Lecture 7: Monte Carlo Methods

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

December 3, 2024

Outline

Outline

Motivation I

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

1 - Simulations using Monte Carlo Methods
2 - 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
3 - Bounds (upper and lower bounds) on probability using inequalities.

Motivation II

Probability

Math

Statistics

Science

Monte Carlo

Computing

Monte Carlo Methods

- One of the top ten algorithms for science and engineering in 20th century
- Monte Carlo Methods, Simplex Method, Fast Fourier Transform, Quicksort, QR Algorithm...

Widely Applications

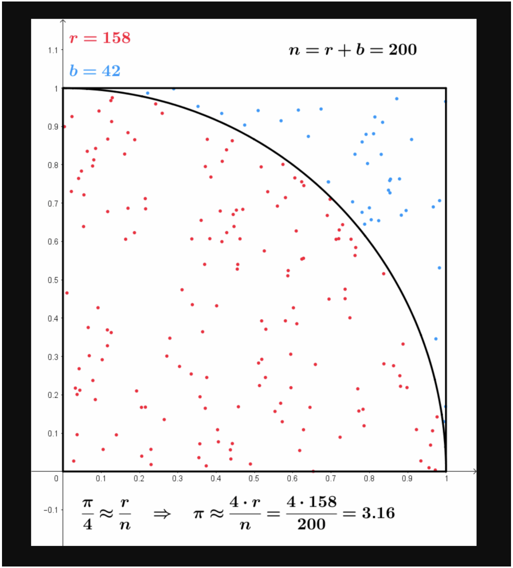
Monte Carlo methods have been used in various tasks, including

- Sampling from the underlying probability distribution $f(x)$ and simulating a random system
- Sampling from posterior distribution for bayesian inference
- Estimation through numerical integration

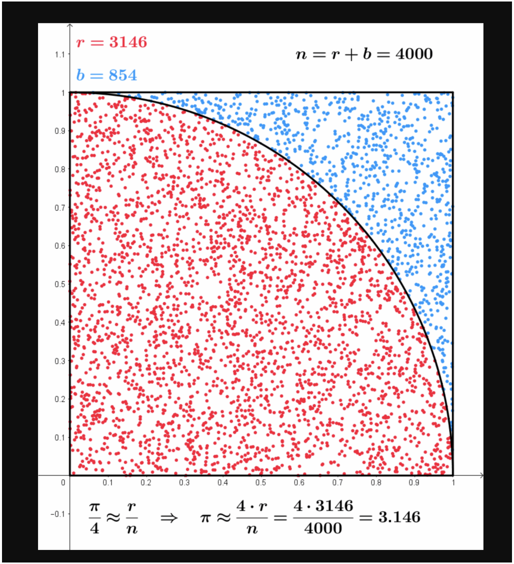
$$c = E_\pi(h(x)) = \int f(x)h(x)dx.$$

- Optimizing a target function to find its maxima or minima

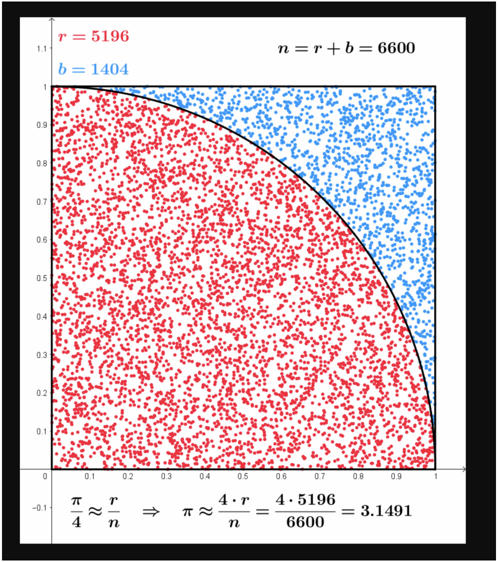
Classical Example: Estimation of π



Classical Example: Estimation of π



Classical Example: Estimation of π



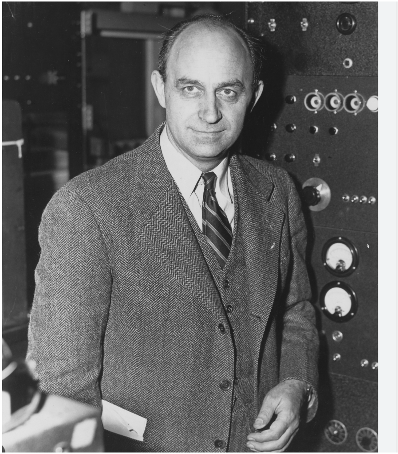
History

Monte Carlo Methods

- 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)

