

Lecture 8: Monte Carlo Methods

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

April 27, 2023

Outline

1 Random Variable Generation

2 Monte Carlo Integration

3 Asymptotic Analysis: Law of Large Numbers

4 Non-asymptotic Analysis: Inequalities

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

- 1 Random Variable Generation
- 2 Monte Carlo Integration
- 3 Asymptotic Analysis: Law of Large Numbers
- 4 Non-asymptotic Analysis: Inequalities

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 $\text{Unif}(0, 1)$ 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 $\text{Unif}(0, 1)$ r.v.
- Other names:
 - ▶ probability integral transform
 - ▶ inverse transform sampling
 - ▶ the quantile transformation
 - ▶ the fundamental theorem of simulation

Inverse Transform Method: Recall

Universality of Uniform

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 $U \sim \text{Unif}(0, 1)$ 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 $F(X) \sim \text{Unif}(0, 1)$.

Algorithm Inverse-Transform Method: PDF Case

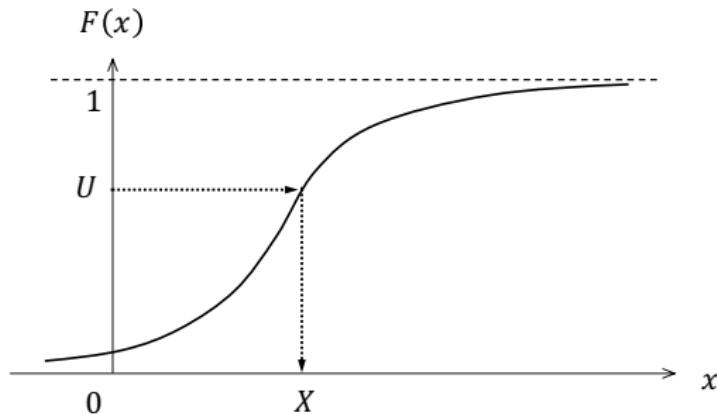
input: Cumulative distribution function F .

output: Random variable X distributed according to F .

1: Generate U from $\text{Unif}(0, 1)$.

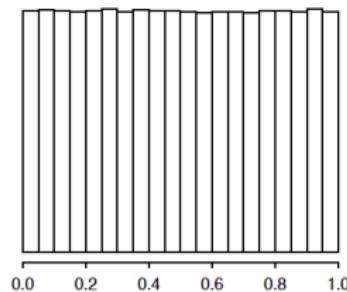
2: $X \leftarrow F^{-1}(U)$

3: **return** X

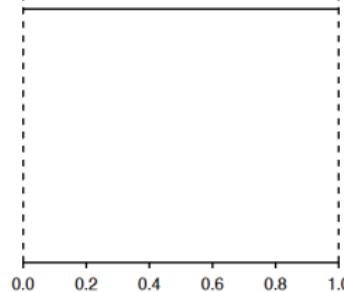


Histogram & PDF: Example

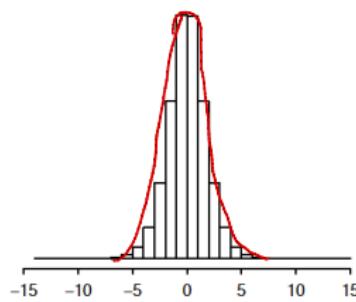
Histogram of u



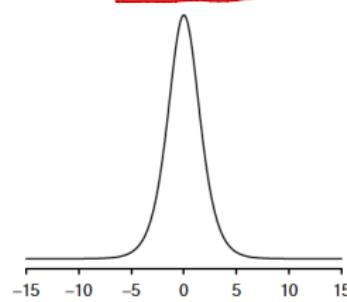
Unif(0,1) PDF



Histogram of $\log(u/(1-u))$



Logistic PDF



Box-Muller: Recall

Let $U \sim \text{Unif}(0, 2\pi)$, and let $T \sim \text{Expo}(1)$ 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

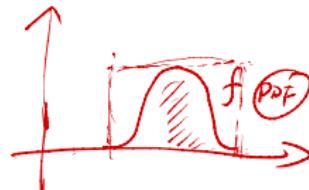
Algorithm Normal Random Variable Generation: Box-Muller Approach

output: Independent standard normal random variables X and Y .

