Lecture 8: Monte Carlo Methods

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

April 27, 2023

Outline

History

- Basic Monte Carlo methods: formally proposed by Stanislaw Ulam & John Von Neumann in 1940s at Los Alamos National Lab (Named after a casino in Monaco)
- Markov Chain Monte Carlo methods: formally proposed by Metropolis et al in 1950s at Los Alamos National Lab

Widely Applications

Monte Carlo methods have been used in various tasks, including

- Simulating a system and sampling from the underlying probability distribution $f(x)$
- Estimating a quantity through Monte Carlo integration

$$c = E_\pi(h(x)) = \int f(x)h(x)dx;$$

- Optimizing a target function to find its maxima or minima
- Learning parameters from a training set to optimize some loss functions

Outline

Sampling

- Assuming an algorithm is available for generating Unif(0, 1) random numbers
- Two elementary methods for generating random variables
  - Inverse-transform method: operates on the CDF
  - The acceptance-rejection method: operates on the PDF (or PMF)

Inverse Transform Method

- Given a Unif(0, 1) r.v., we can construct an r.v. with any continuous distribution we want.
- Conversely, given an r.v. with an arbitrary continuous distribution, we can create a Unif(0, 1) r.v.
- Other names:
  - probability integral transform
  - inverse transform sampling
  - the quantile transformation
  - the fundamental theorem of simulation

Inverse Transform Method: Recall

Theorem

Let F be a CDF which is a continuous function and strictly increasing on the support of the distribution. This ensures that the inverse function F^{-1} exists, as a function from $(0,1)$ to $\mathbb{R}$. We then have the following results.

1. *Let $U \sim \mathrm{Unif}(0,1)$ and $X = F^{-1}(U)$. Then X is an r.v. with CDF F.*
2. *Let X be an r.v. with CDF F. Then $F(X) \sim \mathrm{Unif}(0,1)$.*

Algorithm Inverse-Transform Method: PDF Case

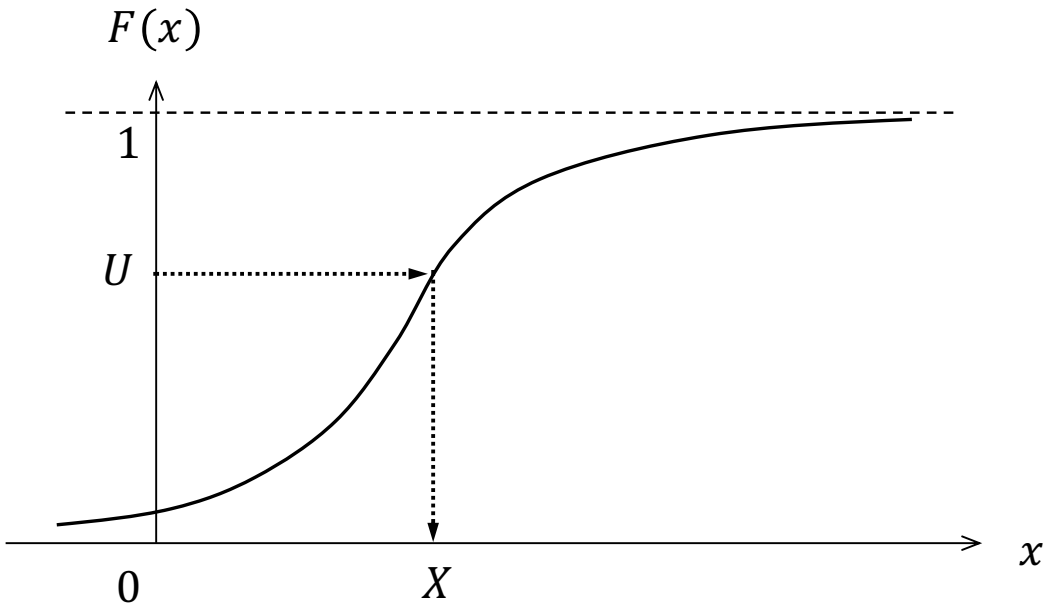
input: Cumulative distribution function F.

output: Random variable X distributed according to F.

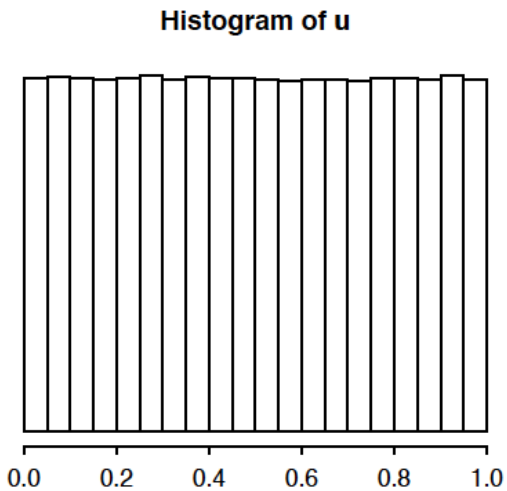
1: Generate U from $Unif(0, 1)$.

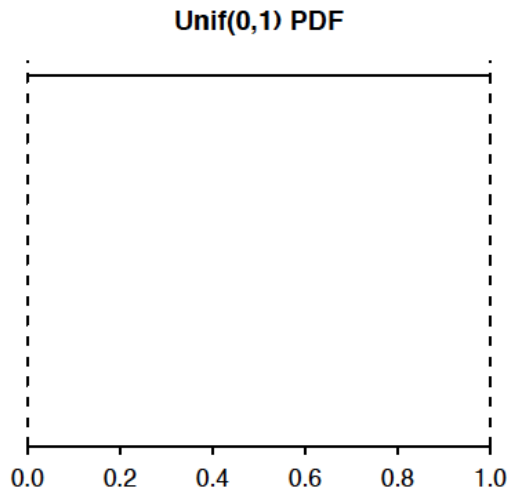
2: $X \leftarrow F^{-1}(U)$

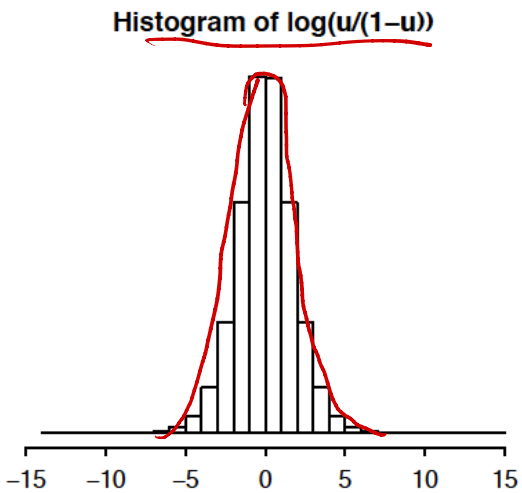
3: **return** X

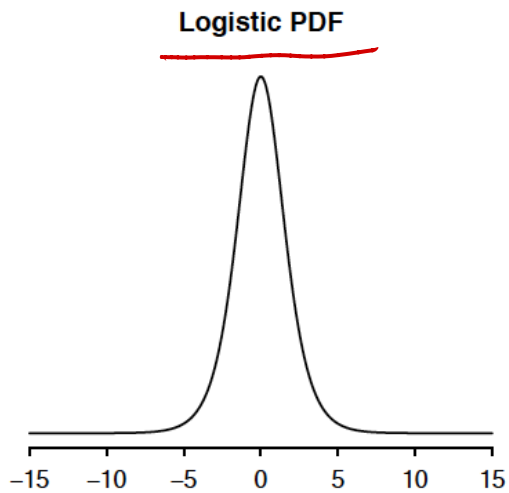


Histogram & PDF: Example









Box-Muller: Recall

Let $U \sim \mathrm{Unif}(0, 2\pi)$, and let $T \sim \mathrm{Expo}(1)$ be independent of U. Define $X = \sqrt{2T}\cos U$ and $Y = \sqrt{2T}\sin U$. Find the joint PDF of (X, Y). Are they independent? What are their marginal distributions?

