1179: Probability
Lecture 25 — Concentration Inequalities

Ping-Chun Hsieh (謝秉均)

December 15, 2021

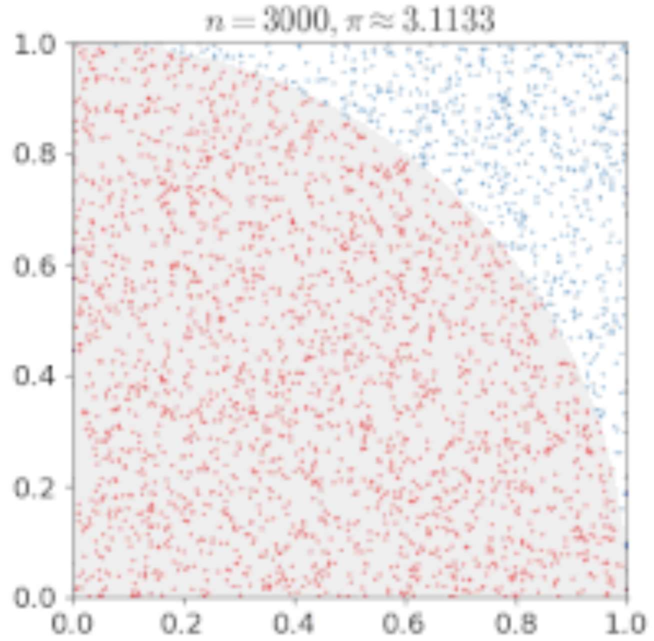
Monte-Carlo Method?

- What is “Monte-Carlo method”?

“… computational algorithms that rely on repeated random sampling to obtain numerical results… use randomness to solve problems that might be deterministic in principle.” (by Wikipedia)

- Math principle behind Monte-Carlo?
- Use Monte-Carlo to estimate π

(Rafael Nadal: 11 titles at Monte Carlo Masters)



This Lecture

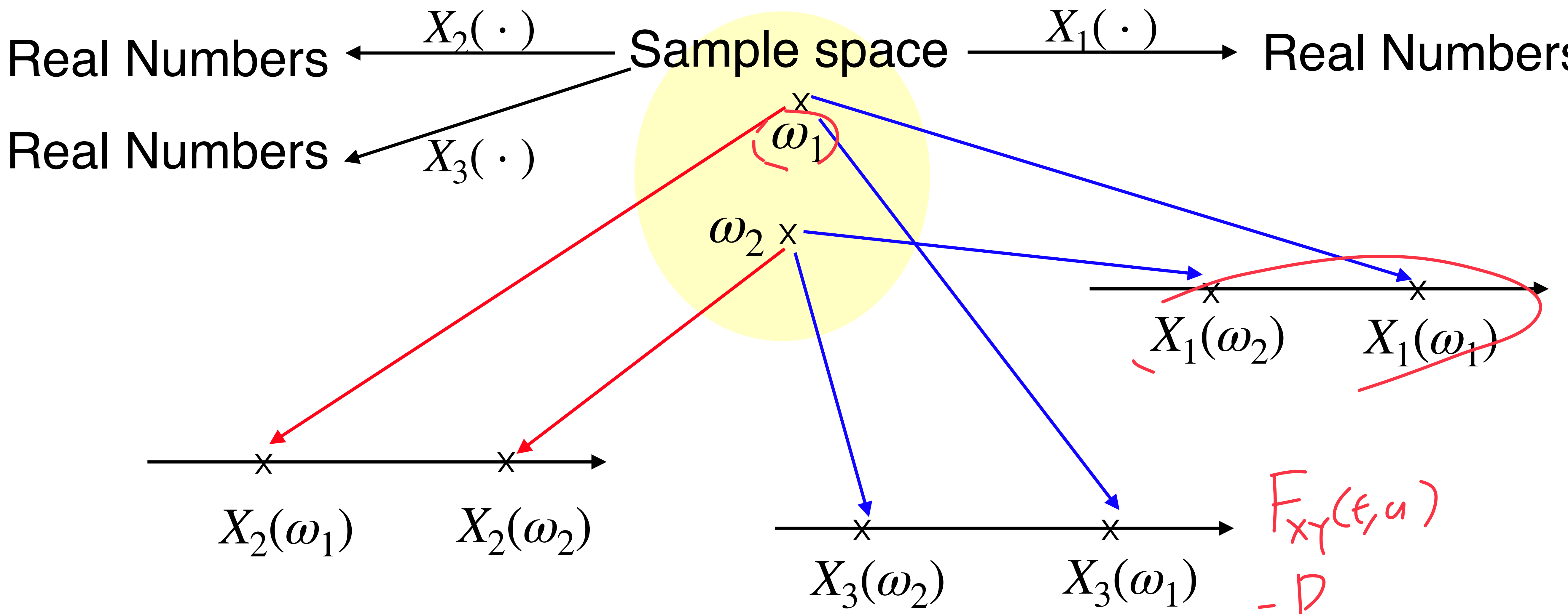
1. Multivariate Random Variables

2. Concentration Inequalities

3. Weak Law of Large Numbers (WLLN)

- Reading material: Chapter 9.1 and 11.3-11.4

A Primer on Multiple Random Variables



$F_{XY}(t,u)$
$= P$

- Could we study the CDF regarding $X_1, X_2,$ and X_3?