- 1: Generate two independent random variables, U_1 and U_2 , from $\text{Unif}(0, 1)$.
 - 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

接收拒绝法



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

(太大 \Rightarrow 样本效率低)



Algorithm Acceptance-Rejection Algorithm

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

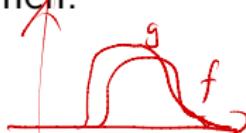
Step 1: Generate $Y \sim g$.

产生新样本

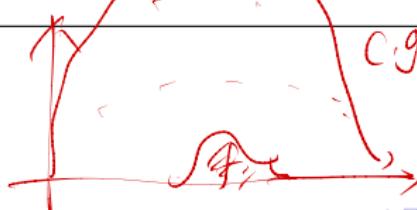
Step 2: Generate $U \sim \text{Unif}[0, 1]$.

$\Rightarrow X \sim f$

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



找到一个易采样的 g , 放大,
直到 Cg 能框住 f



Acceptance-Rejection Method

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

(i) the PDF of generated r.v.s.

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

$$\textcircled{1} P(A|Y=y) = P(U \leq \frac{f(Y)}{c \cdot g(Y)} | Y=y) = P(U \leq \frac{f(y)}{c \cdot g(y)} | Y=y)$$

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) $c \geq 1$

uniform

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

迭代次数 $\sim \text{FS}(\frac{1}{c})$

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

Proof

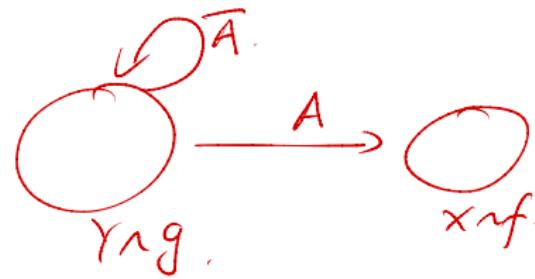
$$\begin{aligned}
 \textcircled{2} \quad P(A) &\stackrel{\text{Defn}}{=} \int \underbrace{P(A|Y=y)}_{\text{Probability}} \cdot g(y) dy = \int \frac{f(y)}{C \cdot g(y)} \cdot g(y) dy \\
 &= \int \frac{f(y)}{C} dy = \frac{1}{C} \underbrace{\int f(y) dy}_{\leq 1} = \frac{1}{C} \leq 1 \Rightarrow (\text{C} \geq 1)
 \end{aligned}$$

$$\begin{aligned}
 \textcircled{3} \quad f_Y(y|A) &= \frac{P(A|Y=y)}{P(A)} \cdot \underbrace{f_Y(y)}_{\text{Probability}} \quad \uparrow \text{概率} \leq 1 \\
 &= \frac{\frac{f(y)}{C \cdot g(y)}}{\frac{1}{C}} \cdot g(y) = \underline{\frac{f(y)}{C}} \cdot g(y) = \underline{f(y)}, \quad \text{by } C \geq 1
 \end{aligned}$$

~~(X) ∩ Ø~~

Proof

(ii)



of iterations $\geq \sim F_{SCP}$)

$$P = P(A) = \frac{1}{C} < 1.$$

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

Example: Beta Distribution

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

Beta 分布在 $(0, 1)$ 上
易采样 $\Rightarrow \text{Unif}(0, 1)$

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

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

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

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

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

PICK $C = \frac{135}{64}$.

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

$$Z \sim N(0, 1)$$

$$(-\infty, +\infty)$$

$$(0, +\infty)$$

Exp(λ) 易采样

②

$$X = |Z|$$

$$(0, +\infty)$$

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

proposal distribution
 $\frac{\text{Exp}(λ)}{\text{Exp}(1)}$ $(0, +\infty)$

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

③

$$X \sim N(0, 1), f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, 0 < x < \infty.$$

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

$$\text{Now } \frac{f(x)}{g(x)} = \frac{1}{\sqrt{\frac{2}{\pi}}} e^{-\frac{1}{2}x^2+x} \quad C \geq \sup_x \frac{f(x)}{g(x)} = \frac{f(x)}{g(x)} = \frac{\sqrt{\frac{2}{\pi}}}{1}.$$

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

Solution

(4)

Step 1: Generate $\gamma \sim \text{Exp}(1)$

2: $---$ $u \sim \text{uniform}$

3: If $u \leq \exp\{-\frac{1}{2}(\gamma-1)^2\}$, set $x = \gamma$ Accept the sample

Otherwise, reject it, return to step 1.