Box-Muller Method

Algorithm Normal Random Variable Generation: Box-Muller Approach

output: Independent standard normal random variables X and Y.

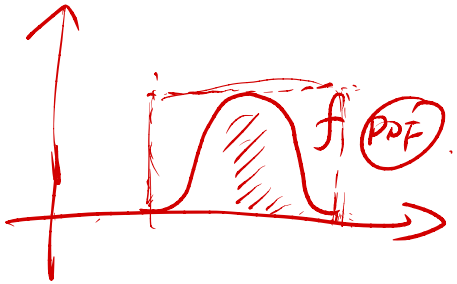
1: Generate two independent random variables, U_1 and U_2, from $Unif(0, 1)$.

2: $X \leftarrow (-2 \ln U_1)^{1/2} \cos(2\pi U_2)$

3: $Y \leftarrow (-2 \ln U_1)^{1/2} \sin(2\pi U_2)$

4: **return** X, Y

Acceptance-Rejection Method



- Suppose one can generate samples (relatively easily) from PDF g
- How can random samples be simulated from PDF f?

Algorithm Acceptance-Rejection Algorithm

Let c denote a constant such that $c \geq \sup_y \frac{f(y)}{g(y)}$. Then:

Step 1: Generate $Y \sim g$.

Step 2: Generate $U \sim Unif[0,1]$.

Step 3: If $U \leq \frac{f(Y)}{c \cdot g(Y)}$, set $X = Y$. Otherwise go back to step 1.

Acceptance-Rejection Method

event $A = "U \le \frac{f(Y)}{c \cdot g(Y)}"$

(i) the PDF of generated r.v.s.. $f_Y(y|A) = \frac{P(A|Y=y)}{P(A)} \cdot f_Y(y)$

① $P(A|Y=y) = P\left(U \le \frac{f(Y)}{c \cdot g(Y)} \,\middle|\, Y=y\right) = P\left(U \le \frac{f(y)}{c \cdot g(y)} \,\middle|\, Y=y\right)$

Theorem

(i) The random variable generated by the Acceptance-Rejection method has the desired PDF f.

(ii) The number of iterations of the algorithm that are needed is a first-success random variable with mean c.

(iii) $c \ge 1$

≤ 1

$\text{Unif}(0,1)$

$= P\left(U \le \frac{f(y)}{c \cdot g(y)}\right) = \frac{f(y)}{c \cdot g(y)}$

$P(U \le x) = x, \quad 0 \le x \le 1$ CDF of $\text{Unif}(0,1)$

Proof

② $P(A) \overset{\text{LOTP}}{=\!=} \int \underline{P(A|Y=y)} \cdot g(y)\,dy = \int \frac{f(y)}{C \cdot g(y)} \cdot g(y)\,dy$

$= \int \frac{f(y)}{C}\,dy = \frac{1}{C}\underline{\int f(y)\,dy} = \frac{1}{C} \le 1 \quad (\underline{C \ge 1})$

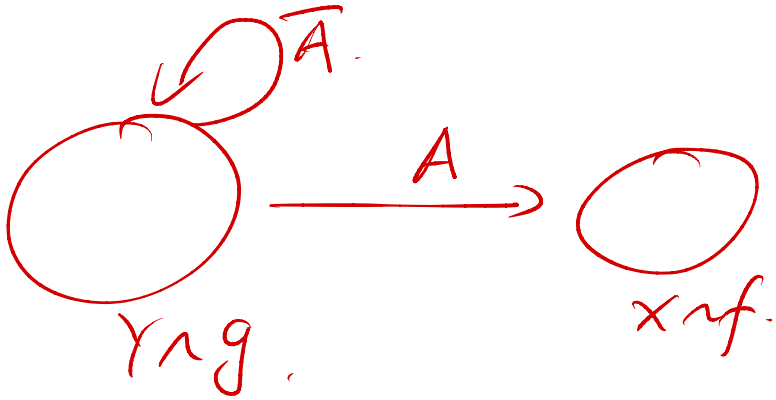
③ $f_Y(y|A) = \frac{P(A|Y=y)}{P(A)} \cdot f_Y(y)$

$= \frac{\frac{f(y)}{C \cdot g(y)}}{\frac{1}{C}} \cdot g(y) = \underline{f(y)}, \forall y.$

$X \sim f$

Proof

(ii)



# of iterations $\sim$ $Z \sim$ FS(p)

$p = p(A) = \frac{1}{C} \leq 1$.

$\Rightarrow E(Z) = \frac{1}{p} = C$

Example: Beta Distribution

① $X \sim \text{Beta}(2,4)$. PDF of X: $f(x) = 20x(1-x)^3$, $0<x<1$.

② Acceptance-Rejection. $Y \sim \text{unif}(0,1)$ $\Rightarrow g(x) = 1$

$$c \geq \sup_y \frac{f(y)}{g(y)} = \sup_y 20y(1-y)^3 \Rightarrow y^* = \frac{1}{4}, \quad 0<x<1.$$

- Use the Acceptance-Rejection Method to generate a random variable with distribution $Beta(2,4)$

$$\Rightarrow c \geq \frac{135}{64}$$

pick $c = \frac{135}{64}$.

$$\Rightarrow \frac{f(y)}{c \cdot g(y)} = \frac{20y(1-y)^3}{\frac{135}{64} \cdot 1} = \frac{256}{27} y(1-y)^3, \quad 0<y<1$$

step 1: Generate $Y \sim \text{unif}(0,1)$

step 2: ---- $U \sim \text{unif}(0,1)$

3: If $U \leq \frac{f(Y)}{c \cdot g(Y)} = \frac{256}{27} Y(1-Y)^3$, set $X = Y$.

otherwise

Solution

Example: Normal Distribution

① $Z \sim N(0,1)$ $(-\infty, +\infty)$

② $X = |Z|$ $(0, +\infty)$

proposal distribution: $\text{Expo}(\lambda)$ $(0,+\infty)$ / $\text{Expo}(1)$

if we obtain X, $Z = \begin{cases} X & \text{w.p. } 0.5 \\ -X & \text{w.p. } 0.5 \end{cases}$, $Z \sim N(0,1)$

- Use the Acceptance-Rejection Method to generate a random variable with distribution $N(0,1)$

③ $X = |Z|$; $f_X(x) = \frac{2}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}$, $0 < x < \infty$.

choose $g \sim \text{Expo}(1)$, $g(x) = e^{-x}$, $0 < x < \infty$.

