# Lecture 2: Inequalities

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

March 9 & 11, 2020



# Outline

## Motivation

If you can not calculate a probability or expectation exactly, then you have three powerful strategies:

- Bounds (upper and lower bounds) on probability using inequalities.
- Approximations using limiting theorems
  - Poisson approximation: The Law of Small Numbers
  - Sample mean limit: The Law of Large Numbers
  - Normal approximation: The Central Limit Theorem
- Simulations using Monte Carlo



## Outline

1. Basic Inequalities
2. Concentration Inequalities
3. References

# Cauchy-Schwarz Inequality

**Theorem**

*For any r.v.s $X$ and $Y$ with finite variances,*

$$|E(XY)| \leq \sqrt{E(X^2)\,E(Y^2)}.$$



# Revisit Correlation

## Jensen's Inequality

If $f$ is a convex function, $0 \le \lambda_1, \lambda_2 \le 1, \lambda_1 + \lambda_2 = 1$, then for any $x_1, x_2$,

$$f(\lambda_1 x_1 + \lambda_2 x_2) \le \lambda_1 f(x_1) + \lambda_2 f(x_2).$$



## Jensen's Inequality

**Theorem**

*Let $X$ be a random variable. If $g$ is a convex function, then $E(g(X)) \ge g(E(X))$. If $g$ is a concave function, then $E(g(X)) \le g(E(X))$. In both cases, the only way that equality can hold is if there are constants $a$ and $b$ such that $g(X) = a + bX$ with probability 1.*

# Quick Examples



# Entropy

- Let $X$ be a discrete r.v. whose distinct possible values are $a_1, a_2, ..., a_n$, with probabilities $p_1, p_2..., p_n$ respectively (so $p_1 + p_2 + \cdots + p_n = 1$).
- The *entropy* of $X$ is defined as follows: $H(X) = \sum_{j=1}^{n} p_j \log_2 (1/p_j)$.
- Using Jensen's inequality, show that the maximum possible entropy for X is when its distribution is uniform over $a_1, a_2, \ldots, a_n$, i.e., $p_j = 1/n$ for all $j$.
- This makes sense intuitively, since learning the value of $X$ conveys the most information on average when $X$ is equally likely to take any of its values, and the least possible information if $X$ is a constant.

# Proof



# Kullback-Leibler Divergence

Let $\mathbf{p} = (p_1, ..., p_n)$ and $\mathbf{r} = (r_1, ..., r_n)$ be two probability vectors (so each is nonnegative and sums to 1). Think of each as a possible PMF for a random variable whose support consists of $n$ distinct values. The *Kullback-Leibler* divergence between $\mathbf{p}$ and $\mathbf{r}$ is defined as

$$D(\mathbf{p}, \mathbf{r}) = \sum_{j=1}^{n} p_j \log_2 (1/r_j) - \sum_{j=1}^{n} p_j \log_2 (1/p_j).$$

Show that the Kullback-Leibler divergence is nonnegative.

# Proof



# Norm Inequality

For a random variable $X$ whose moment of order $r > 0$ is finite, we define the following norm

$$||X_{|}|_r = (\mathbb{E}(|X|^r))^{\frac{1}{r}}.$$

- **The Holder Inequality.** Let $\frac{1}{p} + \frac{1}{q} = 1$. If $\mathbb{E}(|X|^p), \mathbb{E}(|X|^q) < \infty$, then $|\mathbb{E}(XY)| \leq \mathbb{E}|XY| \leq ||X||_p \cdot ||X||_q$.
- **The Lyapunov Inequality.** For $0 < r \leq p$, $||X||_r \leq ||X||_p$.
- **The Minkowski Inequality**. Let $p \geq 1$, $\mathbb{E}(|X|^p), \mathbb{E}(|Y|^p) < \infty$, then $||X + Y||_p \leq ||X||_p + ||Y||_p$.