4: $---$ $u' \sim \text{uniform}$.

$$z' = \begin{cases} x & \text{if } u' \leq \frac{1}{2} \\ -x & \text{if } u' > \frac{1}{2}. \end{cases}$$

$\Rightarrow z \sim N(0, 1)$

Compare
Box-Mullen

Outline

- 1 Random Variable Generation
- 2 Monte Carlo Integration
- 3 Asymptotic Analysis: Law of Large Numbers
- 4 Non-asymptotic Analysis: Inequalities

Monte Carlo Integration

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

$$c = \mathbb{E}_{\underline{f}}(h(\underline{x})) = \int \underline{f(\underline{x})} h(\underline{x}) d\underline{x};$$

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

$$\underline{x_1}, \underline{x_2}, \dots, \underline{x_n} \sim f(\underline{x})$$

$$\underline{h(x_1)}, \dots, \underline{h(x_n)}$$

- One can estimate c by the sample mean:

$$\frac{1}{n} \sum_{j=1}^n h(\underline{x_j})$$

Example: π as An Integration

Evaluate the integration

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

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

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

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

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

$$E[h(x)] \xrightarrow{\text{LOTUS}} \int_0^1 h(x) f(x) dx = \int_0^1 h(x) dx$$

Useful Tools: Importance Sampling

重要性采样
↓
极小概率事件

- 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

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

$E_f[h(y)]$
 $y \sim g$

- 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

$$\frac{f(Y)}{g(Y)} \Rightarrow \text{权重}$$

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

- Hence, given an iid sample Y_1, \dots, Y_n from PDF \underline{g} , our estimator of I becomes

$$\hat{I} = \underbrace{\frac{1}{n} \sum_{j=1}^n \frac{h(Y_j)f(Y_j)}{g(Y_j)}}_{\text{weight.}}$$

Example: Gaussian Tail Probability

Key : probability $\xleftarrow{\text{Indicator variable.}}$ Expectation $\xrightarrow{\text{Expectation}}$

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

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

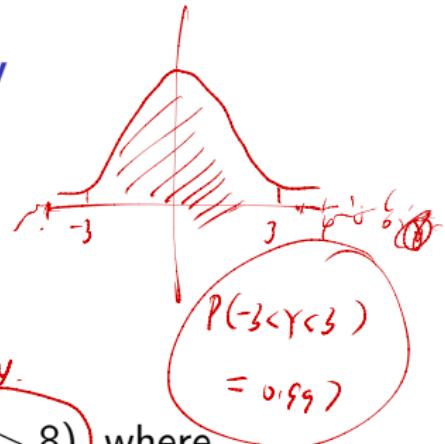
Evaluate the probability of rare event $c = \mathbb{P}(Y > 8)$, where $Y \sim N(0, 1)$.

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

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

$\neq 0$

(2) $\underline{g \sim N(8, 1)}$



Solution

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

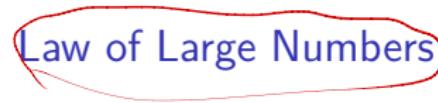
$Y_1, \dots, Y_n \sim g(N(8, 1))$.

$$\begin{aligned}\hat{C}^I &= \frac{1}{n} \sum_{j=1}^n \frac{h(Y_j) f(y_j)}{g(y_j)} = \frac{1}{n} \sum_{j=1}^n I(Y_j > 8) \\ &\quad \cdot \frac{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}y_j^2}}{\frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(y_j - 8)^2}}\end{aligned}$$

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

6.25×10^{-16}

Outline

- 1 Random Variable Generation
- 2 Monte Carlo Integration
- 3 Asymptotic Analysis: Law of Large Numbers

- 4 Non-asymptotic Analysis: Inequalities

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

无偏估计 \Rightarrow Unbiased estimation

Definition

n samples from the same distribution

Let X_1, \dots, 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:

CLT

中心极限定理

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

$\text{Var}(\cdot) \rightarrow 0$

\Rightarrow 随机变量 \rightarrow constant

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

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

$n \rightarrow \infty$

$\Rightarrow \text{Var}(\bar{X}_n) \rightarrow 0$

$\bar{X}_n \rightarrow \mu$

Strong Law of Large Numbers (SLLN)

强大数定理

(Monte Carlo的原理)

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

Theorem

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

$$\left[\frac{1}{n} (f(X_1) + \dots + f(X_n)) \xrightarrow[n \rightarrow \infty]{\text{w.p. 1}} E[f(X)] \right]$$

$X_1, \dots, X_n, X \sim F$

Weak Law of Large Numbers (WLLN)

弱大数定理

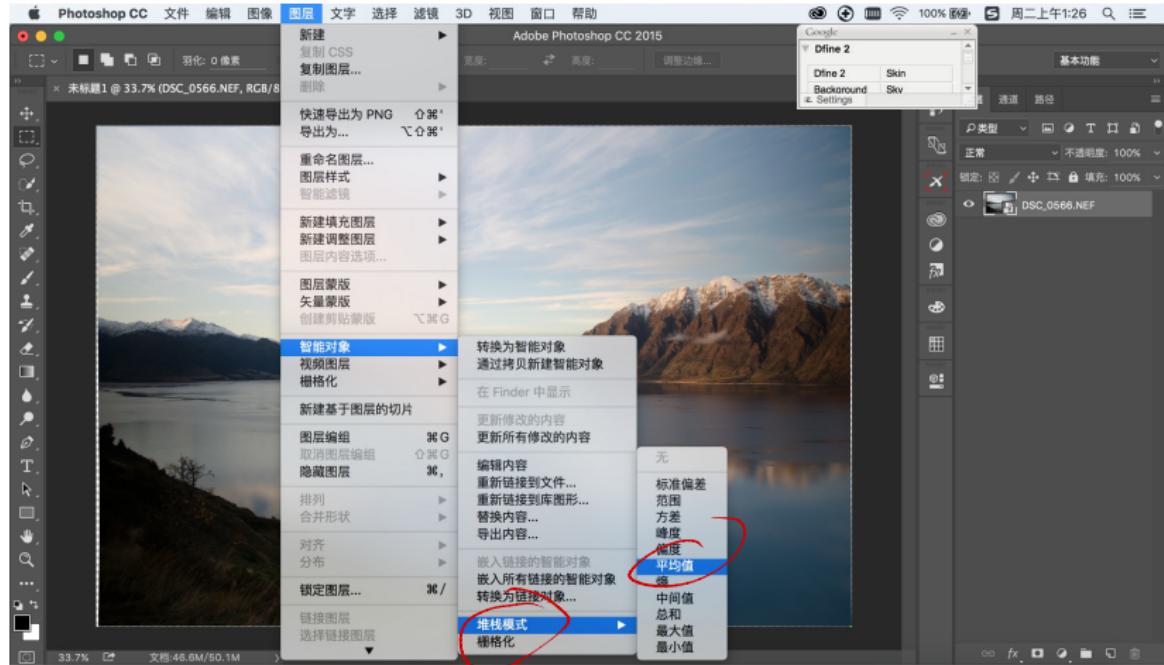
Theorem

For all $\epsilon > 0$, $P(|\bar{X}_n - \mu| > \epsilon) \rightarrow 0$ as $n \rightarrow \infty$. (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

1 Random Variable Generation

2 Monte Carlo Integration

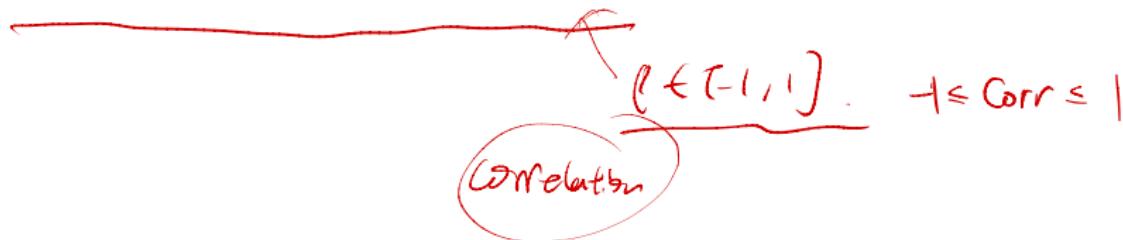
3 Infinite Samples. Asymptotic Analysis: Law of Large Numbers