Now $\frac{f(x)}{g(x)} = \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}x^2 + x}$

$c \geq \sup_x \frac{f(x)}{g(x)} = \frac{f(1)}{g(1)} = \sqrt{\frac{2e}{\pi}}$ ($x^* = 1$)

choose $c = \sqrt{\frac{2e}{\pi}}$ $\Rightarrow \frac{f(x)}{c \cdot g(x)} = \exp\left\{-\frac{1}{2}(x-1)^2\right\}$

Solution

④

Step 1: Generate $Y \sim \text{Expo}(1)$

2: Generate $U \sim \text{unif}(0,1)$

3: If $U \le \exp\{-\frac{1}{2}(Y-1)^2\}$, set $X = Y$ (Accept the sample)

Otherwise, reject it, return to step 1.

4: Generate $U' \sim \text{unif}(0,1)$

$$Z = \begin{cases} X & \text{if } U' \le \frac{1}{2} \\ -X & \text{if } U' > \frac{1}{2} \end{cases}$$

$\Rightarrow Z \sim N(0,1)$ | Compare Box-Muller.

Outline

Monte Carlo Integration

- We want to compute the integral of a function in a very high dimensional space Ω:

$$c = E_f(h(\boldsymbol{x})) = \int f(\boldsymbol{x})h(\boldsymbol{x})d\boldsymbol{x};$$

- This is often estimated by Monte Carlo integration. By drawing n samples (empirical samples) from $f(\mathbf{x})$:

$$x_1, x_2, \ldots, x_n \sim f(\boldsymbol{x})$$

$h(x_1) \ldots h(x_n)$

- One can estimate c by the sample mean:

$$\frac{1}{n}\sum_{j=1}^{n} h(\boldsymbol{x}_j)$$

Example: π as An Integration

Evaluate the integration

$$\int_0^1 \frac{4}{1+x^2} dx.$$

$$\int_0^1 \frac{4}{1+x^2} \cdot 1 \, dx = E_f[h(x)]$$

- $h(x) = 4/(1+x^2)$, $f(x) = 1, 0 < x < 1$.
- $X_1, \ldots, X_n$: samples from Unif$(0,1)$.
- Monte-Carlo Integration:

$$h(x_i) = \frac{4}{1+x_i^2}$$

$$\int_0^1 \frac{4}{1+x^2} dx \approx \frac{1}{n} \sum_{i=1}^{n} \frac{4}{1+X_i^2}.$$

Useful Tools: Importance Sampling

- Standard Monte Carlo integration is great if you can sample from the target distribution (i.e. the desired distribution)
- But what if you can't sample from the target?
- **Importance Sampling**: draw the sample from a proposal distribution and re-weight the integral using importance weights so that the correct distribution is targeted

Importance Sampling

$$I = \int h(y) f(y) dy$$

$E_f[h(Y)]$

- h is some function and f is the PDF of random variable Y
- When the PDF f is difficult to sample from, importance sampling can be used
- Rather than sampling from f , you specify a different PDF g, as the proposal distribution.

$$I = \int h(y) f(y) dy = \int h(y) \frac{f(y)}{g(y)} g(y) dy = \int \frac{h(y) f(y)}{g(y)} g(y) dy$$

Importance Sampling

$$I = E_f[h(Y)] = \int \frac{h(y)f(y)}{g(y)} g(y)dy = E_g\left[\frac{h(Y)f(Y)}{g(Y)}\right]$$

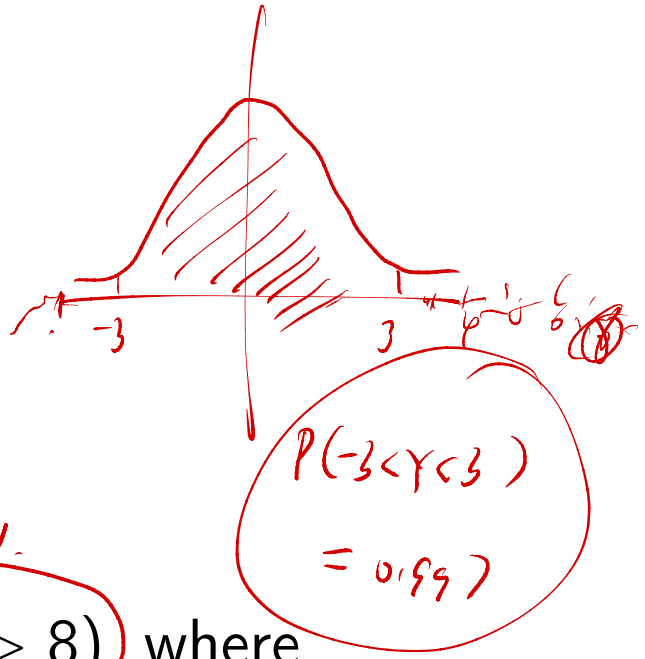
- Hence, given an iid sample $Y_1, \ldots, Y_n$ from PDF g, our estimator of I becomes

$$\hat{I} = \frac{1}{n}\sum_{j=1}^{n} \frac{h(Y_j)f(Y_j)}{g(Y_j)}$$

Example: Gaussian Tail Probability

Key : probability $\longleftrightarrow$ Indicator Variable $\longleftrightarrow$ Expectation

$$I(A) = \begin{cases} 1 & \text{if } A \text{ occurs.} \\ 0 & \text{otherwise} \end{cases}$$



$$c = P(Y>8) = E[I(Y>8)] = \int I(Y>8) f(y) dy.$$

Evaluate the probability of rare event $c = \mathbb{P}(Y > 8)$, where $Y \sim N(0,1)$.

$f \sim N(0,1)$, $h(y) = I(y>8)$.

(1) $\hat{c} = \frac{1}{n} \sum_{j=1}^{n} I(Y_j > 8)$ $\quad Y_1, \dots, Y_n \sim N(0,1)$

$= 0$.

(2) $g \sim N(8,1)$

Solution

② Importance Sampling : Choose $g \sim N(8,1)$.