$F_{X_1X_2X_3}(x_1, x_2, x_3) = P(X_1 \le x_1, X_2 \le x_2, X_3 \le x_3)$

From Bivariate To Multivariate

- Key Idea: "Bivariate" definitions and properties can be directly extended to the "multivariate" cases

- For example:

1. Joint CDF / PMF / PDF
2. Expected value
3. Marginal CDF / PMF / PDF
4. Independence

Joint CDF of Multivariate R.V.s

Joint CDF of 2 Random Variables: Let X and Y be two random variables defined on the same sample space Ω. The joint CDF $F_{XY}(t,u)$ is defined as

$$F_{XY}(t,u) = \underline{P(X \le t, Y \le u)}, \ \ \forall t, u \in \mathbb{R}$$

Joint CDF of n Random Variables: Let $X_1, \cdots, X_n$ be random variables defined on the same sample space Ω. The joint CDF $F(x_1, x_2 \cdots, x_n)$ is defined as

$$F(x_1, x_2 \cdots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \cdots, X_n \le x_n), \ \ \forall x_i \in \mathbb{R}$$

Joint PDF Multivariate R.V.s

Joint PDF of 2 Random Variables: Let X and Y be two continuous random variables. Then, $f_{XY}(x, y)$ is the joint PDF of X and Y if for every subset B of $\mathbb{R}^2$, we have

$$P((X, Y) \in B) = \iint_B f_{XY}(x, y) dx dy$$

Joint PDF of n Random Variables: Let $X_1, \cdots, X_n$ be n continuous random variables. Then, $f(x_1, x_2, \cdots, x_n)$ is the joint PDF of $X_1, \cdots, X_n$ if for every subset B of $\mathbb{R}^n$, we have

$$P((X_1, X_2, \cdots, X_n) \in B) = \int \cdots \int_B f(x_1, x_2, \cdots, x_n) dx_1 dx_2 \cdots dx_n$$

Expected Value

Expected Value of a Function of 2 Continuous RVs:
Let X, Y be 2 continuous random variables with joint PDF $f_{XY}(x, y)$. Let $g(\cdot, \cdot)$ be a function from $\mathbb{R}^2 \to \mathbb{R}$
The expected value of $g(X, Y)$ is

$$E[g(X, Y)] = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy$$

Expected Value of a Function of n Continuous RVs:
Let $X_1, \cdots, X_n$ be n continuous random variables with joint PDF $f(x_1, x_2, \cdots, x_n)$. Let g be a function from $\mathbb{R}^n \to \mathbb{R}$. The expected value of $g(X_1, X_2, \cdots, X_n)$ is

$$E[g(X_1, X_2, \cdots, X_n)] = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \cdots, x_n) f(x_1, \cdots, x_n) dx_1 dx_2 \cdots dx_n$$

Concentration Inequalities

Motivating Example: Tossing Moon Blocks

- 3 possible outcomes: Yes / No / Laughing
- $p = P(\text{outcome is "Yes"})$
- Each toss is <u>independent</u> from other tosses

Suppose the true $p = 0.5$

100 times

$P(\text{see "0" Yes}) = (0.5)^{100}$

- Question: Suppose p is unknown
  - How to learn p?
  - Could we learn anything useful after n experiments?

Concentration Inequalities

Markov's Inequality

For all $\omega \in \Omega$, $X(\omega) \geq 0$

Tail probability

- **Markov's Inequality**: Let X be a nonnegative random variable. Then, for any $t > 0$,