4 Finite Samples. Non-asymptotic Analysis: Inequalities

$$P(|\bar{x}_n - \mu| > \varepsilon) \leq \dots$$

有限采样

Cauchy-Schwarz Inequality: Recall



Theorem

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

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

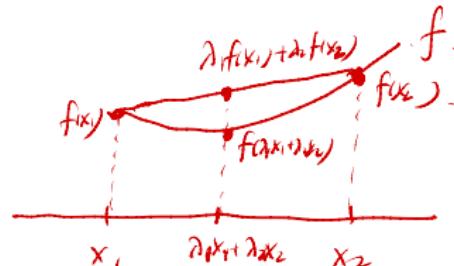
Jensen's Inequality

(x^l)

$$f''(x) > 0$$

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 ,

凸函数



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

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^o - $g(x) = x^2, x \in \mathbb{R}$. ; Convex ; $E[x^2] \geq (E[x])^2$ ✓
 $\text{Var}(x) = E(x^2) - (E[x])^2 \geq 0$

2^o $g(x) = \frac{1}{x}, x > 0$; Concave ; $E\left(\frac{1}{x}\right) \geq \frac{1}{E[x]}$ ✓

3^o - $g(x) = \log x, x > 0$; Concave ; $E[\log x] \leq \log(E[x])$ ✓

Entropy 火熵

正态分布: 期望有约束的情况下
取到最大熵的分布
(continuous)

- Let X be a discrete r.v. whose distinct possible values are a_1, a_2, \dots, a_n , with probabilities p_1, p_2, \dots, p_n respectively (so $p_1 + p_2 + \dots + p_n = 1$).
- 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, \dots, a_n , i.e., $p_j = 1/n$ for all j .
最大熵分布
- 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.

$$P_i = 1 ; P_j = 0 (j \neq i)$$

Proof ① Construct a r.v. Y s.t.

$$Y = \begin{cases} \frac{1}{P_1} & \text{w.p. } P_1 \\ \frac{1}{P_2} & \text{w.p. } P_2 \\ \vdots & \\ \frac{1}{P_n} & \text{w.p. } P_n \end{cases} \quad (P_1 + \dots + P_n = 1)$$

$$E(Y) = \frac{1}{P_1} \cdot P_1 + \frac{1}{P_2} \cdot P_2 + \dots + \frac{1}{P_n} \cdot P_n = n$$

$$\textcircled{2} \quad H(X) = \sum_{j=1}^n P_j \log_2 \frac{1}{P_j} = \underline{E[\log_2 Y]} \leq \log_2(EY) = \log_2 n$$

$$\Theta(P_1, \dots, P_n) \quad H(X) \leq \log_2 n \quad \Rightarrow \quad \underline{\max_{P_1, \dots, P_n} H(X)} \leq \log_2 n.$$

$$\textcircled{3} \quad \underline{\max_{P_1, \dots, P_n} H(X)} \geq \log_2 n \quad [\quad X \sim \text{Dunif}(\frac{1}{n}), P_1 = P_2 = \dots = \frac{1}{n}, H(X) = \log_2 n \quad]$$

$$\textcircled{4} \quad \Rightarrow \quad \max_{P_1, \dots, P_n} H(X) = \log_2 n, \quad X \not\sim \text{Dunif}(\frac{1}{n}).$$

Kullback-Leibler Divergence

Relative Entropy

K-L 散度

伪距离，相对熵…

Let $\mathbf{p} = (p_1, \dots, p_n)$ and $\mathbf{r} = (r_1, \dots, r_n)$ 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

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

不对称,

Show that the Kullback-Leibler divergence is nonnegative

$$\begin{aligned}
 \text{Proof } ① D(p, r) &= \sum_{j=1}^n p_j \log_2 \frac{r_j}{p_j} - \sum_{j=1}^n p_j \log_2 \frac{r_j}{p_j} \\
 &= \sum_{j=1}^n p_j \log_2 \frac{r_j}{p_j} = - \underbrace{\sum_{j=1}^n p_j \log_2 \frac{r_j}{p_j}}_{\text{---}}
 \end{aligned}$$

② Construct a r.v. Y , s.t.