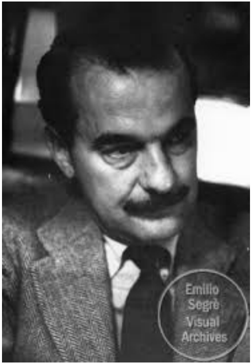
Monte Carlo Trolley

- Analog computer invented by Enrico Fermi in 1946



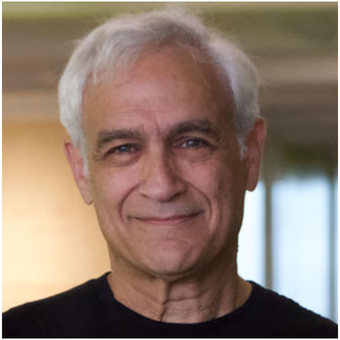
Markov Chain Monte Carlo Methods

- Metropolis-Hastings Algorithm: formally proposed by Nicholas Metropolis et al in 1950s at Los Alamos National Lab, then extended in 1970 by Wilfred Keith Hastings



Markov Chain Monte Carlo Methods

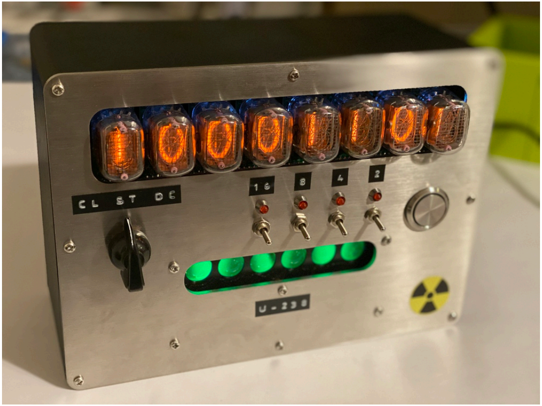
- Gibbs Sampling Algorithm: proposed in 1984 by brothers Stuart Geman (1949-) and Donald Geman (1943-).
- Gibbs sampling is named after the physicist Josiah Willard Gibbs (1839-1903), in reference to an analogy between the sampling algorithm and statistical physics.



Outline

Randomness Generation

- Earlier days: manual techniques including coin flipping, dice rolling, card shuffling, and roulette spinning
- Early days: physical devices including noise diodes and Geiger counters (`https://github.com/nategri/chernobyl_dice`)



Randomness Generation

- The prevailing belief: only mechanical or electronic devices could produce truly random sequences
- The book: *A Million Random Digits With 100,000 Normal Deviates* (based on Uranium radiation)
- Current days: computer simulation with deterministic algorithms, also called pseudorandom number generator

Sampling

- Assuming an algorithm is available for generating Unif(0, 1) random numbers
- Two elementary methods for generating random variables (or samples)
  - 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

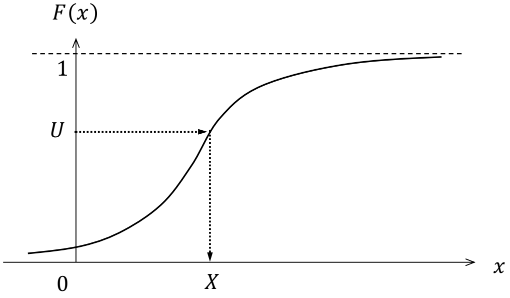
input: Cumulative distribution function F.

output: Random variable X distributed according to F.

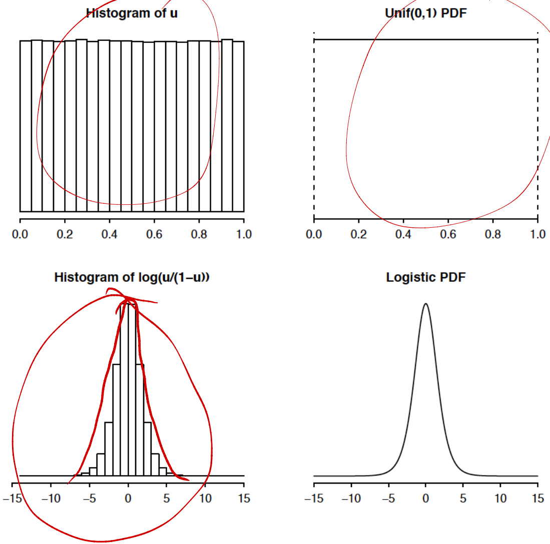
1: Generate U from $\mathrm{Unif}(0, 1)$.

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

3: **return** X



Histogram & PDF: Example



Box-Muller Method: Recall

1°. $U \sim Unif(0,2\pi)$

$= 2\pi\, Unif(0,1)$

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 $Y = \sqrt{2T}\sin U$. Then X and Y are independent, and their marginal distributions are standard normal distribution.

$\Rightarrow U_2 \sim Unif(0,1)$

$U = 2\pi U_2$

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

2°. $T \sim Expo(1)$

CDF of T: $F_T(t) = 1 - e^{-t}$, $t>0$

$\Rightarrow F_T^{-1}(u) = -\ln(1-u)$, $0<u<1$.