$$P(X \geq t) \leq \frac{E[X]}{t}$$

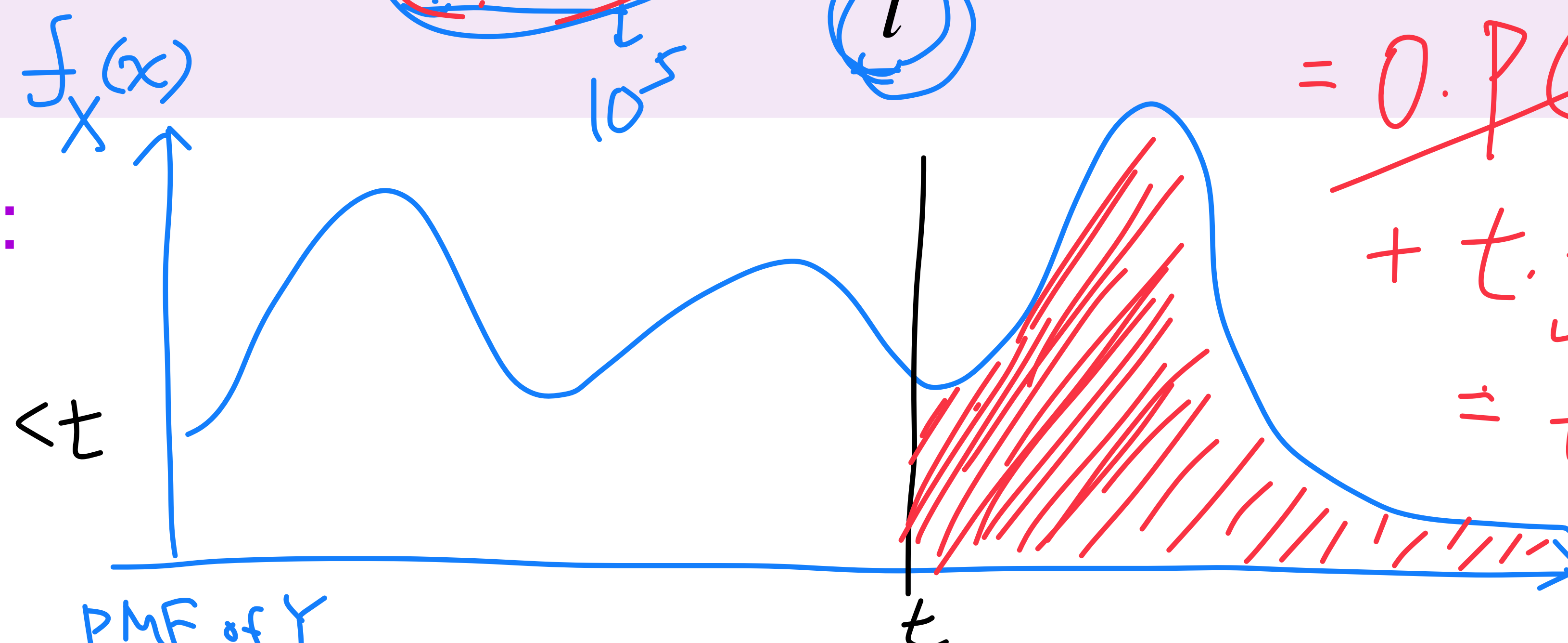
t

10^5

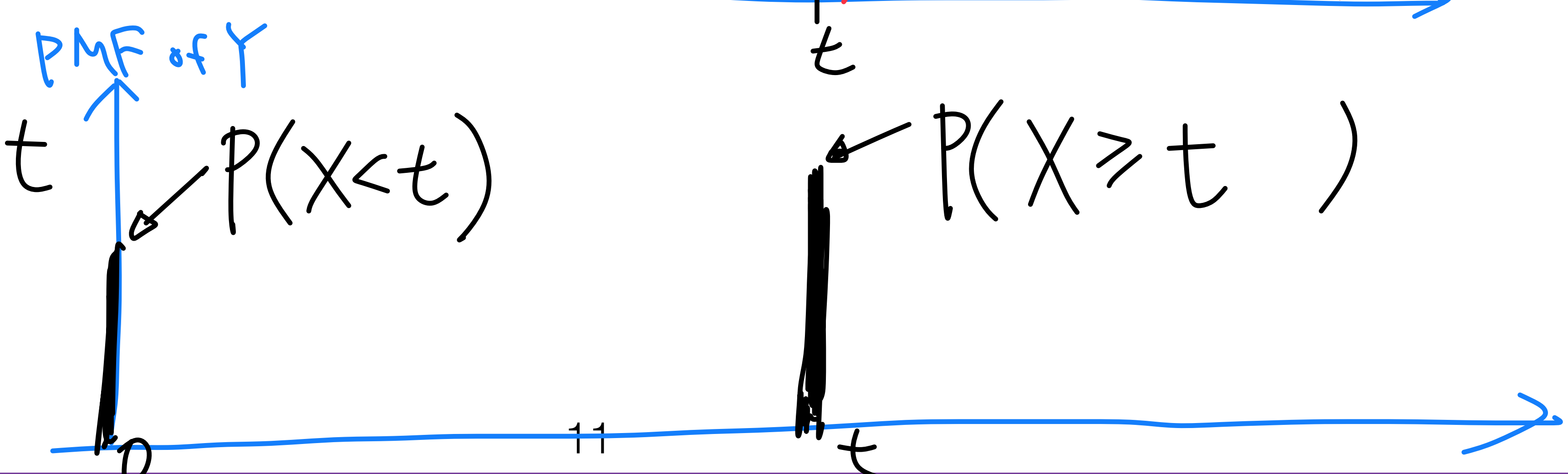
Suppose X is continuous

$$E[X] \geq E[Y]$$
$$= \cancel{0 \cdot P(Y=0)}$$
$$+ \underline{t \cdot P(Y=t)}$$
$$= t \cdot P(X \geq t)$$

- Visualization:



$$Y(\omega) = \begin{cases} 0, & \text{if } X(\omega) < t \\ t, & \text{if } X(\omega) \geq t \end{cases}$$



Proof of Markov's Inequality

- **Markov's Inequality**: Let X be a <u>nonnegative</u> random variable. Then, for any $t > 0$,

