| > \epsilon) \xrightarrow{n \to \infty} 0$$

**Proof.**

Using Chebyshev’s inequality

$$P(|\overline{X}_n - \mu| > \epsilon) \leq \frac{\text{Var}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}$$

since $\text{Var}(\overline{X}_n) = \frac{1}{n^2} \text{Var}(S_n) = \frac{1}{n^2}(\text{Var}(X_1) + \cdots + \text{Var}(X_n)) = \frac{\sigma^2}{n}$.  □
Law of Large Numbers (LLN)

- **Informal statement**: $\overline{X}_n$ “converges” to the expectation $\mu$ as $n \to \infty$, that is, $\overline{X}_n$ is likely to be close to $\mu$.

\[
\text{sample (empirical) average } \approx \text{ population average (expectation)}
\]

- **Precise statement**: For any $\epsilon > 0$,

\[
P(\lvert \overline{X}_n - \mu \rvert > \epsilon) \xrightarrow{n \to \infty} 0
\]

**Proof.**

Using Chebyshev’s inequality

\[
P(\lvert \overline{X}_n - \mu \rvert > \epsilon) \leq \frac{\text{Var}(\overline{X}_n)}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}
\]

since $\text{Var}(\overline{X}_n) = \frac{1}{n^2} \text{Var}(S_n) = \frac{1}{n^2} (\text{Var}(X_1) + \cdots + \text{Var}(X_n)) = \frac{\sigma^2}{n}$. \qed

- Instead of independence, assuming uncorrelated suffices.
Examples of LLN

- Parking lot: 500 spots, 600 permits issued. Suppose each person drives to work with probability 80%, then typically we expect around 480 cars
Examples of LLN

• **Parking lot**: 500 spots, 600 permits issued. Suppose each person drives to work with probability 80%, then typically we expect around 480 cars.

• **Home owner insurance**: Liability of each policy

\[
X = \begin{cases} 
\$100k & \text{w.p. 0.1\% (major)} \\
\$50k & \text{w.p. 0.1\% (substantial)} \\
\$10k & \text{w.p. 1\% (minor)} \\
0 & \text{else}
\end{cases}
\]

Then \( E(X) = \$250 \) is a fair price. The insurance company sets the premium to be $400 to guarantee a decent profit typically.
When does LLN fail?

- By chance: e.g., it is possible, though extremely unlikely, to get all heads in 100 coin flips.
When does LLN fail?

- By chance: e.g., it is possible, though extremely unlikely, to get all heads in 100 coin flips
- $X_1, \ldots, X_n$ are correlated.
When does LLN fail?

- By chance: e.g., it is possible, though extremely unlikely, to get all heads in 100 coin flips
- $X_1, \ldots, X_n$ are correlated. For example:
  - Parking lot: rainy day
  - Home owner insurance: tornado
• **LLN:**

\[ \bar{X}_n = \mu + \text{small error} \]

but it does not say how small the error is, that is, how fast it vanishes as \( n \) grows
Preview: Central Limit Theorem

• **LLN:**

\[ \bar{X}_n = \mu + \text{small error} \]

but it does not say how small the error is, that is, how fast it vanishes as \( n \) grows.

• **CLT:** “small error term” is proportional to \( \frac{\sigma}{\sqrt{n}} \) and approximately Gaussian like:

\[ \bar{X}_n = \mu + \underbrace{\text{small error}}_{\text{approximately } N(0, \frac{\sigma^2}{n})} \]

that is

\[ \bar{X}_n \overset{\text{approx.}}{\sim} N \left( \mu, \frac{\sigma^2}{n} \right) \]
Why do you need to know this?

Question (Lec 16)

Flip a fair coin 100 times. How unlikely is it to get at least 75 heads?
Why do you need to know this?

Question (Lec 16)

Flip a fair coin 100 times. How unlikely is it to get at least 75 heads?

- Normal approximation of $\overline{X}_n$ by $\tilde{X}_n \sim N\left(\frac{1}{2}, \frac{1}{400}\right)$:

$$P(\overline{X}_n \geq 0.75) \approx P(\tilde{X}_n \geq 0.75) = 1 - \Phi(5) = 2.9 \times 10^{-7}$$

- This is justified by CLT for binomial (de Moivre-Laplace theorem), which we proved by brute force (Stirling approximation)
Why do you need to know this?