$U_1 \sim Unif(0,1)$

$1-U_1 \sim Unif(0,1)$

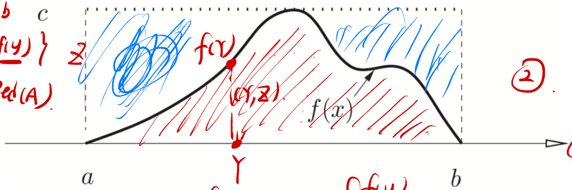
Acceptance-Rejection Method

① object PDF f: $X \in [a,b]$; $c \geq \sup_x f(x)$. $\int_a^b f(x)\,dx = 1$; Area(Red) $= 1$

③ Accept Rule: if $Z \leq f(Y)$, Accept (Y,Z).

$\text{Red}(A) = \{(y,z) : a \leq y \leq b,\ 0 \leq z \leq f(y)\}$

④ $(Y^*, Z^*) \in \text{Red}(A)$



② $Y \sim \text{Unif}(a,b)$, $Z \sim \text{Unif}(0,c)$.

$f_{Y^*,Z^*}(y,z) = \frac{1}{\text{Area(Red)}} = 1 \Rightarrow f_{Y^*}(y) = \int_0^{f(y)} f_{Y^*,Z^*}(y,z)\,dz$, $a \leq y \leq b$

$= \int_0^{f(y)} 1 \cdot dz$

$= f(y)$, $a \leq y \leq b$

Algorithm Acceptance-Rejection Algorithm

Step 1: Generate $Y \sim \text{Unif}(a, b)$.
Step 2: Generate $Z \sim \text{Unif}(0, c)$.
Step 3: If $Z \leq f(Y)$, set $X = Y$. Otherwise go back to step 1.

$\frac{1}{b-a} \cdot \frac{1}{c} = \frac{1}{\text{area(Rectangle)}}$

$Y^* \sim f$

Acceptance-Rejection Method

① $X \in [a,b]$, PDF f: desired.

g: PDF $\quad \phi(x) = C\cdot g(x) \geq f(x)$.

$\Rightarrow C \geq \sup_x \frac{f(x)}{g(x)}, x \in [a,b]$

② $Y \sim g$, $Z \sim \mathrm{unif}(0, C g(Y))$

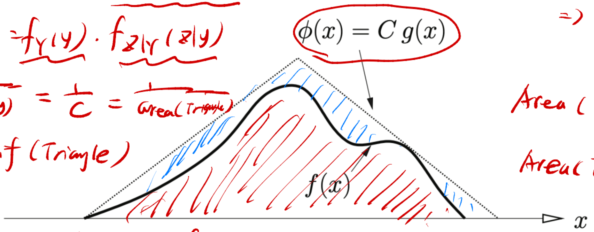
$\Rightarrow f_{Y,Z}(y,z) = f_Y(y) \cdot f_{Z|Y}(z|y)$

$= g(y) \cdot \frac{1}{Cg(y)} = \frac{1}{C} = \frac{1}{\text{area(Triangle)}}$

$(Y, Z) \sim \mathrm{unif}(\text{Triangle})$

③ (Y^*, Z^*)

$\in$ Red(A). $\sim \mathrm{unif}(A)$; $f_{Y^*,Z^*}(y,z) = \frac{1}{\text{area(Red)}} = 1$



Area (Red) = 1

Area(Triangle)

$= \int_a^b \phi(x)\,dx$

$= \int_a^b C\cdot g(x)\,dx$

$= C\int_a^b g(x)\,dx = C \cdot 1 = C$

Algorithm Acceptance-Rejection Algorithm

Step 1: Generate $Y \sim g$. $\quad Z|Y$

Step 2: Generate $Z \sim \mathrm{Unif}(0, c \cdot g(Y))$.

Step 3: If $Z \leq f(Y)$, set $X = Y$. Otherwise go back to step 1.

$\mathrm{Unif}(0, C\cdot g(Y)) \leq f(Y) \Leftrightarrow C\cdot g(Y)\,\mathrm{Unif}(0,1) \leq f(Y)$

$\Leftrightarrow \quad \mathrm{unif}(0,1) \leq \frac{f(Y)}{C\cdot g(Y)}$

$C > 1$

Acceptance-Rejection Method

f, g.

support $\Leftarrow$?

- 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:

$c \geq 1$

Step 1: Generate $Y \sim g$.
Step 2: Generate $U \sim \text{Unif}(0, 1)$.
Step 3: If $U \leq \frac{f(Y)}{c \cdot g(Y)}$, set $X = Y$. Otherwise go back to step 1.