$$P(X \geq t) \leq \frac{E[X]}{t}$$

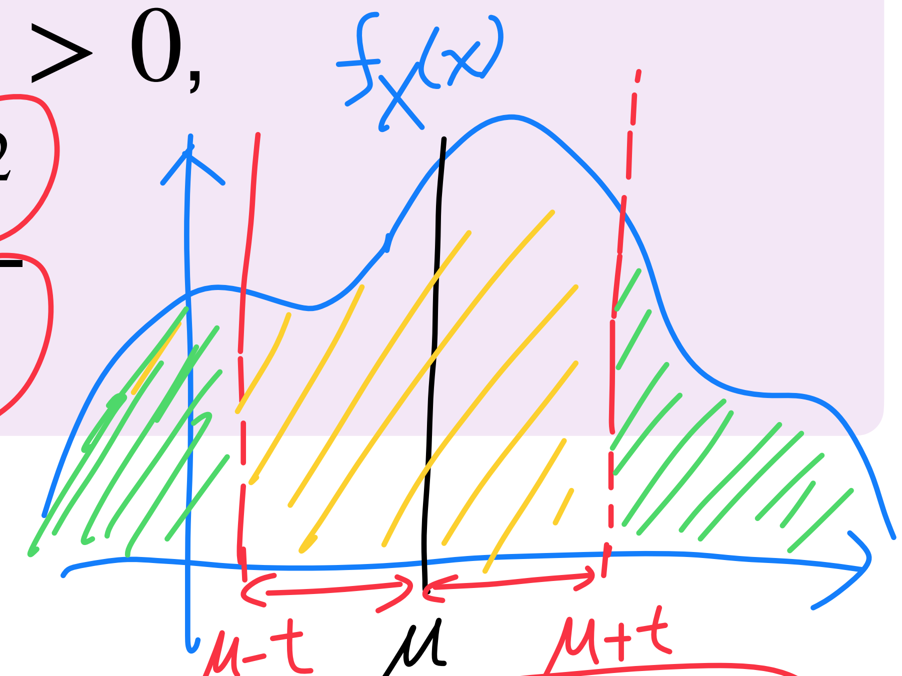
- Proof: (Please refer to the previous page)

Chebyshev's Inequality

- **Chebyshev's Inequality**: Let X be a random variable with mean μ and variance σ^2. Then, for any $t > 0$,

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

Define $Y = |X-\mu|^2$, Y is non-negative



- Proof:

$$\underline{P(|X-\mu| \geq t)}$$
$$= P(|X-\mu|^2 \geq t^2)$$
$$= P(Y \geq t^2)$$

Markov's

$$\leq \frac{E[Y]}{t^2} = \frac{E[|X-\mu|^2]}{t^2} = \frac{\sigma^2}{t^2}$$

Define $Y = |X-\mu|$

$$P(|X-\mu| \geq t)$$
$$= P(Y \geq t)$$
$$\leq \frac{E[Y]}{t}$$

Quick Review: Mean and Variance of Sum of Independent Random Variables

- Example: Each X_i has mean μ_i and variance σ_i^2
  - $X_1, X_2, \cdots, X_n$ are assumed to be independent
  - Question 1: $E[X_1 + X_2 + \cdots + X_n]$ = $\mu_1 + \mu_2 + \cdots + \mu_n$ $\left(\sum_{i=1}^{n} \mu_i\right)$
  - Question 2: $E[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)]$ = $\frac{1}{n}(\mu_1 + \mu_2 + \cdots + \mu_n)$

empirical mean

Quick Review: Mean and Variance of Sum of Independent Random Variables (Cont.)

$X \leftarrow \sigma^2$

$aX \leftarrow a^2\sigma^2$

- Example: Each X_i has mean μ_i and variance σ_i^2
  - $X_1, X_2, \cdots, X_n$ are assumed to be independent
  - Question 3: Var$[X_1 + X_2 + \cdots + X_n]$ = $\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2 \left(\equiv \sum_{i=1}^{n} \sigma_i^2 \right)$
  - Question 4: Var$[\underbrace{\frac{1}{n}(X_1 + X_2 + \cdots + X_n)}_{\text{empirical mean}}]$ = $\left(\frac{1}{n}\right)^2 \cdot \left(\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2\right)$

$$Var[X_1 + X_2] = Var[X_1] + Var[X_2] + 2 \cdot \underbrace{Cov(X_1, X_2)}_{=0}$$

$$= \sigma_1^2 + \sigma_2^2$$

$$Var[\underbrace{X_1 + X_2 + \cdots + X_n}_{X}] = E\left[\left((X_1 + X_2 + \cdots + X_n) - \underbrace{E[X_1 + X_2 + \cdots + X_n]}_{\mu_1 + \mu_2 + \cdots + \mu_n}\right)^2\right]$$

$$= E\left[\left((X_1 - \mu_1) + (X_2 - \mu_2) + \cdots + (X_n - \mu_n)\right)^2\right]$$

$$= \sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$$

Chebyshev's Inequality and Sample Mean

- Example: Tossing moon blocks

- Each toss X_i is Bernoulli with $P(\text{outcome is "Yes"}) = p$
- Each toss is independent from other tosses
- Question: Can we say anything about the sample mean of n tosses $\frac{1}{n}(X_1 + \cdots + X_n)$?

Define $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$

$E[\bar{X}] = \frac{1}{n}(\underbrace{p + p + \cdots + p}_{n \text{ terms}}) = p$

$Var[\bar{X}] = \frac{1}{n^2}(\underbrace{p(1-p) + \cdots + p(1-p)}_{n \text{ terms}}) = \frac{1}{n}p(1-p)$

By Chebyshev's, we have

$$P(|\bar{X} - \underbrace{E[\bar{X}]}_{= p}| \geq t) \leq \frac{\overbrace{Var[\bar{X}]}^{= \frac{1}{n}p(1-p)}}{t^2}$$

Chebyshev's Inequality and Sample Mean (Formally)

- **Chebyshev's and Sample Mean**: Let $X_1, \cdots, X_n$ be a sequence of <u>independent and identically distributed</u> (i.i.d.) random variables with mean μ and variance σ^2. Define $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$. Then, for any $\varepsilon > 0$, we have

empirical mean

$$P(|\bar{X} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{\varepsilon^2 n} = O\left(\frac{1}{n}\right)$$

Any Issue With Chebyshev's Inequality?