Question (Lec 16)

Flip a fair coin 100 times. How unlikely is it to get at least 75 heads?

- Normal approximation of $\bar{X}_n$ by $\tilde{X}_n \sim N\left( \frac{1}{2}, \frac{1}{400} \right)$:

\[
P(\bar{X}_n \geq 0.75) \approx P(\tilde{X}_n \geq 0.75) = 1 - \Phi(5) = 2.9 \times 10^{-7}
\]

- This is justified by **CLT for binomial** (de Moivre-Laplace theorem), which we proved by brute force (Stirling approximation)

Question

Toss a fair die 100 times. How unlikely is it for the sum to exceed 400?
Why do you need to know this?

Question (Lec 16)

Flip a fair coin 100 times. How unlikely is it to get at least 75 heads?

• Normal approximation of $\bar{X}_n$ by $\tilde{X}_n \sim N(\frac{1}{2}, \frac{1}{400})$:

$$P(\bar{X}_n \geq 0.75) \approx P(\tilde{X}_n \geq 0.75) = 1 - \Phi(5) = 2.9 \times 10^{-7}$$

• This is justified by CLT for binomial (de Moivre-Laplace theorem), which we proved by brute force (Stirling approximation)

Question

Toss a fair die 100 times. How unlikely is it for the sum to exceed 400?

• Normal approximation of $\bar{X}_n$ by $\tilde{X}_n \sim N(\frac{7}{2}, \frac{35}{1200})$:

$$P(\bar{X}_n \geq 4) \approx P(\tilde{X}_n \geq 4) = 1 - \Phi(2.93) = 0.17\%$$

• How to justify this?
Universality of Gaussian

Sum of 1 independent fair dice
Universality of Gaussian

Sum of 2 independent fair dice

8/20
Universality of Gaussian

Sum of 3 independent fair dice
Universality of Gaussian

Sum of 4 independent fair dice
Universality of Gaussian

Sum of 5 independent fair dice
Universality of Gaussian

Sum of 6 independent fair dice
Universality of Gaussian

Sum of 1 iid Expo(1)
Universality of Gaussian

Sum of 3 iid Expo(1)
Universality of Gaussian

Sum of 5 iid Expo(1)
Universality of Gaussian

Sum of 7 iid Expo(1)
Universality of Gaussian

Sum of 9 iid Expo(1)
Universality of Gaussian

Sum of 11 iid Expo(1)
Universality of Gaussian

Sum of 13 iid Expo(1)
Universality of Gaussian

Sum of 15 iid Expo(1)
Understanding the distribution of $S_n$

Let $X_1, \ldots, X_n$ be iid, with common PDF $f$.

- Recall (Lec 19) the PDF of $X_1 + X_2$ is given by the convolution $f * f$:
  \[
  (f * f)(x) = \int_{-\infty}^{\infty} f(t)f(x - t)dt
  \]

This is difficult to compute if $n$ is large (which is exactly what we are interested in).

We need better tools for handling convolutions!

▶ Moment generating function turns convolutions into products.
Understanding the distribution of $S_n$

Let $X_1, \ldots, X_n$ be iid, with common PDF $f$.

- Recall (Lec 19) the PDF of $X_1 + X_2$ is given by the convolution $f * f$:
  \[
  (f * f)(x) = \int_{-\infty}^{\infty} f(t)f(x - t)dt
  \]

- The PDF of $S_n = X_1 + \cdots + X_n$ is $n$-fold convolution
  \[
  f * f * \cdots * f
  \]
  $n$ times

This is difficult to compute if $n$ is large (which is exactly what we are interested in)
Understanding the distribution of $S_n$

Let $X_1, \ldots, X_n$ be iid, with common PDF $f$.