$Y_1, \dots Y_n \sim g\ (N(8,1))$.

$$\hat{c}_1 = \frac{1}{n} \sum_{j=1}^{n} \frac{h(Y_j) f(Y_j)}{g(Y_j)} = \frac{1}{n} \sum_{j=1}^{n} I(Y_j > 8) \cdot \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} Y_j^2}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(Y_j - 8)^2}}$$

$$= \frac{1}{n} \sum_{j=1}^{n} I(Y_j > 8) \cdot e^{-8Y_j + 32}$$

6.25×10^{-16}

Outline

Sample Mean: Recall

① $E(\bar{X}_n) = \frac{1}{n}\sum_{j=1}^{n} E(X_j) = \frac{1}{n} \cdot n\mu = \mu$

unbiased estimation

n samples from the same distribution

Definition

Let $X_1, \ldots, X_n$ be i.i.d. random variables with finite mean μ and finite variance σ^2. The *sample mean* $\bar{X}_n$ is defined as follows:

$$\bar{X}_n = \frac{1}{n}\sum_{j=1}^{n} X_j.$$

The sample mean $\bar{X}_n$ is itself an r.v. with mean μ and variance σ^2/n.

② $Var(\bar{X}_n) = \frac{1}{n^2}\sum_{j=1}^{n} Var(X_j) = \frac{1}{n^2} \cdot n\sigma^2 = \frac{1}{n}\sigma^2$

$n \to \infty \Rightarrow Var(\bar{X}_n) \to 0$; $\bar{X}_n \to \mu$.

Strong Law of Large Numbers (SLLN)

$\frac{1}{n}(X_1 + \cdots + X_n) \xrightarrow[\text{w.p. 1}]{n\to\infty} E[X] = \mu$.

Theorem

The sample mean $\bar{X}_n$ converges to the true mean μ pointwise as $n \to \infty$, with probability 1. In other words, the event $\bar{X}_n \to \mu$ has probability 1.

$\frac{1}{n}[f(X_1) + \cdots + f(X_n)] \xrightarrow[\text{w.p. 1}]{n\to\infty} E[f(X)]$

$X_1, \cdots X_n, X \sim F$.

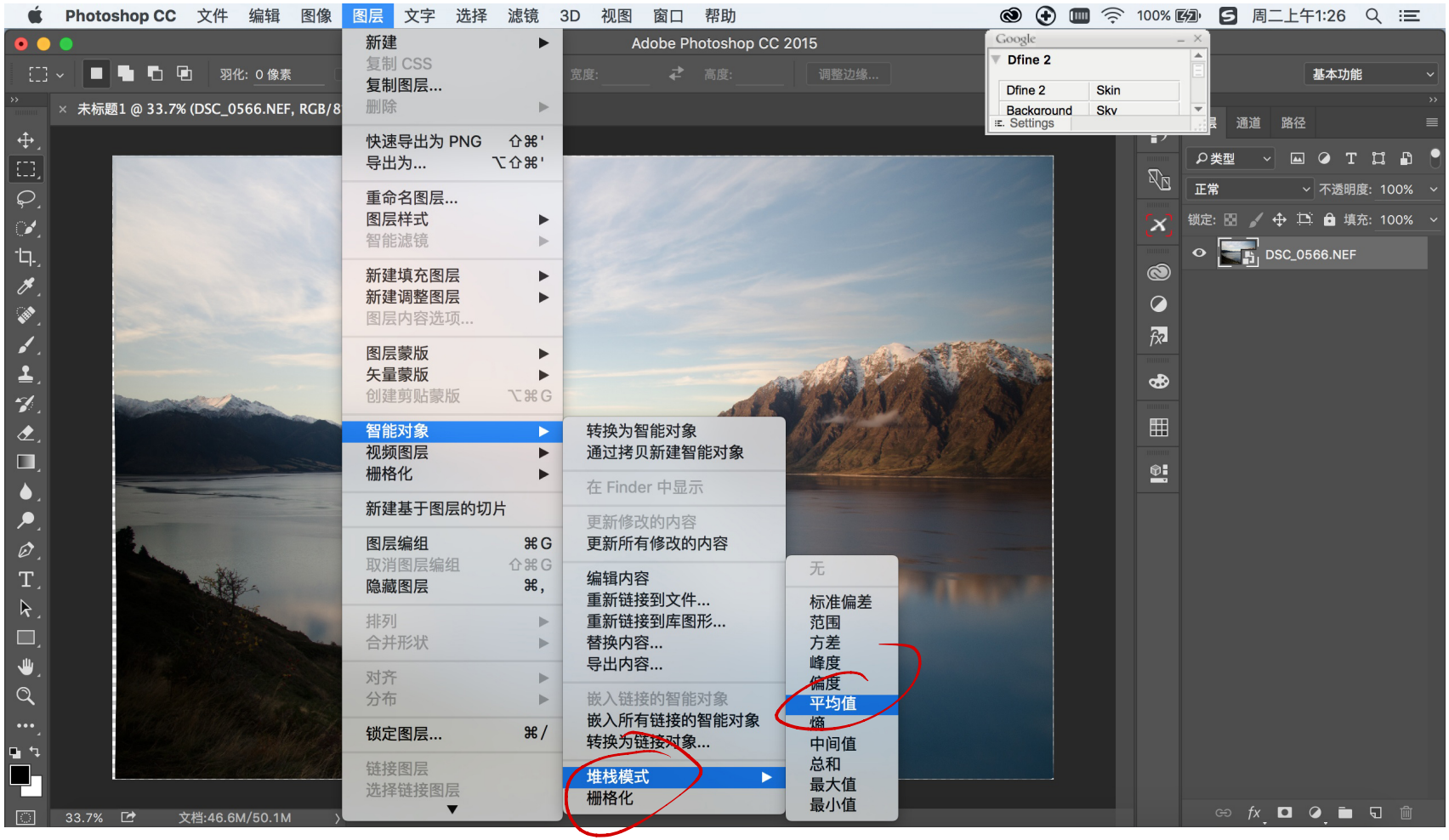
Weak Law of Large Numbers (WLLN)

Theorem

For all $\epsilon > 0$, $P\left(\left|\bar{X}_n - \mu\right| > \epsilon\right) \to 0$ *as* $n \to \infty$. *(This form of convergence is called convergence in probability).*

