DCP 1206: Probability
Lecture 21 — MGF and Concentration Inequalities

Ping-Chun Hsieh

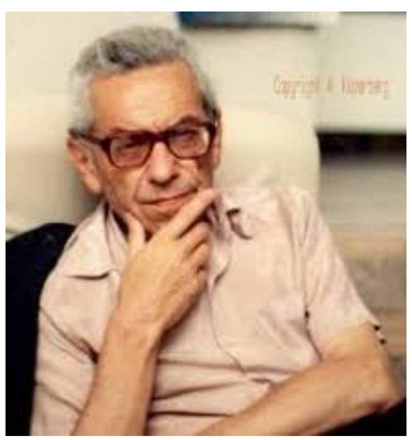
November 29, 2019

Announcements

- No class next Wednesday (12/4)
- HW5 is posted on E3 (Due: 12/11 in class)

Erdős Number

- Six degree of separation?
- In math, Erdős Number embodies a similar principle



Paul Erdős

Richard Bellman
(Erdős number = 1)

Noga Alon
(Erdős number = 1)

Nick Duffield
(Erdős number = 2)

(the most prolific mathematician: 1500+ papers)

The Erdős Number Project: http://www.oakland.edu/enp/

This Lecture

1. Moment Generating Functions (MGF)

2. Concentration Inequalities

- Reading material: Chapter 11.1-11.3

Nice Properties of MGF?

$X \xrightarrow{MGF} E[e^{tX}]$

$\Uparrow$

$M_X(t)$

- **Let X_1, X_2 be two random variables**:

1. ______ Suppose $M_{X_1}(t) = M_{X_2}(t)$, for all $t \in \mathbb{R}$. Do X_1 and X_2 always have the same distribution (i.e. the same CDF)?

2. ______ Could we find moments $E[X_1^n]$ by using $M_{X_1}(t)$?

3. ______ Suppose X_1, X_2 are independent. Could we express $M_{X_1+X_2}(t)$ in terms of $M_{X_1}(t), M_{X_2}(t)$?

$= M_{X_1}(t) \cdot M_{X_2}(t)$

Nice Property (III): Why Is $M_X(t)$ Called the Moment Generating Function?

- **Recall:** What is the "n-th moment" of X? $E[X^n]$

- **Use MGF to Find Moments**: Let X be a random variable with MGF $M_X(t)$. Then, for every $n \in \mathbb{N}$, we have

$$E[X^n] = \frac{d^n}{dt^n} M_X(t)\Big|_{t=0}$$

- **Proof:**

$$M_X(t) = E[e^{tX}] = \sum_{\text{all } x} e^{tx} \cdot p_X(x)$$

$$\frac{d}{dt} M_X(t) = \sum_{\text{all } x} \underbrace{\left(\frac{d}{dt} e^{tx}\right)}_{= x \cdot e^{tx}} \cdot p_X(x) \xrightarrow{t=0} \sum_{\text{all } x} x \cdot 1 \cdot p_X(x) = E[X]$$

$$\frac{d^n}{dt^n} M_X(t) = \sum_{\text{all } x} \left(x^n \cdot e^{tx}\right) p_X(x) \xrightarrow{t=0} \sum_{\text{all } x} x^n \cdot 1 \cdot p_X(x) = E[X^n]$$

Recap: Moment Generating Function (Formally)

- **Moment Generating Function (MGF)**: For a random variable X, define

$$M_X(t) = E[e^{tX}],\ t \in \mathbb{R}$$

If there exists $\delta > 0$ such that $M_X(t) < \infty$ for all $t \in (-\delta, \delta)$, then $M_X(t)$ is called the moment generating function of X

- Question: Why do we emphasize $t \in (-\delta, \delta)$?

For generating moments
(plugging in $t=0$)

Example: Moments of Exp(λ)

$Var[X] = E[X^2] - (E[X])^2$

$= \frac{2}{\lambda^2} - \frac{1}{\lambda^2} = \frac{1}{\lambda^2}$

- Example: Suppose $X \sim \text{Exp}(\lambda)$
  - What is the MGF of X?
  - Use MGF to verify that $E[X] = \frac{1}{\lambda}$ and Var$[X] = \frac{1}{\lambda^2}$?