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

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

$$\begin{aligned}
 \text{③ } \underline{D(p, r)} &= - \sum_{j=1}^n p_j \log_2 \frac{r_j}{p_j} = - \underline{E[\log_2 Y]} \\
 &\geq -\log_2(E(Y)) = -\log_2 1 = 0
 \end{aligned}$$

马尔可夫不等式
Markov's Inequality

$$\frac{f(x) \leq a}{f(x) \leq b}$$

\checkmark $a < b$ tighter

Concentration Inequality 集中不等式

Chebyshev 切比雪夫

Markov 马尔可夫

Lyapunov.

Theorem

For any r.v. X and constant $\underline{a} > 0$,

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

Proof

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

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

LHS	RHS
$I \leq Y$	$1 \leq Y$
LHS	RHS
$0 \leq Y < 1$	$0 \leq Y$

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

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

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

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

Chebyshev's Inequality

切比雪夫不等式

$$\textcircled{D} \quad P(|X-\mu| \geq a) = P(|X-\mu|^2 \geq a^2)$$

$$E[(X-E(X))^2] = \text{Var}(X) \quad \stackrel{\text{Markov's Inequality}}{\leq} \frac{1}{a^2} E(|X-\mu|^2) = \frac{1}{a^2} \sigma^2$$

Theorem

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

$$P(|X-\mu| \geq a) \leq \frac{\sigma^2}{a^2} \quad \uparrow a \quad \downarrow \left(\frac{1}{a^2 \sigma^2}\right)$$

Proof

Application:

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

$$P \leq P(|\bar{X}_n - \mu| \geq a) \leq \frac{1}{a^2} \text{Var}(\bar{X}_n - \mu) = \frac{1}{a^2} \text{Var}(\bar{X}_n)$$

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

$$n \rightarrow \infty$$

$$\downarrow 0$$

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

WTLYN

Chernoff's Inequality $\forall t > 0$;

切尔诺夫不等式

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

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

Theorem

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

$$\overbrace{P(X \geq a)}^{\text{MGF}} \leq \frac{E(e^{tX})}{e^{ta}}.$$

Proof

Chernoff's Technique

切尔诺夫技术

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

$E(e^{tX})$

 e^{ta}
 $f(t)$

Theorem

For any r.v. X and constants a ,

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

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

最小值
↑
 $E(e^{tX})$

 e^{ta}

$$\textcircled{2} \quad P(X \leq a), \quad t < 0 \quad \Rightarrow P(tX \geq ta) = P(tX \geq -ta)$$

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

Proof

Example: Normal Distribution

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

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

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

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

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

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

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

Solution

Hoeffding Lemma

霍夫丁引理

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 $\lambda > 0$,

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

MGF

Hoeffding Bound

Theorem

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

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

X_n

Application: Parameter Estimation

$$\hat{f} = 0.05$$

常用: 95% 置信区间 ($\delta = 5\%$)

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

Definition 置信区间

\hat{p} : estimation of p .

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

↓
估值

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

$$\Leftrightarrow \Pr(|\hat{p} - p| \leq \epsilon) \geq 1 - \delta$$

$$\Leftrightarrow \hat{p} - \epsilon \leq p \leq \hat{p} + \epsilon$$

$$\Leftrightarrow -\epsilon \leq p - \hat{p} \leq \epsilon$$

$$\Leftrightarrow -\epsilon \leq \hat{p} - p \leq \epsilon$$

$$\Leftrightarrow |\hat{p} - p| \leq \epsilon$$

ϵ 可用 δ 表达

$$\Leftrightarrow \Pr(|\hat{p} - p| \geq \epsilon) \leq \delta$$

Application: Parameter Estimation

① $X_i \sim \text{Bern}(p)$. $E(X_i) = p$; $0 \leq X_i \leq 1$

$(X_i = 1 \text{ or } 0)$
 $(a=0; b=1)$

$$E(\hat{p}) = p; \text{ LLN: } n \rightarrow \infty; \hat{p} \rightarrow p.$$

② $\Pr(|\hat{p} - p| \geq \varepsilon) = \Pr\left(\left|\frac{1}{N} \sum_{i=1}^N X_i - p\right| \geq \varepsilon\right)$

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 i th result is head, otherwise 0. We estimate p by using $\hat{p} = \frac{X_1 + \dots + X_N}{N}$. Find the confidence interval for p .