Acceptance-Rejection Method

A^c

A

X

$Y \sim g$

\# of iteration

$N \sim FS(p)$

$p = P(A) = \frac{1}{c}$

Theorem

$E(N) = \frac{1}{p} = c$

(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) $c \geq 1$

Proof

1° event $A =$ " $U \le \frac{f(Y)}{c\cdot g(Y)}$ " ; $f_Y(y|A)$

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

($U \sim \text{unif}(0,1)$, $U \perp Y$.)

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

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

($P(U \le t) = t,\ 0 \le t \le 1$)

2° $P(A) \overset{\text{LOTP}}{=} \int P(A|Y=y)\cdot f_Y(y)\cdot dy$ $\quad$ ($Y \sim g$)

$= \int \frac{f(y)}{c\cdot \cancel{g(y)}} \cdot \cancel{g(y)}\cdot dy = \frac{1}{c}\int f(y)\,dy = \frac{1}{c} \le 1$

$\Rightarrow f_Y(y|A) = \frac{P(A|Y=y)}{P(A)} f_Y(y) = \frac{\frac{f(y)}{\cancel{c\cdot g(y)}}}{\frac{1}{\cancel{c}}} \cancel{g(y)} = f(y)$

$c \ge 1$

Proof

Example: Beta Distribution

- An r.v. X is said to have the *Beta distribution* with parameters a and b, $a > 0$ and $b > 0$, if its PDF is

a=1, b=1

$$f(x) = \frac{1}{\beta(a,b)} x^{a-1} (1-x)^{b-1}, \; 0 < x < 1,$$

where the constant $\beta(a,b)$ is chosen to make the PDF integrate to 1. We write this as $X \sim \mathrm{Beta}(a,b)$.

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

Solution

object PDF $f(x) = 20x(1-x)^3 \ , 0<x<1$

① g: $\text{Unif}(0,1)$, $g(x) = 1$, $0<x<1$.

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

$\Rightarrow c \geq \frac{135}{64} > 1$. choose $c = \frac{135}{64}$.

② $\Rightarrow 0<y<1$, $\frac{f(y)}{c\cdot g(y)} = \frac{20y(1-y)^3}{\frac{135}{64}\cdot 1} = \frac{256}{27}y(1-y)^3$

Step 1 : Generate $Y \sim \text{Unif}(0,1)$.

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 reject Y, Go back to step 1.

Solution

Example: Normal Distribution

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

$X = |Z|$ $(0, +\infty)$

$P(X \le x) = P(|Z| \le x) = 2P(0 \le Z \le x)$

$= 2\int_0^x \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} dz$ $\Rightarrow f_X(x) = \sqrt{\frac{2}{\pi}} e^{-\frac{1}{2}x^2}, \ 0 < x < \infty$

② $g \sim \text{Expo}(1)$. $g(x) = e^{-x}, \ 0 < x < \infty$

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

$c \ge \sup_y \frac{f(y)}{g(y)} = \sup_y \sqrt{\frac{2}{\pi}} e^{\underbrace{-\frac{1}{2}y^2 + y}} = \sqrt{\frac{2e}{\pi}}$ $(y^* = 1)$

choose $c = \sqrt{\frac{2e}{\pi}}$.

$\Rightarrow \frac{f(y)}{c \cdot g(y)} = e^{\{y - \frac{1}{2}y^2 - \frac{1}{2}\}} = e^{-\frac{1}{2}(y-1)^2}$

Solution

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

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

3: If $U \le e^{-\frac{1}{2}(Y-1)^2}$, set $X = Y$.

otherwise return to step 1.

$X = |Z|$.

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

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

$Z \sim N(0,1)$

Box-Muller

v.s.

Acceptance - Rejection

Pros / cons.

Solution

Outline

Monte Carlo Integration

$X_1, \ldots X_n$: $E(X) \approx \frac{1}{n}(X_1 + \cdots + X_n)$

- We can use the sample mean to approximate the expectation:

$g(X_1) \ldots g(X_n)$

$$E[g(X)] \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i).$$

$X \sim \mathrm{unif}(a,b)$

- Now we have integration

$f(x)$ PDF

$$\int_a^b g(x)dx = (b-a)\int_a^b g(x)\cdot\frac{1}{b-a}dx. = (b-a)\cdot\int_a^b g(x)f(x)dx$$

- Drawing n samples (empirical samples) from Unif(a, b):

$$X_1, X_2, \ldots, X_n \sim \mathrm{Unif}(a,b).$$

$= (b-a)\cdot E[g(X)]$

- Monte Carlo Integration:

$\approx (b-a)\cdot\frac{1}{n}\sum_{i=1}^{n} g(X_i)$

$$\int_a^b g(x)dx \approx \frac{1}{n}\sum_{i=1}^{n} g(X_i)(b-a).$$

$X_1, \ldots X_n \sim \mathrm{unif}(a,b)$

Monte Carlo Integration

Example: π as An Integration

Evaluate the integration

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

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

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

Example

Evaluate the integration

$$\int_0^4 \sqrt{x+\sqrt{x+\sqrt{x+\sqrt{x}}}}\,dx.$$

- Corresponding

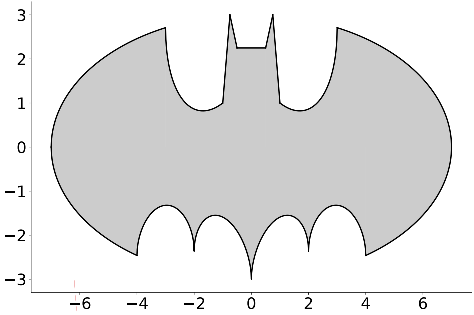
$$g(x) = \sqrt{x+\sqrt{x+\sqrt{x+\sqrt{x}}}}.$$

- $X_1, \ldots, X_n$: samples from Unif$(0,4)$.
- Monte Carlo Integration:

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

Example: Area of Batman Curve

- Challenging and Fun
- `https://mathworld.wolfram.com/BatmanCurve.html`



Example: Estimation of Probability

- Indicator: bridge between expectation and probability
- Given event A:

$$I_A(x) = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{Otherwise} \end{cases}.$$

- For random variable X:

$$\begin{aligned} P(X \in A) &= 1 \cdot P(X \in A) + 0 \cdot P(X \notin A) \\ &= E(I_A(X)) \\ &\approx \frac{1}{n} \sum_{i=1}^{n} I_A(X_i). \end{aligned}$$

$X_1, \dots, X_n \sim X$

Example: Estimation of π

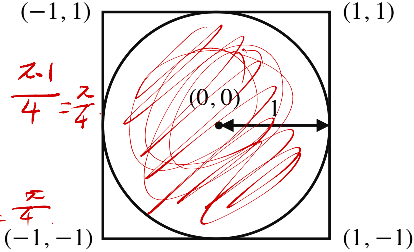
① n generate points $(x_1, y_1) \dots (x_n, y_n)$

$-1 \le x_i \le 1$

$-1 \le y_i \le 1$

② event A_i = " the ith point lands within the circle".

$\Leftrightarrow$ " $\{(x_i, y_i) : x_i^2 + y_i^2 \le 1\}$ "



③ $I_{A_i} = Z_i$

$\Rightarrow P(Z_i = 1) = P(A_i) = \frac{\pi \cdot 1}{4} = \frac{\pi}{4}$

$P(Z_i = 0) = 1 - \frac{\pi}{4}$

$\Rightarrow E(Z_i) = P(Z_i = 1) = \frac{\pi}{4}$

④ $Z_1, \dots, Z_n \sim Z$

$\Rightarrow E(Z) = \frac{\pi}{4}$ $\Rightarrow \pi = 4E(Z) \approx 4 \cdot \frac{1}{n}(Z_1 + \dots + Z_n)$

完

Example: Estimation of π

Example: Estimation of π

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

$$H = E_f[h(Y)] = \int h(y)f(y)dy$$

- 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.

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

Importance Sampling

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

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

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

Example: Gaussian Tail Probability

$P(-3 < Y < 3)$

$= 0.997$

Method 1 : $C = P(Y>8) = E[I(Y>8)]$

$Y_1, \dots Y_n \sim f.$ $f \sim N(0,1)$.

$\approx \frac{1}{n} \sum_{j=1}^{n} I(Y_j > 8)$

$h(y) = I(y>8)$

$= \begin{cases} 1 & \text{if } y>8 \\ 0 & \text{otherwise.} \end{cases}$

10^{-16}

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

Method 2 : choose $g \sim N(8,1)$, $Y_1, \dots Y_n \sim g$

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

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

$n = 50000$; $c \approx 6.25 \times 10^{-16}$

Solution

Outline

Sample Mean: Recall

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.

Strong Law of Large Numbers (SLLN)

① $X_1, \ldots X_n$ i.i.d. r.v. g : continuous function

$g(X_1) \ldots g(X_n)$ i.i.d.

$E[g(X)] = \int_a^b g(x) \frac{1}{b-a} dx$

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.

$E[X]$

② By SLLN $\frac{g(X_1) + \cdots + g(X_n)}{n} \xrightarrow[n \to \infty]{w.p.1} E[g(X)]$

$= \int_a^b g(x) \frac{1}{b-a} dx$

$\Rightarrow \frac{(b-a)}{n} \sum_{i=1}^{n} g(X_i) \xrightarrow[n \to \infty]{w.p.1} \int_a^b g(x) dx$

Weak Law of Large Numbers (WLLN)

$X_n \xrightarrow[n\to\infty]{P} X \iff \forall \varepsilon>0, \lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0$

$X_1, X_2, \ldots X_n$

$X_n = \begin{cases} 1 & \text{w.p. } \frac{1}{n} \\ 0 & \text{w.p. } 1-\frac{1}{n} \end{cases}$

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).