## Markov's Inequality

**Theorem**

*For any r.v. $X$ and constant $a > 0$,*

$$P(|X| \geq a) \leq \frac{E|X|}{a}.$$



## Proof

## Chebyshev's Inequality

### Theorem

*Let $X$ have mean $\mu$ and variance $\sigma^2$. Then for any $a > 0$,*

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



## Proof

# Chernoff's Inequality

### Theorem

*For any r.v. $X$ and constants $a > 0$ and $t > 0$,*

$$P\left(X \geq a\right) \leq \frac{E\left(e^{tX}\right)}{e^{ta}}.$$



# Proof

## Chernoff's Technique

### Theorem

*For any r.v. $X$ and constants $a$,*

$$P(X \geq a) \leq \inf_{t>0} \frac{E\left(e^{tX}\right)}{e^{ta}}$$
$$P(X \leq a) \leq \inf_{t<0} \frac{E\left(e^{tX}\right)}{e^{ta}}.$$



## Proof

## Example: Normal Distribution

Given $X \sim \mathcal{N}(\mu, \sigma^2)$, for arbitrary constant $a > \mu$, find the Chernoff bound on $P(X > a)$.



## Solution

# Example: Poisson Distribution

Given $X \sim Pois(\lambda)$, for arbitrary constant $b > 0$, find the Chernoff bound on $P(X > b)$.



# Solution

# Outline





# Hoeffding Lemma

**Lemma**

*Let the random variable $X$ satisfy $\mathbb{E}(X) = 0$ and $a \leq X \leq b$, where $a$ and $b$ are constants. Then for any $\lambda > 0$,*

$$\mathbb{E}(e^{\lambda X}) \leq e^{\frac{1}{8}\lambda^2(b-a)^2}.$$

## Useful Analysis Tools

- Jensen's inequality: if $f$ is convex, $0 \le \lambda_1, \lambda_2 \le 1, \lambda_1 + \lambda_2 = 1$, then for any $x_1, x_2$,

$$f(\lambda_1 x_1 + \lambda_2 x_2) \le \lambda_1 f(x_1) + \lambda_2 f(x_2).$$

- Taylor's theorem or Taylor's expansion: If all the derivatives of a function $f(x)$ exist at point $a$, then for any positive integer $k$, there exist a real number $\theta$ between $a$ and $x$ such that

$$f(x) = f(a) + \cdots + \frac{f^{(k)}(a)}{k!}(x-a)^k + \frac{f^{(k+1)}(\theta)}{(k+1)!}(x-a)^{k+1}.$$



## Proof

# Proof



# Proof

# Hoeffding Bound

## Theorem

*Let the random variables $X_1, X_2, \ldots, X_n$ be independent with $E(X_i) = \mu$, $a \leq X_i \leq b$ for each $i = 1, \ldots, n$, where $a, b$ are constants. Then for any $\epsilon \geq 0$,*

$$\mathbb{P}(|\frac{1}{n}\sum_{i=1}^{n} X_i - \mu| \geq \epsilon) \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}}.$$



# Proof

# Proof



# Proof

## More General Hoeffding Bound

**Theorem**

*Let the random variables $X_1, X_2, \ldots, X_n$ be independent, with $a_k \le X_k \le b_k$ for each $k$, where $a_k, b_k$ are constants. Let $S_n = \sum_{k=1}^{n} X_k$ and let $\mu = \mathbb{E}(S_n)$. Then for any $t \ge 0$,*

$$\mathbb{P}(|S_n - \mu| \ge t) \le 2e^{-\frac{2t^2}{\sum_{k=1}^{n}(b_k - a_k)^2}}.$$



## Application: Parameter Estimation

Instead of predicting a single value for the parameter, we given an interval that is likely to contain the parameter:

**Definition**

A $1 - \delta$ confidence interval for a parameter $p$ is an interval $[\hat{p} - \epsilon, \hat{p} + \epsilon]$ such that

$$Pr\left(p \in [\hat{p} - \epsilon, \hat{p} + \epsilon]\right) \ge 1 - \delta.$$

# Application: Parameter Estimation

Tossing a coin with probability $p$ landing heads and probability $1 - p$ landing tails. $p$ is unknown and we need to estimate its value from experiments results. We toss such coin $N$ times, Let $X_i = 1$ if the $i$th result is head, otherwise 0. We estimate $p$ by using $\hat{p} = \frac{X_1 + \ldots + X_N}{N}$. Find the confidence interval for $p$.