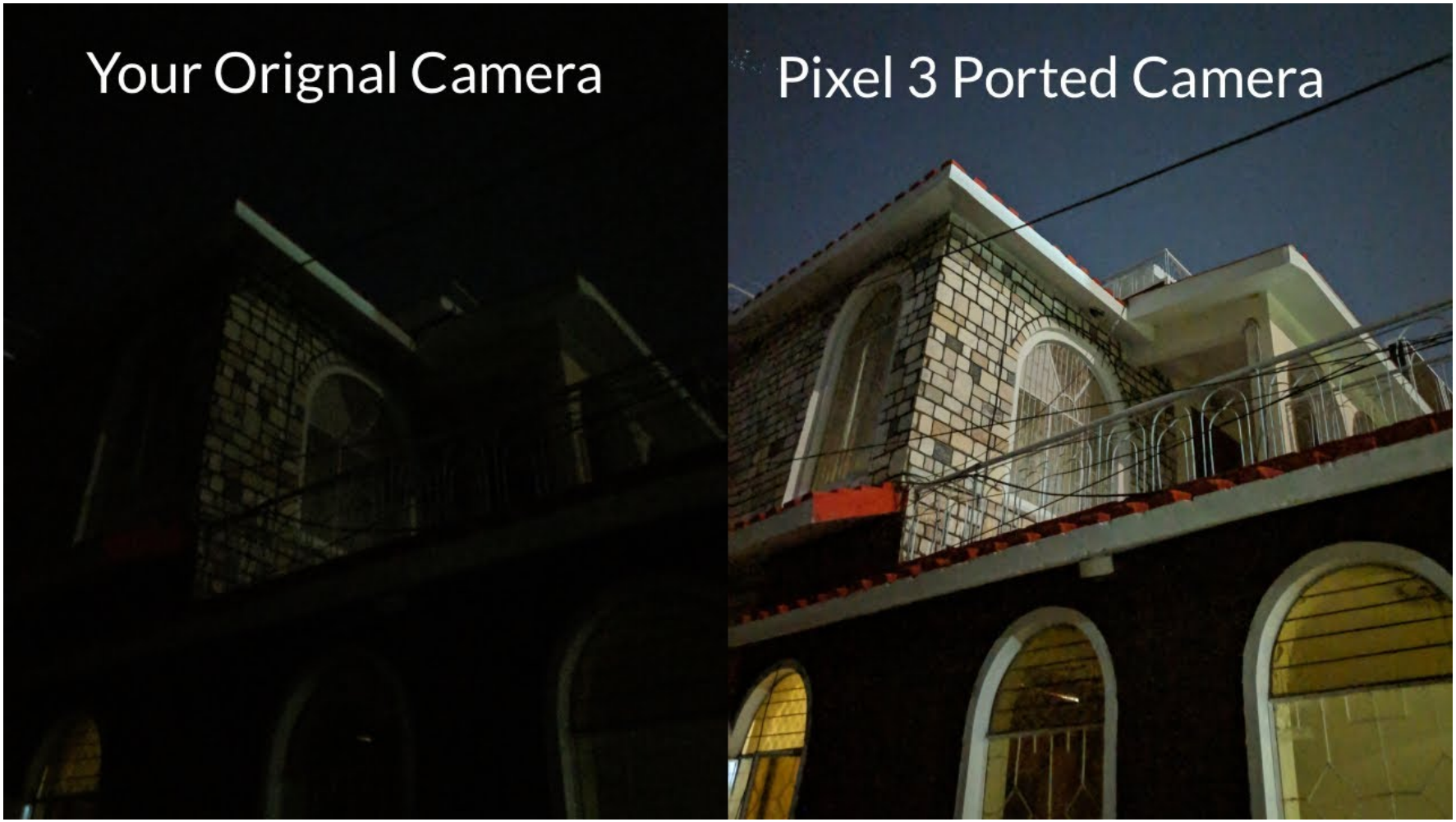
Widely Applications: Photo Stacking with PC

Widely Applications: Photo Stacking with PC



Widely Applications: Night Model with Smart Phone

Widely Applications: Photo Stacking with Smart Phone



Widely Applications: Photo Stacking with Smart Phone

Widely Applications: Photo Stacking with Smart Phone

Outline

Infinite Samples.

Finite Samples

$0 \leq P(|\bar{X}_n - \mu| > \varepsilon) \leq 0$

Cauchy-Schwarz Inequality: Recall

$\rho \in [-1, 1]$.

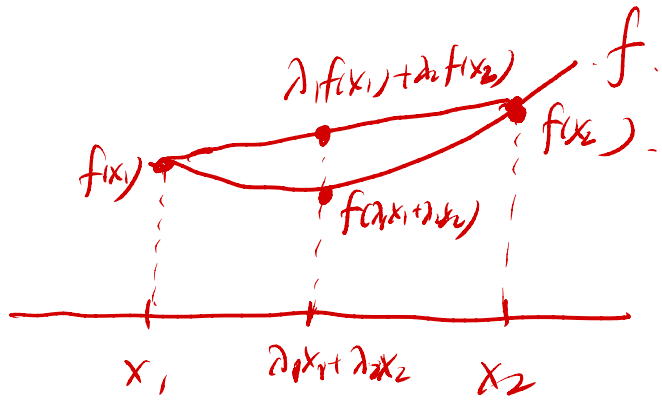
correlation

Theorem

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

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

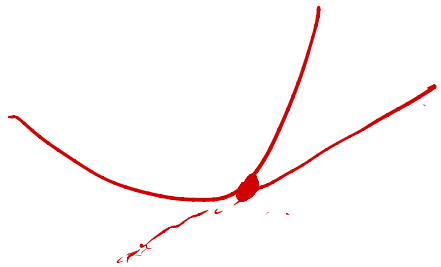
Jensen's Inequality



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

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

Jensen's Inequality



Theorem

Let X be a random variable. If g is a convex function, then $E(g(X)) \geq g(E(X))$. If g is a concave function, then $E(g(X)) \leq g(E(X))$. 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

g is convex ; $E[g(x)] \geq g[E(x)]$

Concave ; $\leq$

1° $g(x)=x^2$, $x \in R$. ; Convex ; $E[x^2] \geq (E(x))^2$ ✓

$Var(x) = E(x^2) - (Ex)^2 \geq 0$

2° $g(x)=\frac{1}{x}$, $x>0$; Convex ; $E(\frac{1}{x}) \geq \frac{1}{E(x)}$ ✓

3° $g(x)=\log x$, $x>0$; Concave ; $E(\log x) \leq \log(E(x))$ ✓

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

① Construct a r.v. Y s.t.

$$Y = \begin{cases} \frac{1}{p_1} & \text{w.p. } p_1 \\ \frac{1}{p_2} & \text{w.p. } p_2 \\ \vdots & \\ \frac{1}{p_n} & \text{w.p. } p_n \end{cases}$$

$$E(Y) = \frac{1}{p_1} \cdot p_1 + \frac{1}{p_2} \cdot p_2 + \cdots + \frac{1}{p_n} \cdot p_n = \underline{n}$$

② $H(X) = \sum_{j=1}^{n} p_j \log_2 \frac{1}{p_j} = \underline{E[\log_2 Y]} \le \log_2(E[Y]) = \log_2 n$

$\forall (p_1, \cdots p_n)$ $H(X) \le \log_2 n \Rightarrow \max_{p_1, \cdots p_n} H(X) \underline{\le \log_2 n}$

③ $\max_{p_1, \cdots p_n} H(X) \underline{\ge \log_2 n}$ [$X \sim \underline{\text{Dunif}(\frac{1}{n})}$, $p_1 = p_2 = \cdots = \frac{1}{n}$. $H(X) = \log_2 n$]

④ $\Rightarrow \max_{p_1, \cdots p_n} H(X) = \log_2 n$, $X^* \sim \text{Dunif}(\frac{1}{n})$.

Kullback-Leibler Divergence

Relative Entropy

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

$D(\mathbf{p}, \mathbf{r}) \neq_n D(\mathbf{r}, \mathbf{p})$

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

Show that the Kullback-Leibler divergence is nonnegative.

Proof

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

$= \sum_{j=1}^{n} p_j \log_2 \frac{p_j}{r_j} = -\sum_{j=1}^{n} p_j \log_2 \frac{r_j}{p_j}$.