- Recall (Lec 19) the PDF of $X_1 + X_2$ is given by the convolution $f * f$:
  \[
  (f * f)(x) = \int_{-\infty}^{\infty} f(t) f(x - t) \, dt
  \]

- The PDF of $S_n = X_1 + \cdots + X_n$ is $n$-fold convolution
  \[
  f * f * \cdots * f
  \]

  $n$ times

  This is difficult to compute if $n$ is large (which is exactly what we are interested in)

- We need better tools for handling convolutions!
Understanding the distribution of $S_n$

Let $X_1, \ldots, X_n$ be iid, with common PDF $f$.

- Recall (Lec 19) the PDF of $X_1 + X_2$ is given by the convolution $f * f$:
  \[
  (f * f)(x) = \int_{-\infty}^{\infty} f(t) f(x - t) dt
  \]

- The PDF of $S_n = X_1 + \cdots + X_n$ is $n$-fold convolution
  \[
  f * f * \cdots * f
  \]

This is difficult to compute if $n$ is large (which is exactly what we are interested in)

- We need better tools for handling convolutions!
  - Moment generating function turns convolutions into products.
Moment Generating Function (MGF)
Definition

- The Moment Generating Function (MGF) of a random variable $X$ is defined as:
  \[ M_X(t) = E(e^{tX}), \]
  which is a function of $t \in \mathbb{R}$.
- The $k$th moment of $X$ is
  \[ E(X^k) \]
Why MGF?

- MGF provides a unified way to calculate all moments
- MGF helps us to prove general CLT, going beyond the binomial case
- MGF helps establish sharp concentration inequalities: Chernoff inequality (refined version of Chebyshev inequality) — see HW
From MGF to moments

- Recall Taylor expansion of $e^{tx}$ at $x = 0$:

$$e^{tx} = \sum_{k \geq 0} \frac{t^k}{k!} x^k$$
From MGF to moments

- Recall Taylor expansion of $e^{tx}$ at $x = 0$:

$$e^{tx} = \sum_{k \geq 0} \frac{t^k}{k!} x^k$$

- Replace $x$ by random variable $X$ and take expectation:

$$M_X(t) = E(e^{tx}) = E\left(\sum_{k \geq 0} \frac{t^k}{k!} X^k\right) = \sum_{k \geq 0} \frac{t^k}{k!} E(X^k)$$
From MGF to moments

- Recall **Taylor expansion** of $e^{tx}$ at $x = 0$:

$$e^{tx} = \sum_{k \geq 0} \frac{t^k}{k!} x^k$$

- Replace $x$ by random variable $X$ and take expectation:

$$M_X(t) = E(e^{tX}) = E \left( \sum_{k \geq 0} \frac{t^k}{k!} X^k \right) = \sum_{k \geq 0} \frac{t^k}{k!} E(X^k)$$

- Compare with Taylor expansion of $M_X(t)$ at $t = 0$:

$$M_X(t) = \sum_{k \geq 0} \frac{t^k}{k!} \underbrace{M_X^{(k)}(0)}_{=E(X^k)}$$

that is $M_X(0) = 1$, $M'_X(0) = E(X)$, $M''_X(0) = E(X^2)$, …
From MGF to moments

- The previous formal derivation can be rigorously justified if MGF is finite in a neighborhood near zero.

Summary:

\[ E(X^k) = M_X^{(k)}(0) \]

where:
- \( E(X^k) \) is the \( k \)th moment
- \( M_X^{(k)}(0) \) is the \( k \)th derivative of the MGF at 0

And:

\[ M_X(t) = \sum_{k \geq 0} \frac{E(X^k)}{k!} t^k \]
Example: Bernoulli

For $X \sim \text{Bern}(p)$, we have

$$M_X(t) = E(e^{tX}) = (1 - p) \cdot e^0 + p \cdot e^t = 1 - p + pe^t$$

Then

$$E(X^k) = M_X^{(k)}(0) = p, \quad k \geq 1$$
Example: Bernoulli

For $X \sim \text{Bern}(p)$, we have

$$M_X(t) = E(e^{tX}) = (1 - p) \cdot e^0 + p \cdot e^t = 1 - p + pe^t$$

Then

$$E(X^k) = M_X^{(k)}(0) = p, \quad k \geq 1$$

This is of course obvious because $X \in \{0, 1\}$ so $X^k = X$. 