$M_X(t) = E[e^{tX}]$

$= \int_0^\infty \underbrace{(e^{tx})}_{\text{function value}} \underbrace{(\lambda e^{-\lambda x})}_{\text{PDF}} dx = \int_0^\infty \lambda \cdot e^{(t-\lambda)x} dx$

$= \frac{\lambda}{t-\lambda} \cdot e^{(t-\lambda)x} \Big|_0^\infty$

$= \frac{\lambda}{t-\lambda}(0 - 1)$

$= \frac{\lambda}{\lambda - t} \quad (t < \lambda)$

$E[X] = \frac{d}{dt} M_X(t) \Big|_{t=0} = \frac{d}{dt}\left(\frac{\lambda}{\lambda - t}\right)\Big|_{t=0}$

$= \frac{\lambda}{(\lambda-t)^2}\Big|_{t=0} = \frac{1}{\lambda}$

$E[X^2] = \frac{d^2}{dt^2} M_X(t)\Big|_{t=0} = \frac{2\lambda}{(\lambda-t)^3}\Big|_{t=0} = \frac{2}{\lambda^2}$

2. Concentration Inequalities

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

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

Concentration Inequalities

Markov's Inequality

$E[X] \geq E[Y]$

$\|$

$E[Y] = t(P(X \geq t)) \cancel{+ 0 \cdot (1-P(X\geq t))}$

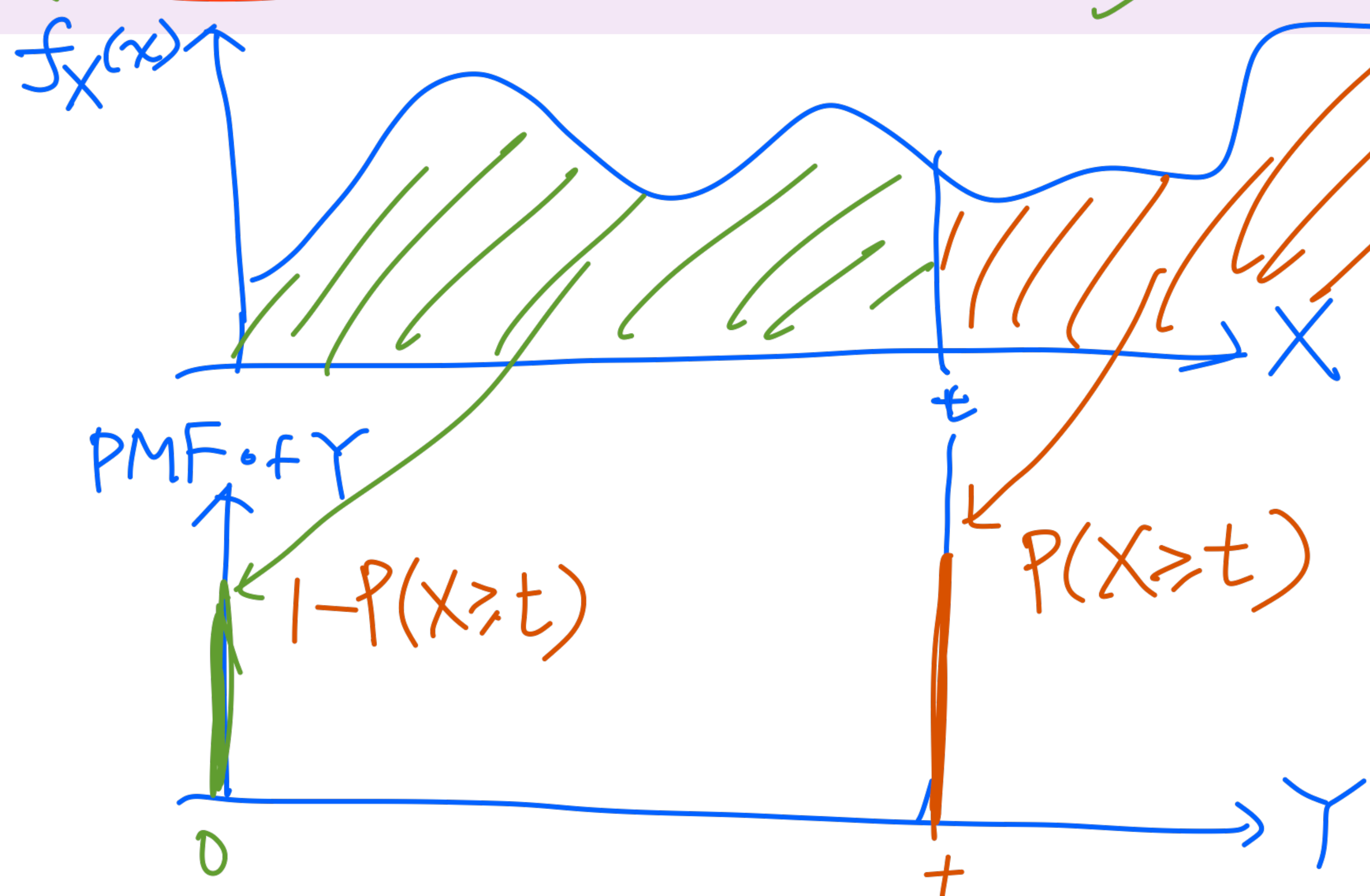
- **Markov's Inequality**: Let X be a nonnegative random variable. Then, for any $t > 0$,

