Lecture 1: Probability & Statistics Review

Ziyu Shao

School of Information Science and Technology
ShanghaiTech University

February 26, 2025

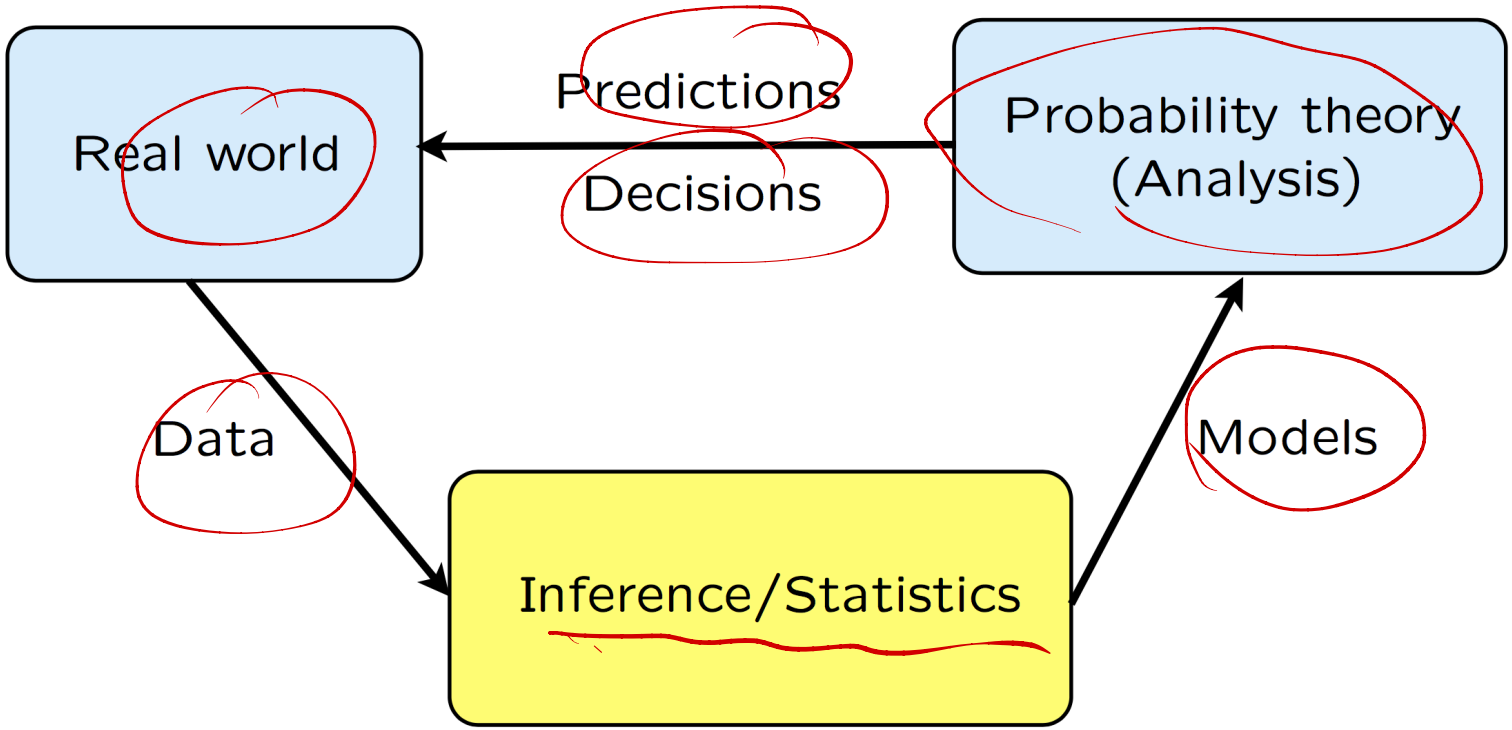
Outline

1. Stochastic Laws in Random Worlds
2. Statistical Inference
3. Conjugate Prior: A Weapon of Bayesian
4. Conditional Expectation
5. Monte Carlo Method: Another Weapon of Bayesian
6. Sampling: Random Variable Generation
7. Sampling: Random Vector Generation
8. Monte Carlo Integration
9. Performance Analysis of Monte Carlo Integration
10. References