- Example: $X_1, \cdots, X_n$ are i.i.d. Bernoulli with parameter 0.5
  - $E[X_i] =$ ________ and Var$[X_i] =$ ________
  - Chebyshev's: $P(|\bar{X} - \mu| \geq \varepsilon) \leq \dfrac{\sigma^2}{\varepsilon^2 n}$
  - Let's plot $P(|\bar{X} - \mu| \geq \varepsilon)$ for small ε

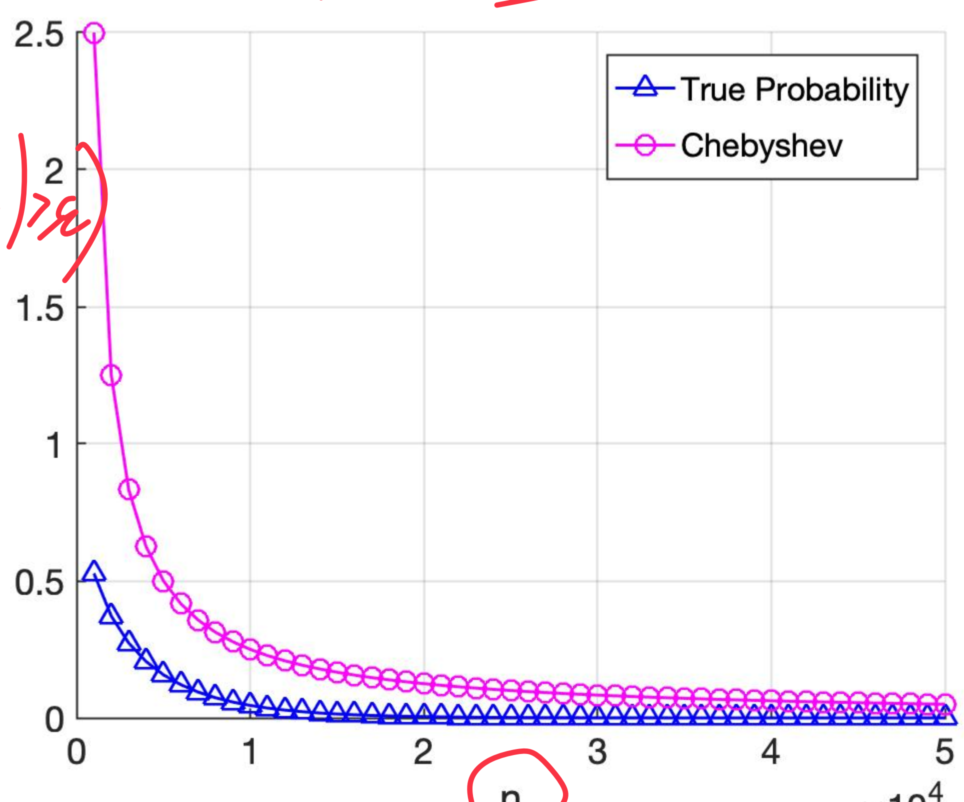
$\bar{X} = \frac{1}{n}(X_1 + X_2 + \cdots + X_n)$