$X_n \xrightarrow{P} 0$;

$\forall \varepsilon>0$, $P(|X_n - 0| > \varepsilon) = P(|X_n| > \varepsilon)$

$= P(X_n > \varepsilon) = \begin{cases} 0 & \varepsilon \geq 1 \\ \frac{1}{n} & 0<\varepsilon<1 \end{cases}$

$= P(X_n = 1)$

$= \frac{1}{n}$

$\Leftarrow$

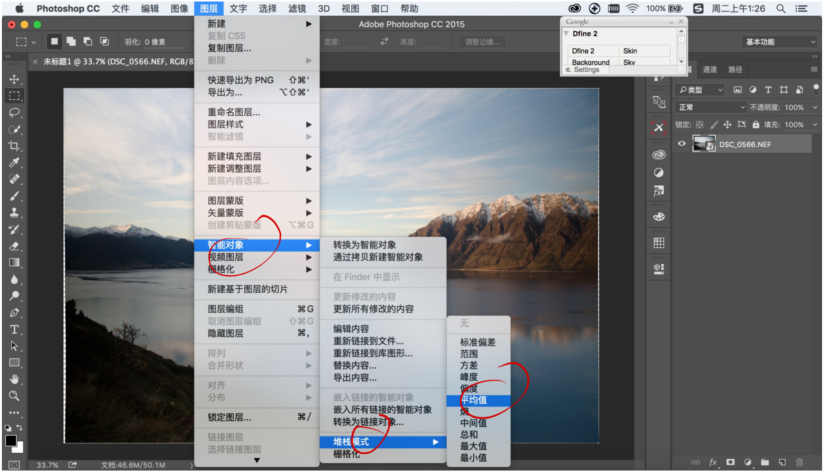
$0<\varepsilon<1$

$\lim_{n\to\infty} P(|X_n - 0| > \varepsilon) = \lim_{n\to\infty} \frac{1}{n} = 0$

$\varepsilon \geq 1$, $\lim_{n\to\infty} P(|X_n - 0| > \varepsilon) = 0$

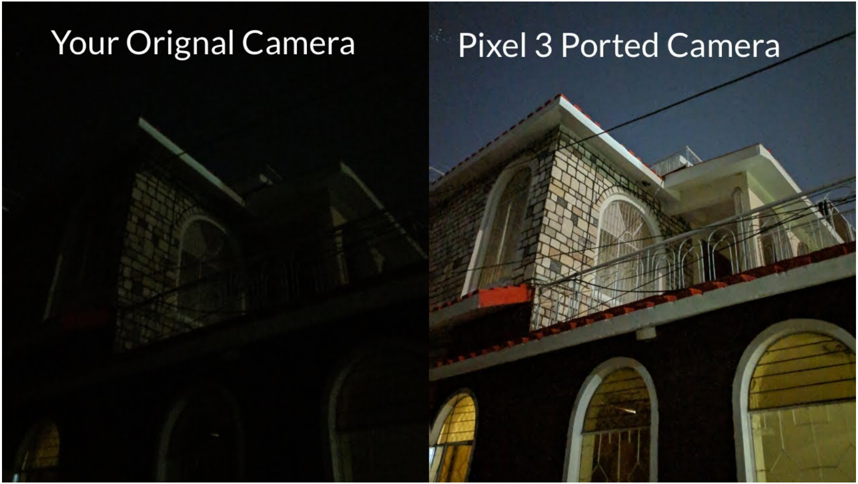
Widely Applications: Photo Stacking with PC

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

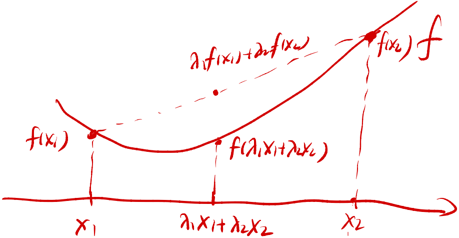
Cauchy-Schwarz Inequality: Recall

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



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 ; $E[g(X)] \geq g[E(X)]$; $g''(\cdot) \geq 0$.
concave ; $\leq$; $g''(\cdot) \leq 0$.

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

$var(X) = E(X^2) - (E(X))^2 \geq 0$

2°. $g(x) = \frac{1}{x}$, $x > 0$, convex, $\Rightarrow$ $E[\frac{1}{X}] \geq \frac{1}{E(X)}$, ✓

3°. $g(x) = \log x$, $x > 0$, concave $\Rightarrow$ $E[\log X] \leq \log(E(X))$.