② Construct a r.v. Y s.t.

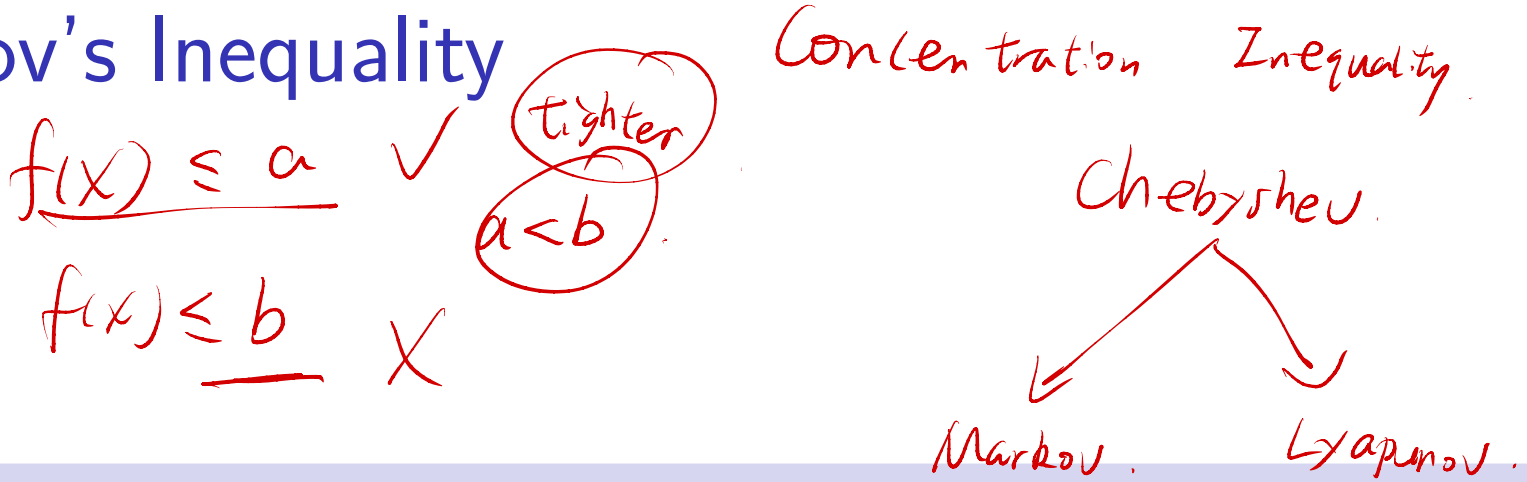
$P(Y = \frac{r_j}{p_j}) = p_j \quad , j = 1, 2 \cdots n$.

$\Rightarrow E(Y) = \sum_{j=1}^{n} \frac{r_j}{p_j} \cdot p_j = \sum_{j=1}^{n} r_j = 1$.

③ $D(p,r) = -\sum_{j=1}^{n} p_j \log_2 \frac{r_j}{p_j} = -E[\log_2 Y]$

$\geq -\log_2(E(Y)) = -\log_2 1 = 0$

Markov's Inequality



Theorem

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

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

Proof

$P(|X| \geq a) \leq \frac{1}{a} E|X|. \quad a > 0$

① $Y = \frac{|X|}{a} \geq 0$; $I(Y \geq 1) \leq Y$. → Indicator of "$Y \geq 1$"

if $Y \geq 1$	LHS	RHS
	1	$\leq Y$ ✓
$0 \leq Y < 1$	LHS	RHS
	0	$\leq Y$ ✓

$\Rightarrow \quad E[I(Y \geq 1)] \leq E[Y]$

$\Rightarrow \quad P(Y \geq 1) \leq E[Y] = \frac{1}{a} E|X|.$

↓

$P(\frac{|X|}{a} \geq 1)$

↓

$P(|X| \geq a)$

② $P(|X| \geq a) \leq \frac{1}{a} E(|X|).$

Chebyshev's Inequality

① $P(|X-\mu| \geq a) = P(|X-\mu|^2 \geq a^2)$

Markov's Inequality

$\leq \frac{1}{a^2} E(|X-\mu|^2) = \frac{1}{a^2}\sigma^2$

Theorem

Let X have mean μ and variance σ^2. Then for any $a > 0$,

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

$\uparrow a \quad \downarrow \frac{1}{a^2}\sigma^2$

Proof

Application:

$\bar{X}_n$: sample mean ; $E(\bar{X}_n) = \mu$; $Var(\bar{X}_n) = \frac{1}{n}\sigma^2$

$$0 \le P(|\bar{X}_n - \mu| \ge a) \le \frac{1}{a^2} Var(\bar{X}_n - \mu) = \frac{1}{a^2} Var(\bar{X}_n)$$

$$= \frac{1}{a^2} \cdot \frac{1}{n}\sigma^2$$

$n \to \infty$

$\downarrow$

0

$$\Rightarrow \lim_{n\to\infty} P(|\bar{X}_n - \mu| \ge a) = 0$$

WLLN

Chernoff's Inequality

$\forall\, t > 0$;

$$P(X \geq a) = P(e^{tX} \geq e^{ta})$$

$$\leq \frac{E[e^{tX}]}{e^{ta}}.$$

Theorem

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

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

$E(e^{tX})$ → MGF

Proof

Chernoff's Technique

① $P(X \geq a) \leq \frac{E[e^{tX}]}{e^{ta}}$ $\forall t > 0$

$f(t)$

Theorem

For any r.v. X and constants a,

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

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

② $P(X \leq a)$, $\forall t < 0$ $\Rightarrow P(tX \geq ta)$

$= P(tX \geq ta)$

$\Rightarrow P(e^{tX} \geq e^{ta})$

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

Proof

Example: Normal Distribution

① MGF of X: $M_{G(X)} = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

② $P(X > a) \le \inf_{t>0} \frac{E[e^{tx}]}{e^{ta}} = \inf_{t>0} f(t)$

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

③ $f(t) = \frac{1}{e^{ta}} E[e^{tx}] = \frac{1}{e^{ta}} M_{G(X)} = e^{\mu t + \frac{1}{2}\sigma^2 t^2 - at}$

$\mu t + \frac{1}{2}\sigma^2 t^2 - at = \frac{1}{2}\sigma^2 \left[\left(t + \frac{\mu - a}{\sigma^2} \right)^2 - \frac{(\mu - a)^2}{\sigma^4} \right]$

$\Rightarrow t^* = \frac{a - \mu}{\sigma^2} > 0$

④ $P(X > a) \le f(t^*) = e^{-\frac{(a-\mu)^2}{2\sigma^2}}$

$P(X - \mu > a - \mu) =$

Solution

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 $\lambda > 0$,

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

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 $\epsilon \geq 0$,

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

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

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

Application: Parameter Estimation

① $X_i \sim Bern(p)$. $E(X_i) = p$; $0 \le X_i \le 1$ ($X_i = 1$ or 0) ($a=0$; $b=1$)

$E(\hat{p}) = p$; LLN : $N \to \infty$; $\hat{p} \to p$.