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

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

Solution

$$\textcircled{4} \quad \Pr(|\hat{P} - P| \geq \varepsilon) \leq \delta$$

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

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

$$\textcircled{5} \quad (\hat{P} - \varepsilon, \hat{P} + \varepsilon)$$

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

± δ Confidence Interval.

Application Example: Monte Carlo Method for Estimation π

①

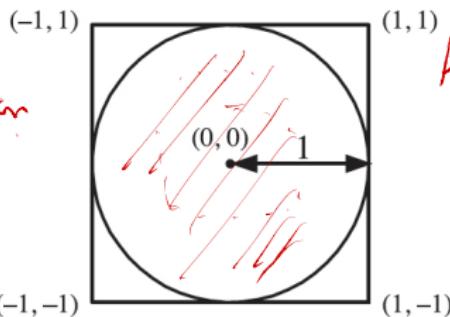
square $[-1, 1] \times [-1, 1]$

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

②

$$\pi = 4 \Pr(A)$$

$$= 4 \cdot E(I_A) \quad \xrightarrow{\text{indicator}}$$



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

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 $I_{A_i} = \begin{cases} 1 & \text{"the } i^{\text{th}} \text{ point landing in the circle"} \\ 0 & \text{otherwise} \end{cases}$

Example: Monte Carlo Method for Estimation π

$$P(Z_i=1) = \frac{1}{4}, \quad E(Z_i) = \frac{1}{4}, \quad 0 \leq Z_i \leq 1 \quad Z_i = 1 \text{ or } 0$$

④ Suppose we generate n points.

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

⑤ $\hat{\pi} = 4W$

$n \rightarrow \infty, \hat{\pi} \rightarrow \pi$ $\Rightarrow \Pr(|\hat{\pi} - \pi| \geq \varepsilon) = \Pr(|4W - \pi| \geq \varepsilon)$

Extends $= \Pr\left(|W - \frac{1}{4}| \geq \frac{\varepsilon}{4}\right) = \Pr\left(\left|\frac{1}{n} \sum_{i=1}^n Z_i - \frac{1}{4}\right| \geq \frac{\varepsilon}{4}\right)$

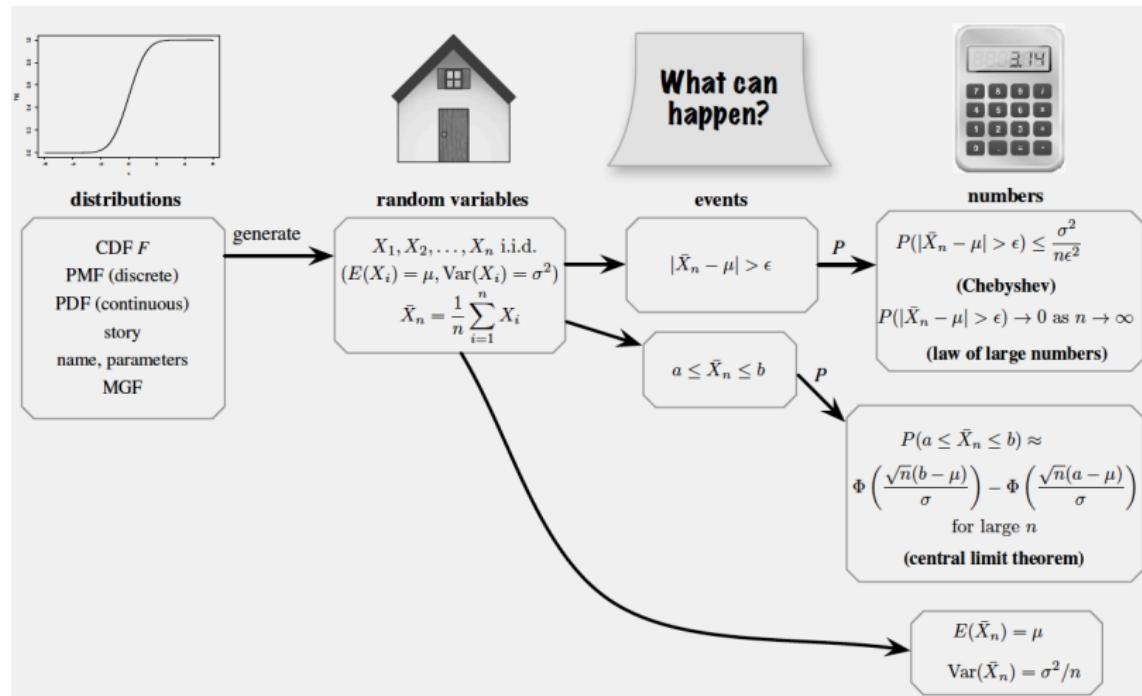
Hoeffding Bound $\leq 2e^{-\frac{2n(\frac{\varepsilon}{4})^2}{(1-0)^2}} = \underline{2e^{-\frac{1}{8}n\varepsilon^2}} = \delta$