Entropy

- Let X be a discrete r.v. whose distinct possible values are $a_1, a_2, \ldots, a_n$, with probabilities $p_1, p_2 \ldots, p_n$ respectively (so $p_1 + p_2 + \cdots + p_n = 1$).
- The *entropy* of X is defined as follows:
  $H(X) = \sum_{j=1}^{n} p_j \log_2 (1/p_j)$. $E[\log Y]$
- 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

$$Y = \begin{cases} \frac{1}{p_1} & \text{w.p. } p_1 \\ \frac{1}{p_2} & \text{w.p. } p_2 \\ \vdots & \\ \frac{1}{p_n} & \text{w.p. } p_n \end{cases} \Rightarrow E(Y) = \frac{1}{p_1}\cdot p_1 + \frac{1}{p_2}\cdot p_2 + \cdots + \frac{1}{p_n}\cdot p_n = n$$

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

$\forall p_1, \ldots, p_n$, $p_1 + \cdots + p_n = 1$ $\Rightarrow$ $\max_{p_1, \ldots, p_n} H(X) \le \log_2 n$

③ when $X \sim \mathrm{Unif}(\frac{1}{n})$, $p_1 = p_2 = \cdots = p_n = \frac{1}{n}$, $H(X) = \sum_{j=1}^{n} \frac{1}{n} \log_2 n = \log_2 n$

$\Rightarrow \max_{p_1, \ldots, p_n} H(X) \ge \log_2 n$ $\Rightarrow \max_{p_1, \ldots, p_n} H(X) = \log_2 n$

Kullback-Leibler Divergence

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

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

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

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

③ $D(p,r) = -E[\log_2 Y] \geq -\log_2(E[Y]) = -\log_2 1 = 0$

Markov's Inequality

Chebyshev

Markov Lyapunov

$P(|X - E(x)| \geq a)$

$a \uparrow$ $\propto$ $\frac{1}{a}$

Prob $\downarrow$ $\frac{1}{a^2}$

$\frac{1}{e^a}$

Theorem

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

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

$O(\frac{1}{a})$

Proof

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

① $\underline{Y = \frac{1}{a}|X| \geq 0}$; $\underline{I(Y \geq 1) \leq Y}$

		LHS		RHS.
$Y \geq 1$		1	$\overset{\checkmark}{\leq}$	Y.
$0 \leq Y < 1$		0	$\overset{\checkmark}{\leq}$	Y.

② $\underline{E[I(Y \geq 1)]} \leq E[Y]$

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

$\Updownarrow$

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

$\Updownarrow$

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

Chebyshev's Inequality

Markov's inequality

$P(|X-\mu| \geq a) = P(|X-\mu|^2 \geq a^2) \leq \frac{1}{a^2} E[|X-\mu|^2]$

$= \frac{1}{a^2} Var(X)$

$= \frac{1}{a^2} \sigma^2$.

Theorem

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

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

$O(\frac{1}{a^2})$

Application : $X_1, \ldots X_n$ i.i.d. (μ, σ^2) Sample mean $\bar{X}_n$. $E(\bar{X}_n) = \mu$

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

$P(|\bar{X}_n - \mu| \geq a) \leq \frac{1}{a^2} Var(\bar{X}_n) = \frac{\sigma^2}{a^2 \cdot n} \xrightarrow[n\to\infty]{} 0$

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

Proof

Chernoff's Inequality

$= P(tX \geq ta)$

$\forall t > 0$, $P(X \geq a) = P(e^{tX} \geq e^{ta})$

Markov's inequality

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

Theorem

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

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

MGF.

$\forall t > 0$, $P(X \geq a) \leq f(t)$

$\Rightarrow P(X \geq a) \leq \inf_{t > 0} f(t)$

Proof

Chernoff's Technique

$\forall t<0, \ P(X \le a)$

$= P(tX \ge ta)$

$= P(e^{tX} \ge e^{ta})$

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

Theorem

For any r.v. X and constants a,

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

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

Proof

Example: Normal Distribution

$= E[e^{tX}]$

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

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

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

$f(t) = \frac{M_X(t)}{e^{ta}} = e^{\frac{1}{2}\sigma^2 t^2 + (\mu - a)t}$

$= e^{\frac{1}{2}\sigma^2 \left[(t + \frac{\mu - a}{\sigma^2})^2 - \frac{(\mu-a)^2}{\sigma^4} \right]}$; $\Rightarrow t^* = \frac{a-\mu}{\sigma^2} > 0$

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

$a = \mu + \varepsilon$ $\Rightarrow P(X > \mu + \varepsilon) \leq e^{-\frac{\varepsilon^2}{2\sigma^2}}$

Solution

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 $\epsilon \geq 0$,

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

$O(e^{-n\varepsilon^2})$

Application: Parameter Estimation

$\hat{p} - \varepsilon \le p \le \hat{p} + \varepsilon \iff -\varepsilon \le p - \hat{p} \le \varepsilon$