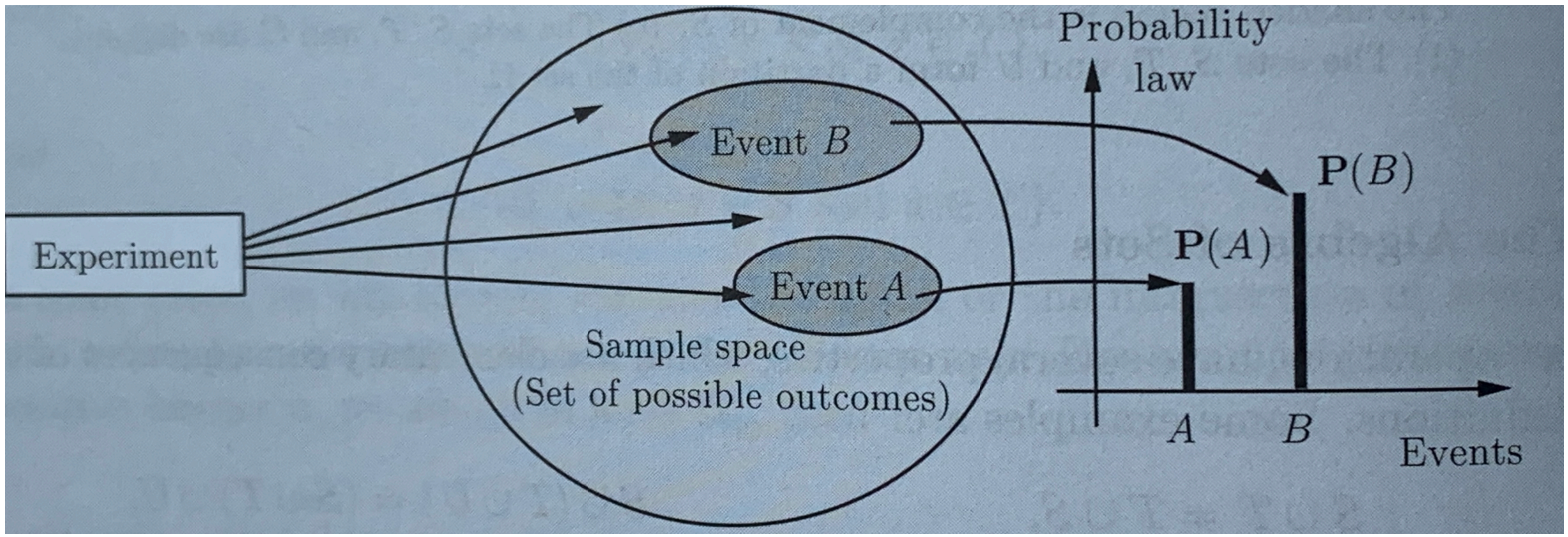
Outline

1. Stochastic Laws in Random Worlds
2. Statistical Inference
3. Conjugate Prior: A Weapon of Bayesian
4. Conditional Expectation
5. Monte Carlo Method: Another Weapon of Bayesian
6. Sampling: Random Variable Generation
7. Sampling: Random Vector Generation
8. Monte Carlo Integration
9. Performance Analysis of Monte Carlo Integration
10. References