Y

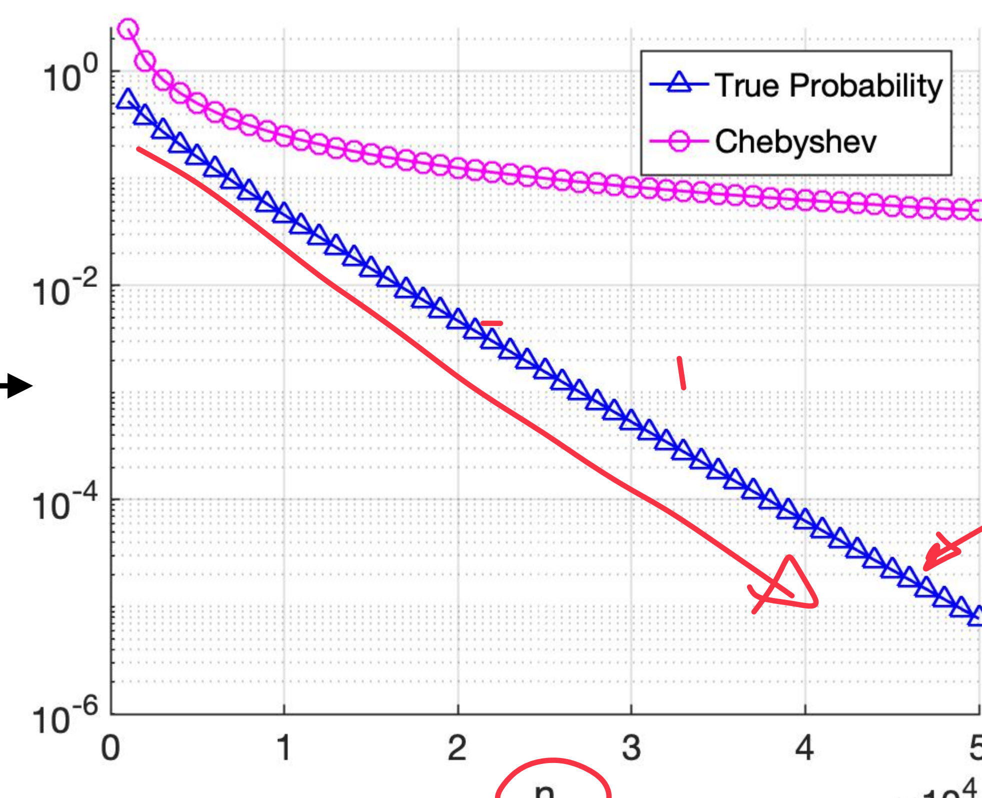
$\varepsilon = 0.01$

$\frac{\sigma^2}{\varepsilon^2 n}$

$P(|\bar{X} - \mu| \geq \varepsilon)$



log scale



Chernoff Bound

- **Chernoff Bound**: Let X be a random variable with MGF $M_X(t)$. Suppose $M_X(t)$ exists for all t in some set S. Then, for any $t > 0$ and $t \in S$, for any $a \in \mathbb{R}$, we have

$$P(X \geq a) \leq e^{-ta} \cdot M_X(t)$$

$t=1$
$t=2$
e^x

tail probability

- Proof:

$$P(X \geq a) = P(\underbrace{e^{tX}}_{Y} \geq e^{ta}), \quad (t>0)$$

$$P(X \geq a) = P(3X+3 \geq 3a+3)$$

$$\leq \frac{E[Y]}{e^{ta}} = \frac{E[e^{tX}]}{e^{ta}} = \frac{M_X(t)}{e^{ta}}$$

Optimizing the Chernoff Bound

- **Chernoff Bound**: Let X be a random variable with MGF $M_X(t)$. Suppose $M_X(t)$ exists for all t in some set S. Then, for any $t > 0$ and $t \in S$, for any $a \in \mathbb{R}$, we have

$$P(X \geq a) \leq e^{-\phi(a)},$$

where $\phi(a) = \max_{t>0, t\in S} (ta - \ln M_X(t))$

- **Proof:** $P(X \geq a) \leq e^{-ta} \cdot M_X(t) = e^{-\boxed{(ta - \ln M_X(t))}}$

$\phi_t(a) = (ta - \ln M_X(t))$

$$\Rightarrow \quad P(X \geq a) \leq e^{-\left(\max_{t>0, t\in S} \phi_t(a)\right)}$$

Example: Chernoff Bound for Bernoulli R.V.s

- Example: Suppose $X \sim$ Bernoulli(p)
  - What is $M_X(t)$?
  - What is the Chernoff bound for X? ($P(X \geq a) \leq e^{-ta} \cdot M_X(t)$)

Example: Optimizing Chernoff Bound for Bernoulli R.V.s

- Example: Suppose $X \sim$ Bernoulli(p)
  - How to optimize the Chernoff bound for X?
    $(P(X \geq a) \leq e^{-\phi(a)}, \phi(a) = \max_{t>0, t \in S} (ta - \ln M_X(t)))$

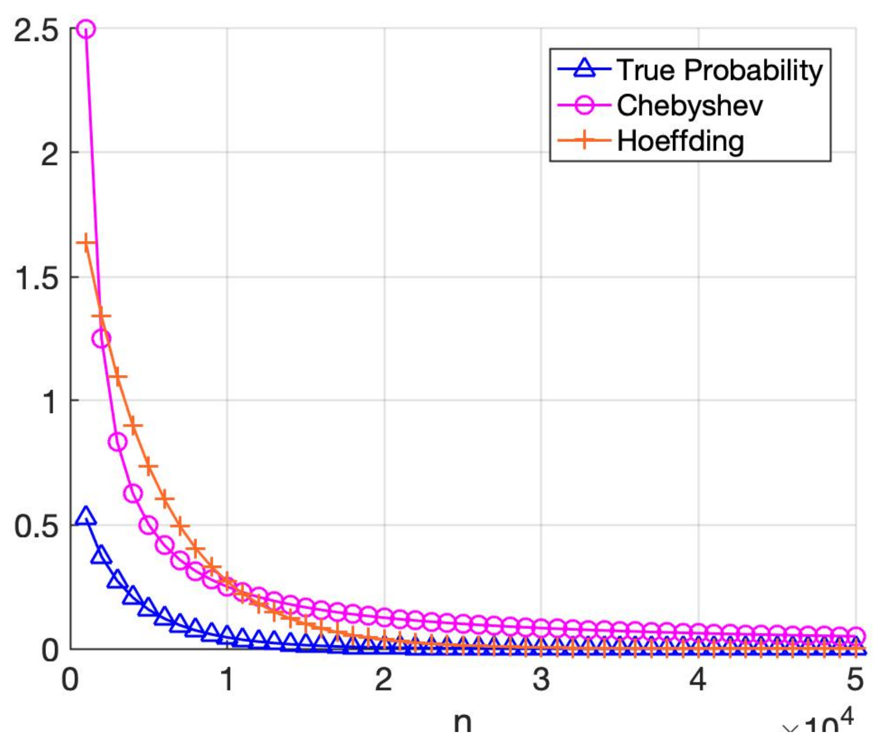
<u>How about applying Chernoff bound to</u>
<u>“sum of independent random variables”?</u>