② $Pr(|\hat{p} - p| \ge \varepsilon) = Pr(|\frac{1}{N}\sum_{i=1}^{N} X_i - p| \ge \varepsilon)$ By Hoeffding Bound.

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 ith 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.

$\le 2e^{-\frac{2N\varepsilon^2}{(b-a)^2}} = 2e^{-2N\varepsilon^2}$

③ $2e^{-2N\varepsilon^2} = \delta \Rightarrow \varepsilon = \sqrt{\frac{\ln(\frac{2}{\delta})}{2N}}$

Solution

④ $Pr(|\hat{p}-p| \geq \varepsilon) \leq \delta$

$\Rightarrow Pr(|\hat{p}-p| < \varepsilon) > 1-\delta$

$\Rightarrow Pr(p \in (\hat{p}-\varepsilon, \hat{p}+\varepsilon)) > 1-\delta$

⑤ $(\hat{p}-\varepsilon, \hat{p}+\varepsilon)$

$$\left(\frac{1}{N}\sum_{i=1}^{N} x_i - \sqrt{\frac{\ln(\frac{2}{\delta})}{2N}}, \frac{1}{N}\sum_{i=1}^{N} x_i + \sqrt{\frac{\ln(\frac{2}{\delta})}{2N}}\right)$$

$1-\delta$ Confidence Interval.

Application Example: Monte Carlo Method for Estimation π

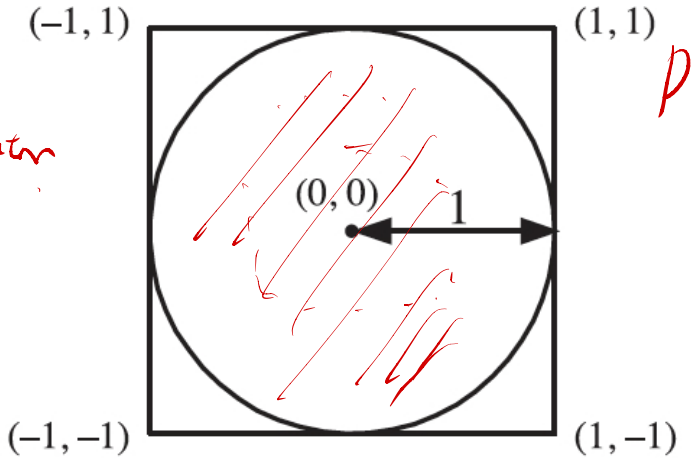
① square $[-1,1] \times [-1,1]$

Circle : $\{(x,y) : x^2+y^2 \le 1\}$

② $\pi = 4\Pr(A)$ → indicator

$= 4 \cdot E(I_A)$

$\approx 4\left(\frac{1}{n}\sum_{i=1}^{n} Z_i\right)$



A = "point landing in the circle"

$\Pr(A) = \frac{\text{area(circle)}}{\text{area(square)}} = \frac{\pi}{4}$

- A point chosen uniformly at random in the square has probability $\pi/4$ of landing in the circle

③ $Z_i = I\{\text{"the } i\text{th point landing in the circle"}\}$

Example: Monte Carlo Method for Estimation π

$P(Z_i=1) = \frac{\pi}{4}$; $E(Z_i) = \frac{\pi}{4}$. $0 \le Z_i \le 1$ $Z_i = 1$ or 0

④ Suppose we generate n points.

$W \triangleq \frac{1}{n}\sum_{i=1}^{n} Z_i$ $E(W) = \frac{\pi}{4}$.

⑤ $\hat{\pi} = 4W$.

$n \to \infty$, $\hat{\pi} \to \pi$

Finite n ;

$Pr(|\hat{\pi} - \pi| \ge \varepsilon) = Pr(|4W - \pi| \ge \varepsilon)$

$= Pr(|W - \frac{\pi}{4}| \ge \frac{\varepsilon}{4}) = Pr(|\frac{1}{n}\sum_{i=1}^{n} Z_i - \frac{\pi}{4}| \ge \frac{\varepsilon}{4})$