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

- Visualization:

$$Y = \begin{cases} t, & X \geq t \\ 0, & 0 \leq X < t \end{cases}$$

(Y is discrete)



Proof of Markov's Inequality

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

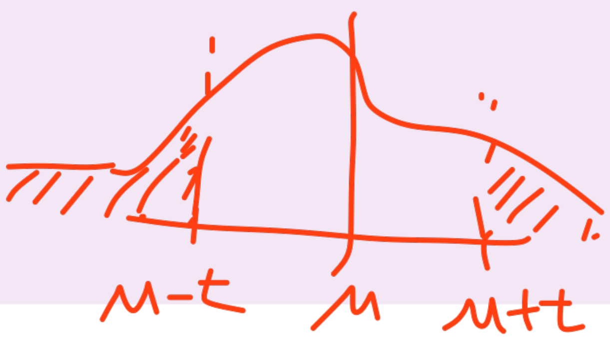
$$P(X \geq t) \leq \frac{E[X]}{t}$$

- Proof:

Please see 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 $t > 0$,



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

- Proof:

Define $Y = (X-\mu)^2 \geq 0$

By Markov's inequality:

$$\underbrace{P(Y \geq t^2)}_{\substack{\| \\ P(|X-\mu| \geq t)}} \leq \frac{E[Y]}{t^2} = \frac{E[(X-\mu)^2]}{t^2} = \frac{\sigma^2}{t^2}$$

tail probability

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: $\underbrace{E[X_1 + X_2 + \cdots + X_n]}$ = $\mu_1 + \mu_2 + \cdots + \mu_n$
  - Question 2: $\underbrace{E[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)]}$ = $\frac{1}{n}(\mu_1 + \mu_2 + \cdots + \mu_n)$

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

- 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$[X_1 + X_2 + \cdots + X_n]$ = $\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2$
  - Question 4: Var$[\frac{1}{n}(X_1 + X_2 + \cdots + X_n)]$ = $\frac{1}{n^2}(\sigma_1^2 + \sigma_2^2 + \cdots + \sigma_n^2)$

$$= E\Big[\big(X_1 + X_2 + \cdots + X_n - \underbrace{E[X_1 + \cdots + X_n]}\big)^2\Big]$$

$$= E\Big[\big((X_1 - E[X_1]) + (X_2 - E[X_2]) + \cdots + (X_n - E[X_n])\big)^2\Big]$$

$$= \mathrm{Var}(X_1) + \mathrm{Var}(X_n) + \cdots + \mathrm{Var}(X_n)$$

$$+\, 2 \cdot \big(\underbrace{\mathrm{Cov}(X_1, X_2) + \mathrm{Cov}(X_1, X_3) + \cdots}_{=0 \text{ (by independance)}}\big)$$