$\iff -\varepsilon \le \hat{p} - p \le \varepsilon$

$\iff |\hat{p} - p| \le \varepsilon$

Instead of predicting a single value $\hat{p}$ for the parameter p, 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

$\delta = 0.05.$

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

$Pr(|\hat{p} - p| \le \varepsilon) \ge 1 - \delta$

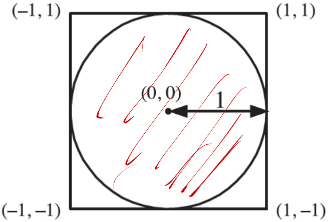
$Pr(|\hat{p} - p| > \varepsilon) \le \delta$

Application Example: Monte Carlo Method for Estimation π

① (x_i, y_i) $(-1 \le x_i \le 1,\ -1 \le y_i \le 1)$

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

$Z_i = I_{\{(x_i,y_i):\, x_i^2+y_i^2 \le 1\}}$



$P(Z_i = 1) = \frac{\pi}{4}$

$E(Z_i) = \frac{\pi}{4}$

② $W \triangleq \frac{1}{n}\sum_{i=1}^{n} Z_i$

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

$E(W) = \frac{\pi}{4}$

$\hat{\pi} = 4W = 4 \cdot \frac{1}{n} \sum_{i=1}^{n} Z_i$

Example: Monte Carlo Method for Estimation π

③ $n \to \infty$, $\hat{\pi} \to \pi$

n is finite

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

$Z_1, \ldots, Z_n, Z$ i.i.d. Bern $(\frac{\pi}{4})$

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

Hoeffding's inequality

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

$a = 0$, $b = 1$

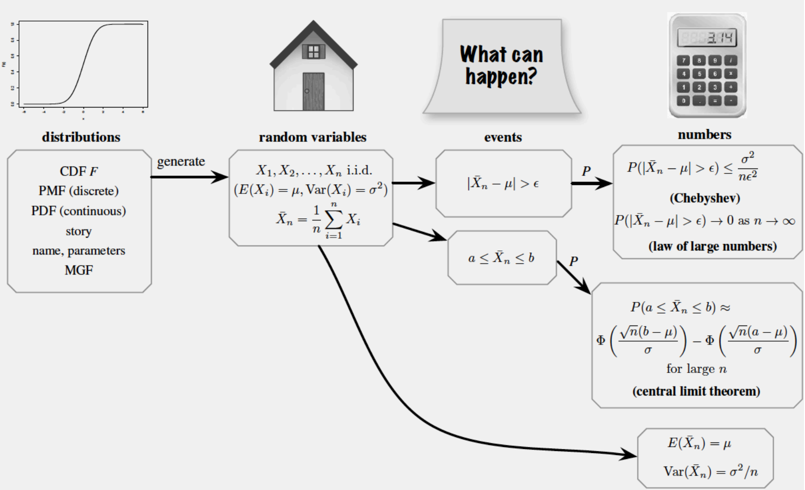
$$\Rightarrow \varepsilon = \sqrt{\frac{8\log(\frac{2}{\delta})}{n}}$$

, $\delta = 0.05$

$$\Rightarrow \Pr\left(\pi \in \left(\hat{\pi} - \sqrt{\frac{8\log(\frac{2}{\delta})}{n}}, \hat{\pi} + \sqrt{\frac{8\log(\frac{2}{\delta})}{n}}\right)\right) \geq 1 - \delta$$

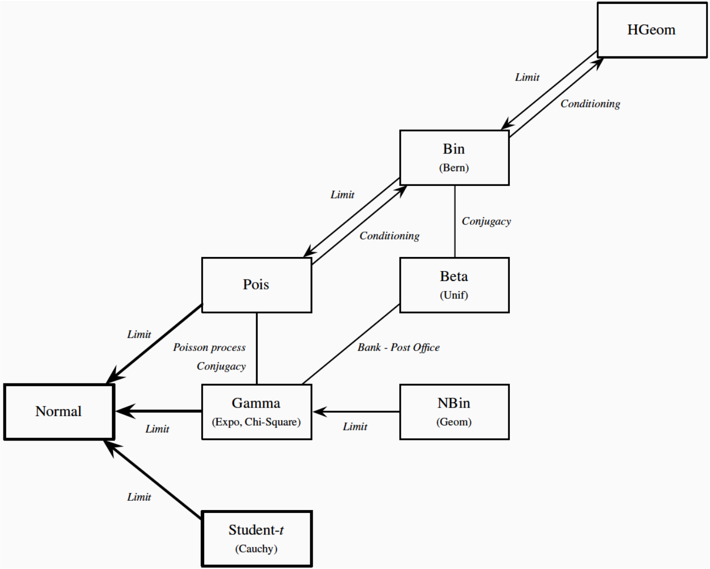
Example: Monte Carlo Method for Estimation π

Summary 1



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