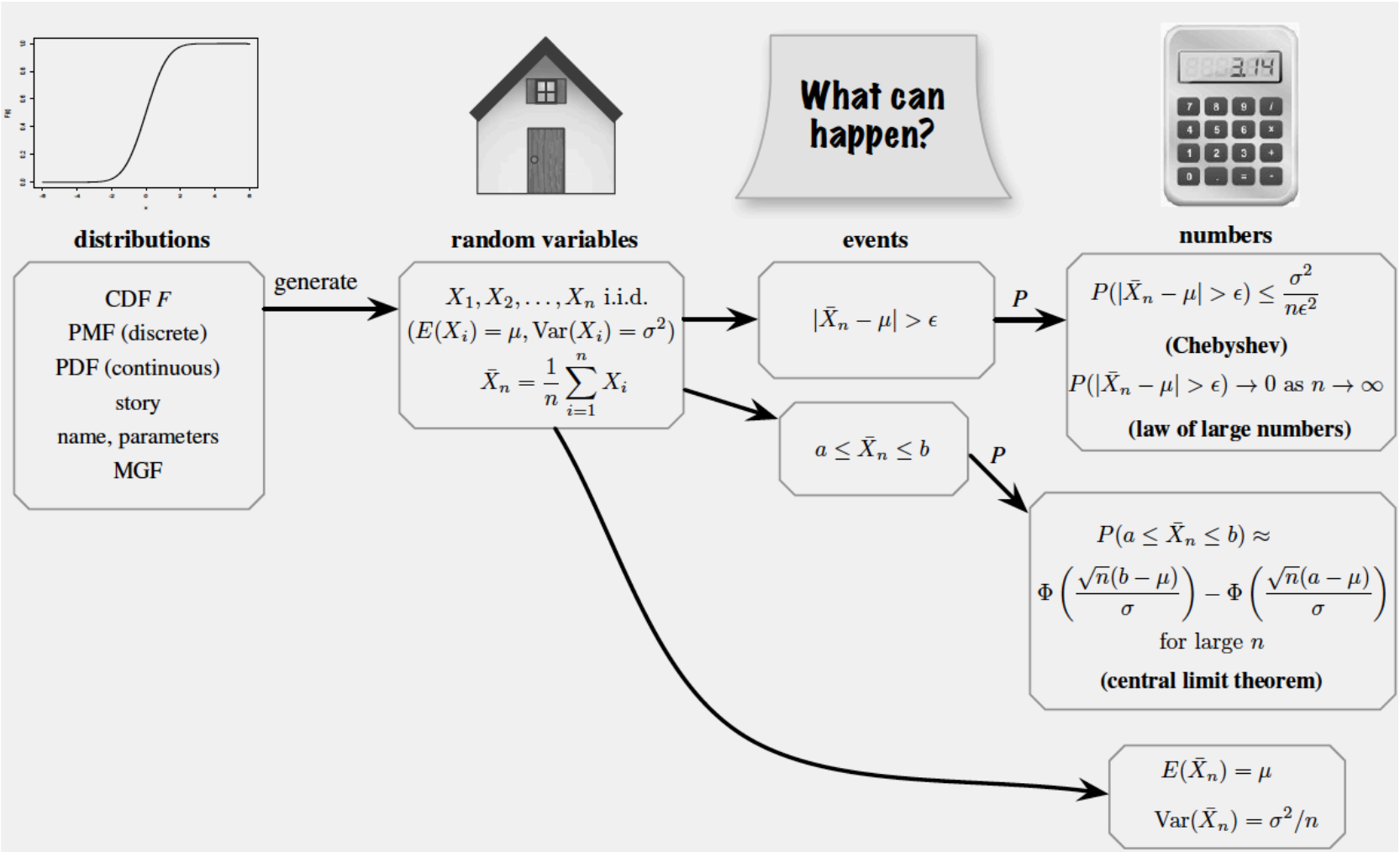
Hoeffding Bound

$\le 2e^{-\frac{2n(\frac{\varepsilon}{4})^2}{(1-0)^2}} = 2e^{-\frac{1}{8}n\varepsilon^2} = \delta$

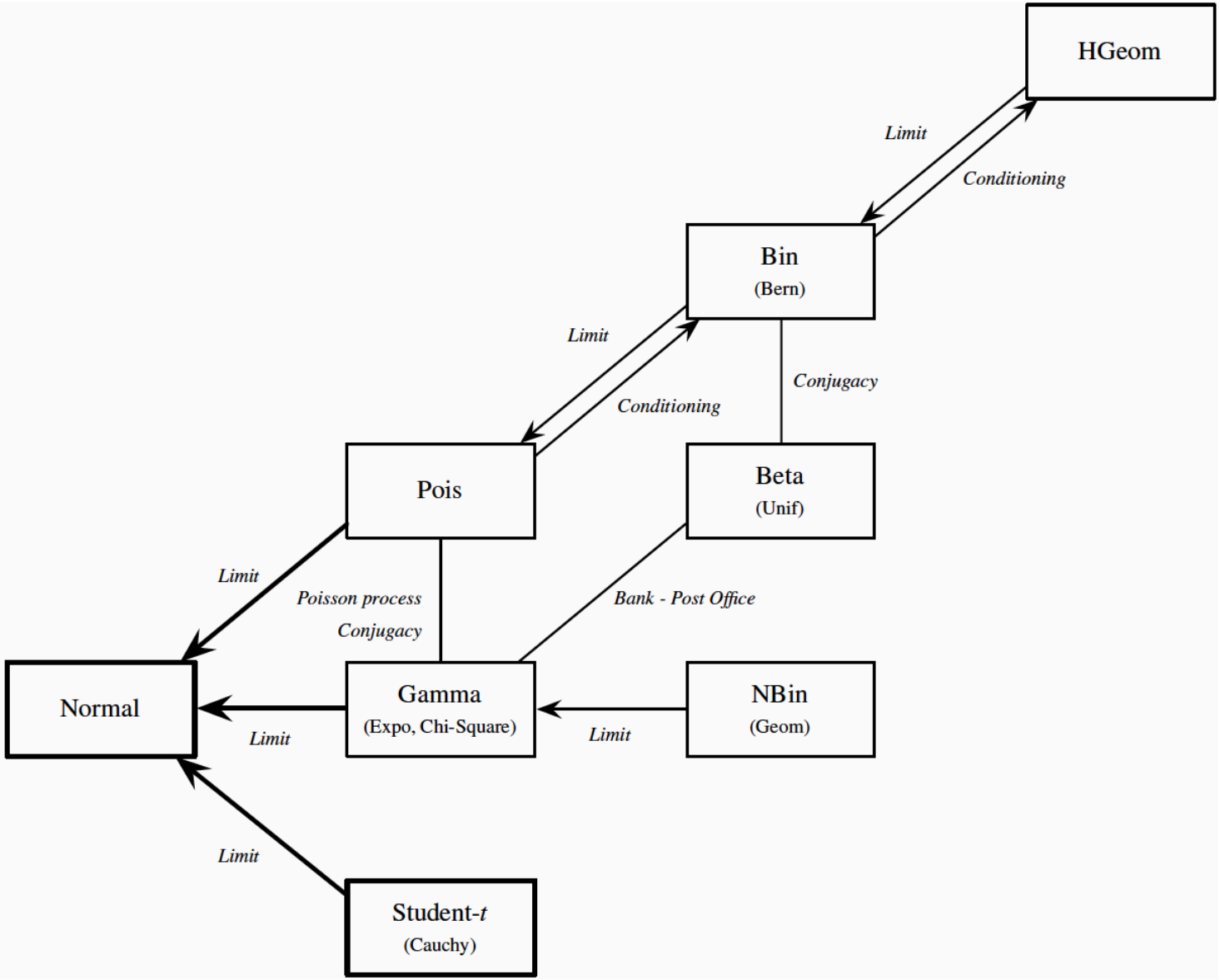
⑥ $\varepsilon = \sqrt{\frac{8\ln(\frac{2}{\delta})}{n}}$ $\Rightarrow Pr(\pi \in (\hat{\pi} - \sqrt{\frac{8\ln(\frac{2}{\delta})}{n}}, \hat{\pi} + \sqrt{\frac{8\ln(\frac{2}{\delta})}{n}})) \ge 1 - \delta$

Example: Monte Carlo Method for Estimation π

Summary 1


distributions
random variables
events
numbers
What can happen?
CDF F
PMF (discrete)
PDF (continuous)
story
name, parameters
MGF
generate
$X_1, X_2, \ldots, X_n$ i.i.d.
$(E(X_i) = \mu, \mathrm{Var}(X_i) = \sigma^2)$
$\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$
$|\bar{X}_n - \mu| > \epsilon$
$a \le \bar{X}_n \le b$
P
$P(|\bar{X}_n - \mu| > \epsilon) \le \frac{\sigma^2}{n\epsilon^2}$
(Chebyshev)
$P(|\bar{X}_n - \mu| > \epsilon) \to 0$ as $n \to \infty$
(law of large numbers)
$P(a \le \bar{X}_n \le b) \approx$
$\Phi\left(\frac{\sqrt{n}(b-\mu)}{\sigma}\right) - \Phi\left(\frac{\sqrt{n}(a-\mu)}{\sigma}\right)$
for large n
(central limit theorem)
$E(\bar{X}_n) = \mu$
$\mathrm{Var}(\bar{X}_n) = \sigma^2/n$

Summary 2


HGeom
Limit
Conditioning
Bin
(Bern)
Limit
Conditioning
Conjugacy
Pois
Beta
(Unif)
Limit
Poisson process
Conjugacy
Bank - Post Office
Normal
Gamma
(Expo, Chi-Square)
NBin
(Geom)
Limit
Limit
Limit
Student-t
(Cauchy)

References

- Chapter 10 of **BH**
- Chapter 5 of **BT**