Chebyshev's Inequality and Sample Mean

$Var[X_1] = p(1-p)$

- 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 $\frac{1}{n}(X_1 + \cdots + X_n)$?

$$P\left(\left|\frac{1}{n}(X_1+\cdots+X_n) - \underbrace{E\left[\frac{1}{n}(X_1+\cdots+X_n)\right]}_{=\,p}\right| \ge t\right) \le \frac{\overbrace{Var\left[\frac{1}{n}(X_1+\cdots+X_n)\right]}^{=\,\frac{p(1-p)}{n}}}{t^2}$$

$$\Rightarrow P\left(\left|\text{sample mean} - p\right| \ge t\right) \le \frac{p(1-p)}{nt^2}$$

fix $t=0.01$

Chebyshev's Inequality and Sample Mean (Formally)

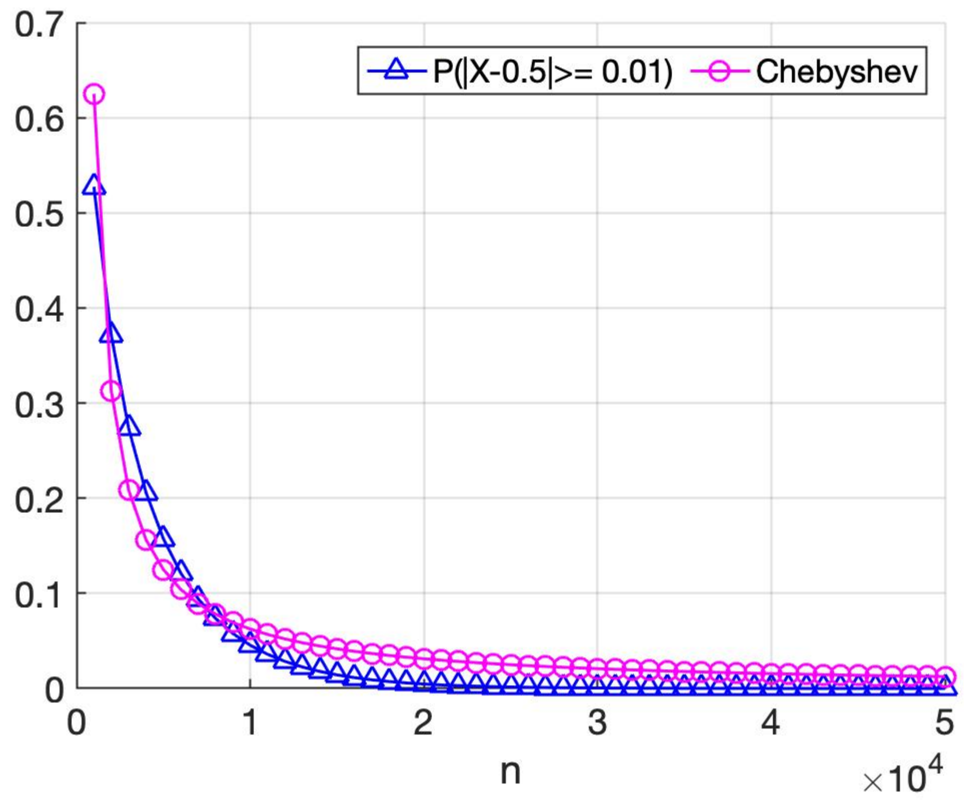
- **Chebyshev's and Sample Mean**: Let $X_1, \cdots, X_n$ be a sequence of independent and identically distributed (i.i.d.) random variables with mean μ and variance σ^2. Define $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$. Then, for any $\varepsilon > 0$, we have