Example: standard normal

- For $X \sim N(0, 1)$, we have

$$M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} e^{tx} \, dx$$

$$= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-t)^2/2 + t^2/2} \, dx = e^{t^2/2}$$
Example: standard normal

• For $X \sim N(0, 1)$, we have

\[
M_X(t) = E(e^{tX}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} e^{tx} \, dx
= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-t)^2/2 + t^2/2} \, dx = e^{t^2/2}
\]

• Taylor expansion at zero:

\[
e^{t^2/2} = \sum_{k \geq 0} \frac{1}{k!} \left(\frac{t^2}{2}\right)^k = \sum_{k \geq 0} \frac{t^{2k}}{2^k k!}
\]

• Moments of standard normal:

\[
E(X^{2k+1}) = 0 \quad \text{(by symmetry too)}
\]

\[
E(X^{2k}) = \frac{(2k)!}{2^k k!}
\]
Key property of MGF: scaling and shifting

- For any constant $a, b$:

$$M_{aX+b}(t) = M_X(at)e^{bt}$$

**Proof:**

$$M_{aX+b}(t) = E(e^{(aX+b)t}) = E(e^{atX})e^{bt} = M_X(at)e^{bt}$$

- Application: Find MGF of $X \sim N(\mu, \sigma^2)$. 

Solution: Write $X = \mu + \sigma Z$, where $Z \sim N(0, 1)$. Then

$$M_X(t) = e^{\mu t}M_Z(\sigma t) = e^{\mu t} + \sigma^2 t^2/2$$
Key property of MGF: scaling and shifting

- For any constant $a, b$:

\[ M_{aX+b}(t) = M_X(at)e^{bt} \]

**Proof:**

\[
M_{aX+b}(t) = E(e^{(aX+b)t}) = E(e^{atX})e^{bt} \]

- Application: Find MGF of $X \sim N(\mu, \sigma^2)$.

**Solution:** Write $X = \mu + \sigma Z$, where $Z \sim N(0, 1)$. Then

\[ M_X(t) = e^{\mu t}M_Z(\sigma t) = e^{\mu t+\sigma^2 t^2/2} \]
Key property of MGF: sum of independent RVs

- Let $X$ and $Y$ be independent. Then

$$M_{X+Y}(t) = M_X(t) M_Y(t)$$
Key property of MGF: sum of independent RVs

- Let $X$ and $Y$ be independent. Then

\[ M_{X+Y}(t) = M_X(t) M_Y(t) \]

Proof:

\[ M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX} e^{tY}) \overset{\text{independence}}{=} E(e^{tX}) E(e^{tY}) \]

\[ \overset{\text{independence}}{=} \left( M_X(t) \right) \left( M_Y(t) \right) \]
Key property of MGF: sum of independent RVs

• Let $X$ and $Y$ be independent. Then

\[ M_{X+Y}(t) = M_X(t) M_Y(t) \]

Proof:

\[ M_{X+Y}(t) = E(e^{t(X+Y)}) = E(e^{tX} e^{tY}) \overset{\text{independence}}{=} E(e^{tX}) E(e^{tY}) = M_X(t) M_Y(t) \]

• Let $X_1, \ldots, X_n$ be iid and $S_n = X_1 + \cdots + X_n$. Then

\[ M_{S_n}(t) = (M_{X_1}(t))^n \]
Example: Binomial

For $X \sim \text{Bin}(n, p)$, write

$$X = X_1 + X_2 + \cdots + X_n,$$

where $X_i$’s are iid $\text{Bern}(p)$, whose MGF is $1 - p + pe^t$. Then

$$M_X(t) = E(e^{tX}) = (1 - p + pe^t)^n.$$
Example: Binomial

For $X \sim \text{Bin}(n, p)$, write

$$X = X_1 + X_2 + \cdots + X_n,$$

where $X_i$’s are iid $	ext{Bern}(p)$, whose MGF is $1 - p + pe^t$. Then

$$M_X(t) = E(e^{tX}) = (1 - p + pe^t)^n.$$

**Exercise:** Derive this using the binomial PMF.