Hoeffding's Inequality (Formally)

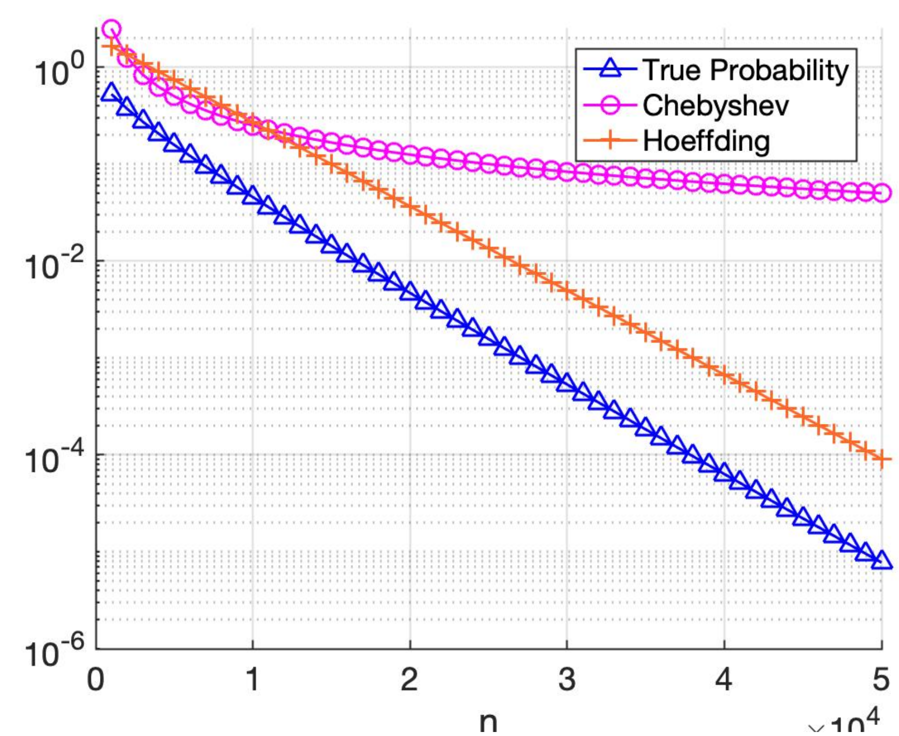
- **Hoeffding's Inequality (For Bernoulli)**: Let $X_1, \cdots, X_n$ be a sequence of i.i.d. Bernoulli random variables with parameter p. Define $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$. Then, for any $\varepsilon > 0$, we have

$$P(|\bar{X} - p| \geq \varepsilon) \leq 2\exp(-2n\varepsilon^2)$$

$\varepsilon = 0.01$



log scale



Proof of Hoeffding's Inequality (Positive Part)

$$P(\bar{X} - p \geq \varepsilon) \leq \exp(-2n\varepsilon^2)$$

- [Hint] Chernoff bound: $P(X \geq a) \leq e^{-ta} \cdot M_X(t)$

$$P(\bar{X} - p \geq \varepsilon) \leq$$

Hoeffding's Lemma

- **Hoeffding's Lemma**: Let Z be a random variable with $E[Z] = 0$, and $Z \in [a, b]$ with probability 1. Then, for any $t > 0$, we have
$$E[e^{tZ}] \leq \exp\left(\frac{t^2(b-a)^2}{8}\right)$$

- Question: If $Z \sim$ Bernoulli(p), then $E[e^{t(Z-p)}] \leq$

Proof of Hoeffding's Inequality (Negative Part)

$$P(\bar{X} - p \leq -\varepsilon) = P(p - \bar{X} \geq \varepsilon) \leq \exp(-2n\varepsilon^2)$$

- [Hint] Chernoff bound: $P(X \geq a) \leq e^{-ta} \cdot M_X(t)$

$$P(p - \bar{X} \geq \varepsilon) \leq$$

Weak Law of Large Numbers (WLLN)

Review: Chebyshev's and Sample Mean: $n \to \infty$

- **Chebyshev's and Sample Mean**: Let $X_1, \cdots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2. Define $S_n = (X_1 + \cdots + X_n)$. Then, for any $\varepsilon > 0$, we have

$$P\left(\left| \frac{S_n}{n} - \mu \right| \geq \varepsilon \right) \leq \frac{\sigma^2}{\varepsilon^2 n}$$

- What if we let $n \to \infty$?

The Weak Law of Large Numbers (WLLN)

- **The Weak Law of Large Numbers (Khinchin's Law)**: Let $X_1, \cdots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean μ. Define $S_n = (X_1 + \cdots + X_n)$. Then, for every $\varepsilon > 0$, we have

$$P\left(\left| \frac{S_n}{n} - \mu \right| \geq \varepsilon \right) \to 0 \quad \text{as } n \to \infty$$

- Question: Any change in technical conditions (cf: Chebyshev's)?

- Question: What does "convergence" mean here?