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

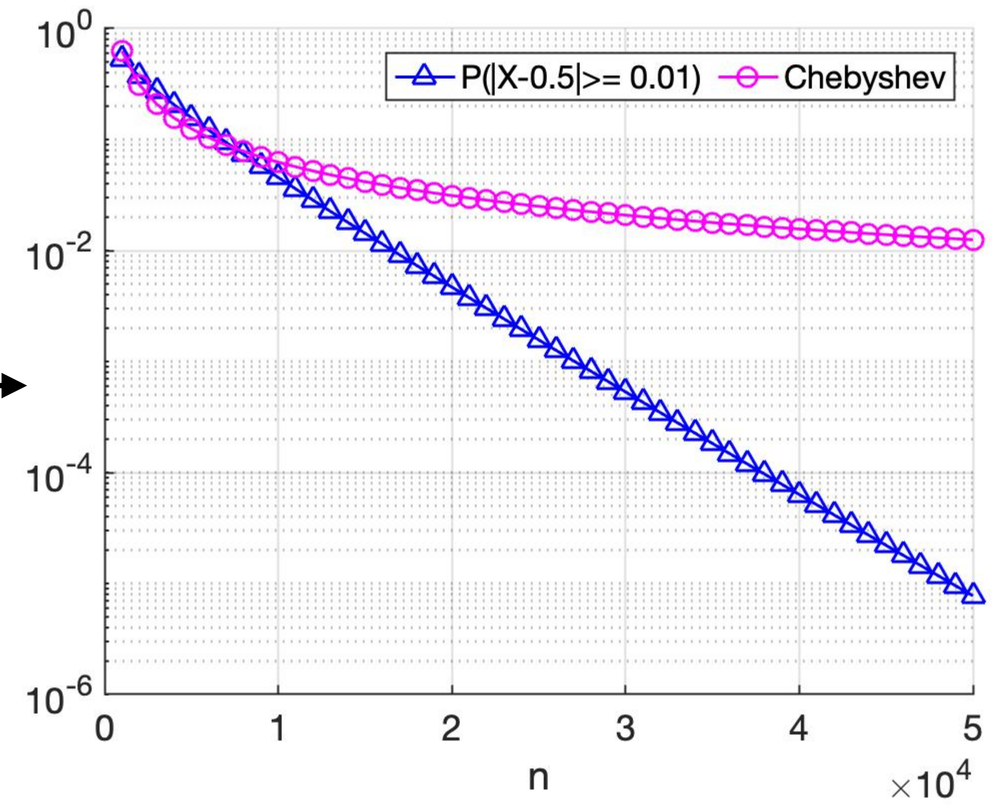
Any Issue With Chebyshev's Inequality?

- 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

$\underline{\varepsilon = 0.01}$



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

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

- Proof:

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:

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

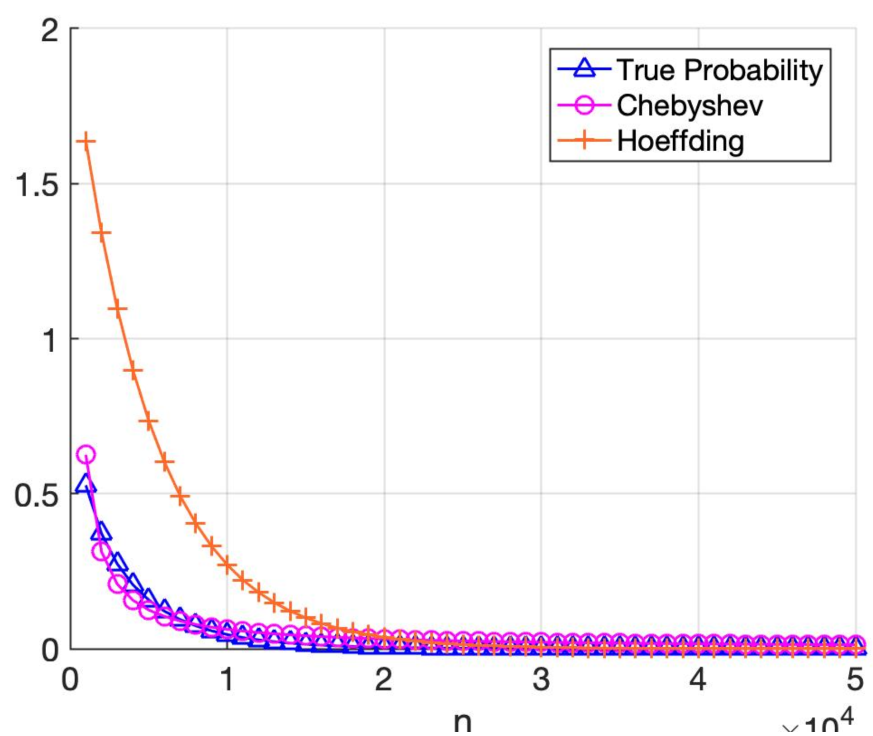
How about applying Chernoff bound to sum of independent random variables?