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

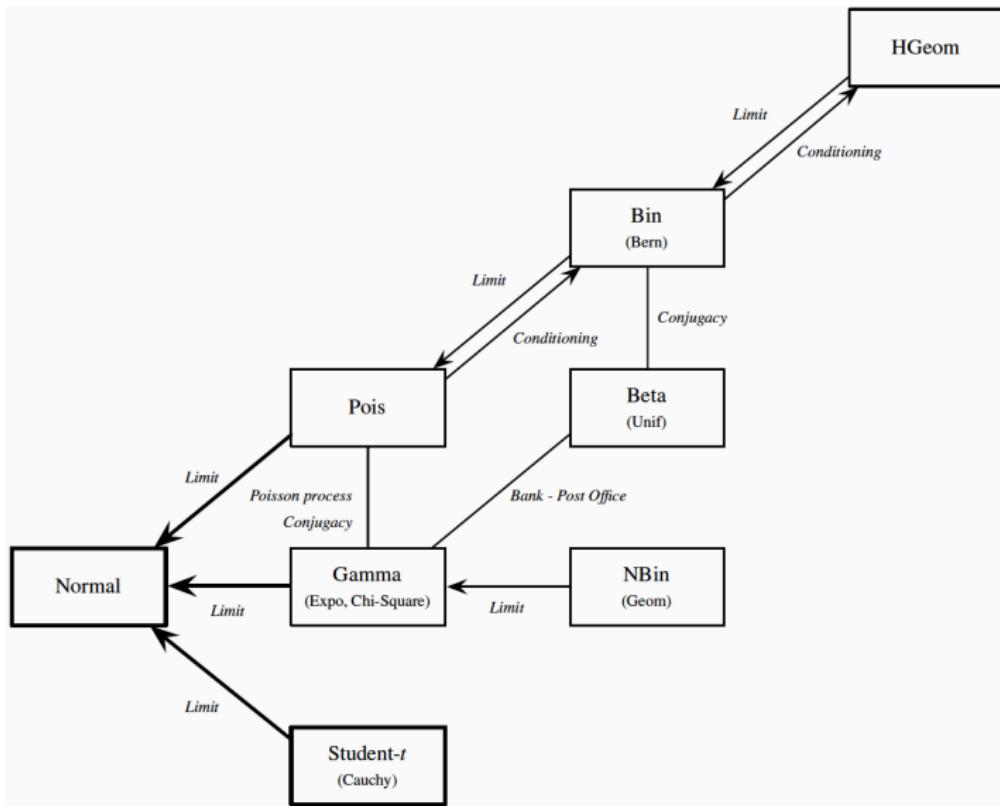
$n \uparrow, \text{置信区间} \downarrow, \hat{\pi} \rightarrow \pi > 1 - \delta$

Example: Monte Carlo Method for Estimation π

Summary 1



Summary 2



References

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