Convergence in Probability

- **Convergence of a Deterministic Sequence**: Let $a_1, a_2 \cdots$ be a sequence of real numbers. We say that a_n converges to a if for every $\varepsilon > 0$, there exists N_0 such that

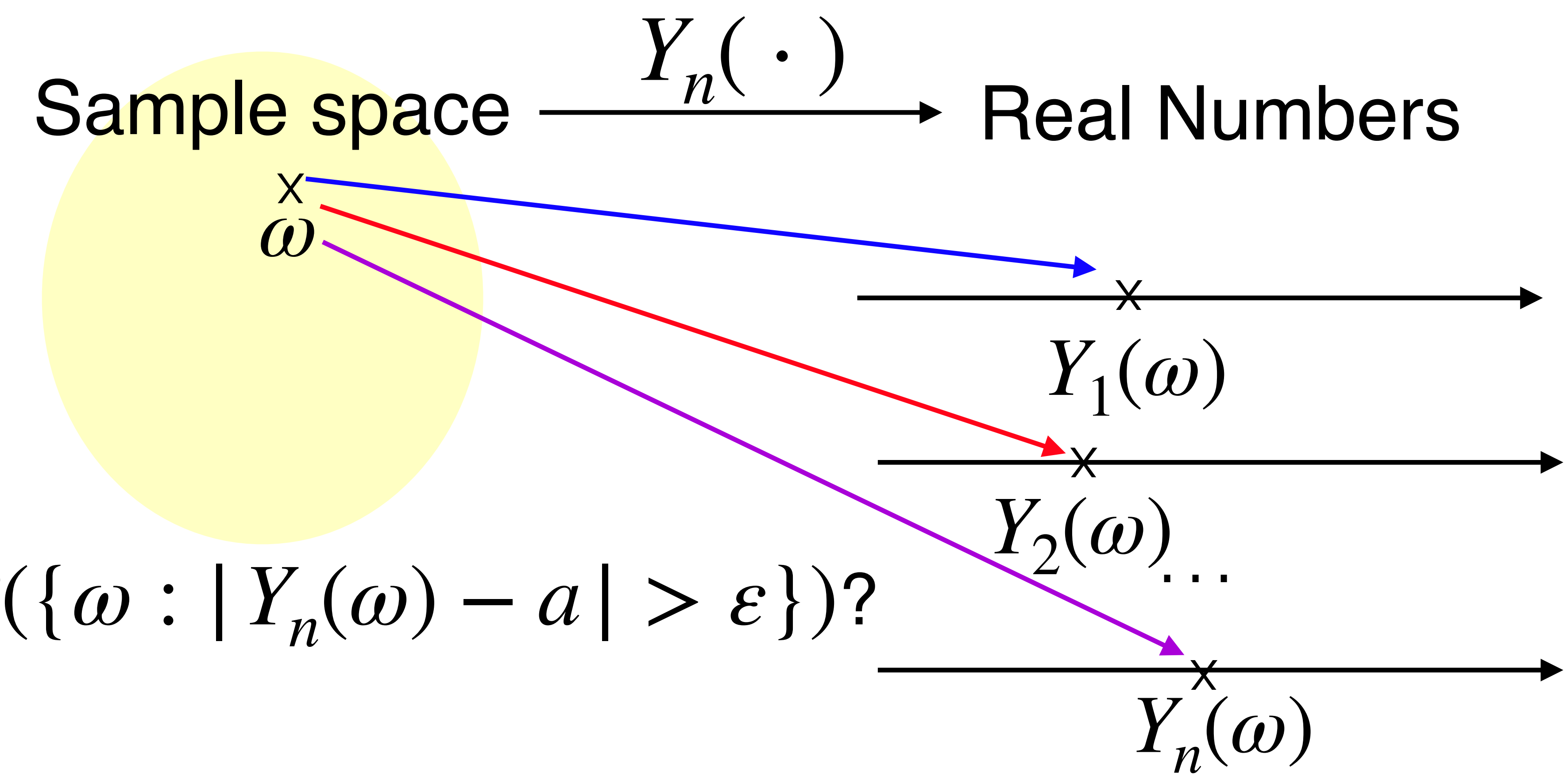
$$|a_n - a| \leq \varepsilon \qquad \text{for all } n \geq N_0$$

- **Convergence to a Scalar in Probability**: Let $Y_1, Y_2 \cdots$ be a sequence of random variables, and let a be a real number. We say that Y_n converges to a in probability if for every $\varepsilon > 0$,

- Question: How to interpret this definition?

Recall: Random Variables Defined on Ω

- $Y_1, Y_2, \cdots, Y_n, \cdots$ are defined on the <u>same sample space Ω</u>



- How to interpret $P(\{\omega : |Y_n(\omega) - a| > \varepsilon\})$?

- How about $\lim_{n\to\infty} P(\{\omega : |Y_n(\omega) - a| > \varepsilon\}) = 0$?

Example: Convergence in Probability

- Example: Consider a sequence of r.v.s Y_n

$$P(Y_n = y) = \begin{cases} 1 - \frac{1}{n} & , \text{ if } y = 0 \\ \frac{1}{n} & , \text{ if } y = n^2 \\ 0 & , \text{ otherwise} \end{cases}$$

- For every $\varepsilon > 0$, can we find $P(|Y_n - 0| > \varepsilon)$?
- How about $\lim_{n\to\infty} P(|Y_n - 0| > \varepsilon)$?