Hoeffding's Inequality (Formally)

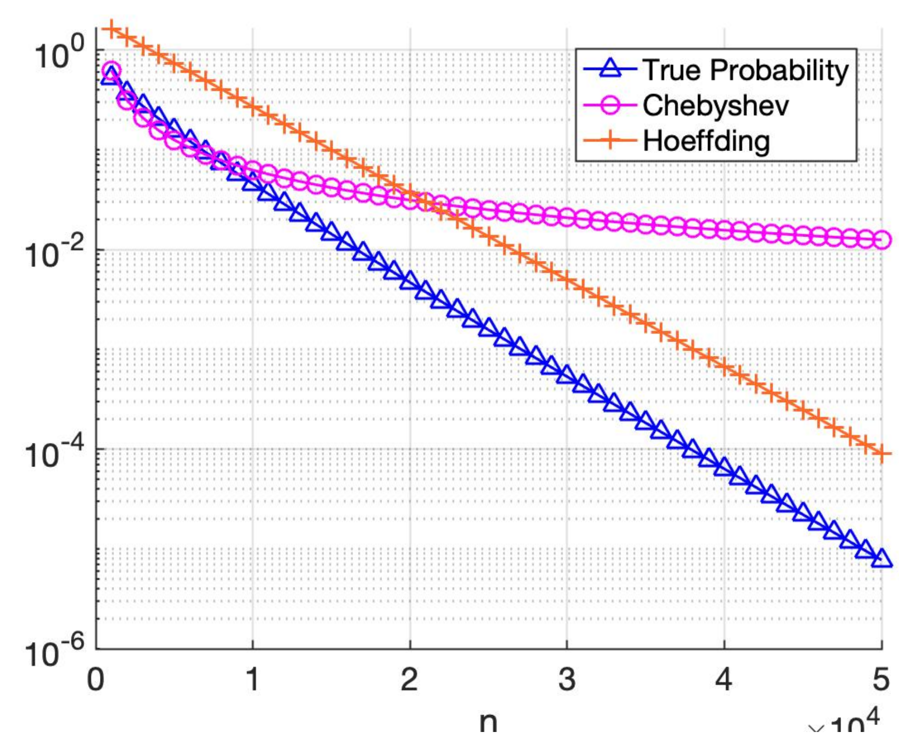
- **Hoeffding's Inequality (For Bernoulli)**: Let $X_1, \cdots, X_n$ be a sequence of i.i.d. Bernoulli random variables with parameter p. Define $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$. Then, for any $\varepsilon > 0$, we have

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

$\varepsilon = 0.01$



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
$$E[e^{tZ}] \le \exp\left(\frac{t^2(b-a)^2}{8}\right)$$

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

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

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

Next Lecture

- Law of Large Numbers

1-Minute Summary

1. Moment Generating Functions (MGF)

- Find $E[X^n]$ using MGF

2. Concentration Inequalities

- Markov’s and Chebyshev’s Inequalities
- Chernoff Bound and Hoeffding’s Inequality