# Solution

# Application: Monte Carlo Method for Estimation $\pi$

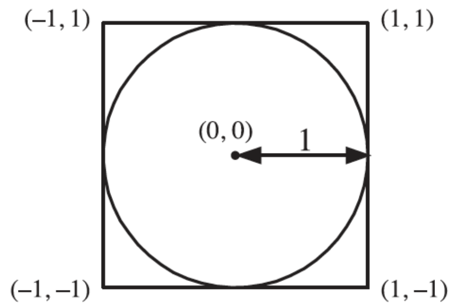



**Figure 11.1:** A point chosen uniformly at random in the square has probability $\pi/4$ of landing in the circle.



# Application: Monte Carlo Method for Estimation $\pi$

# Application: Monte Carlo Method for Estimation $\pi$



# Advanced Topics

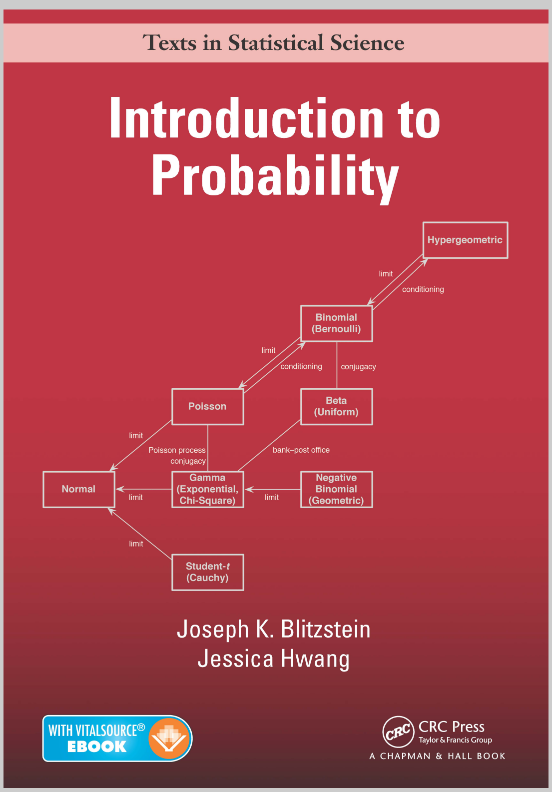
- From independent case to dependent case
- Martingale inequalities
- Logarithmic Sobolev inequalities
- Transportation method

# Outline





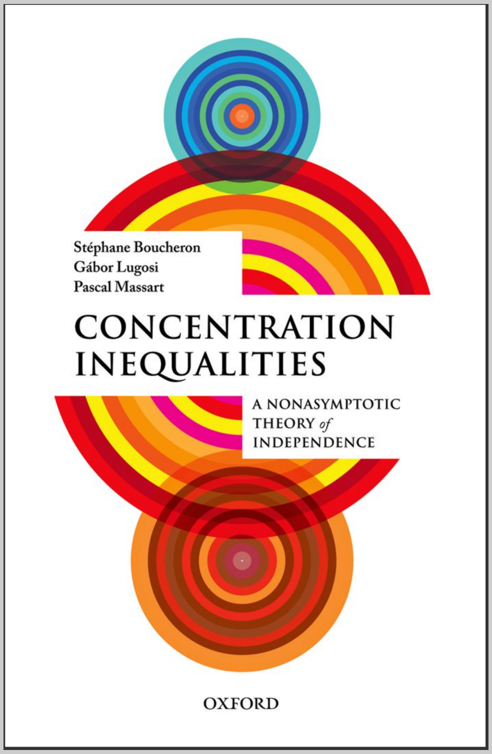
# BH




**Joseph K. Blitzstein & Jessica Hwang**

- Introduction to Probability
- Chapman & Hall/CRC, 2014.
- Chapman & Hall/CRC, 2019.
- Chapter 10

# BLM




**Stephane Boucheron & Gabor Lugosi & Pascal Massart**

- Concentration Inequalities
- Oxford, 2013