Understanding Random Worlds: Framework



Understanding Random Worlds: Modeling



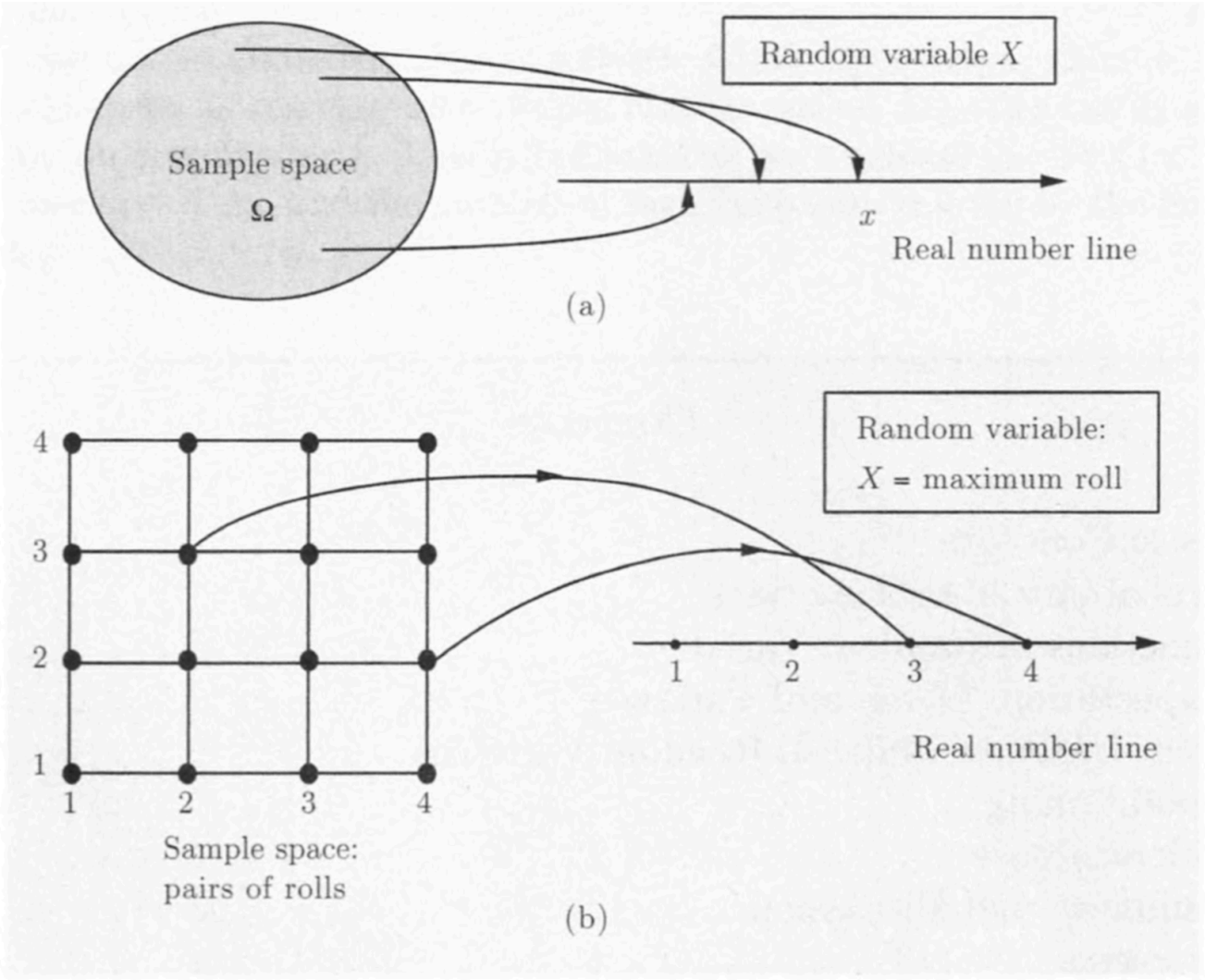
Bayes' Rule

Theorem

For any events A and B with positive probabilities,

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Understanding Random Worlds: Modeling



General Bayes' Rule

	Y **discrete**	Y **continuous**
X **discrete**	$P(Y=y\|X=x)=\frac{P(X=x\|Y=y)P(Y=y)}{P(X=x)}$	$f_Y(y\|X=x)=\frac{P(X=x\|Y=y)f_Y(y)}{P(X=x)}$
X **continuous**	$P(Y=y\|X=x)=\frac{f_X(x\|Y=y)P(Y=y)}{f_X(x)}$	$f_{Y\|X}(y\|x)=\frac{f_{X\|Y}(x\|y)f_Y(y)}{f_X(x)}$

General LOTP

	Y **discrete**	Y **continuous**
X **discrete**	$P(X=x)=\sum_y P(X=x\|Y=y)P(Y=y)$	$P(X=x)=\int_{-\infty}^{\infty} P(X=x\|Y=y)f_Y(y)dy$
X **continuous**	$f_X(x)=\sum_y f_X(x\|Y=y)P(Y=y)$	$f_X(x)=\int_{-\infty}^{\infty} f_{X\|Y}(x\|y)f_Y(y)dy$

Outline

1. Stochastic Laws in Random Worlds
2. Statistical Inference
3. Conjugate Prior: A Weapon of Bayesian
4. Conditional Expectation
5. Monte Carlo Method: Another Weapon of Bayesian
6. Sampling: Random Variable Generation
7. Sampling: Random Vector Generation
8. Monte Carlo Integration
9. Performance Analysis of Monte Carlo Integration
10. References

Important Concepts

- Population: distribution $\mathcal{F}$
- Sample: random vector $\boldsymbol{X} = (X_1, \ldots, X_n)$ where n is the sample size
- Random Sample: $\{X_i\}$ are i.i.d random variables and $X_i \sim \mathcal{F}$
- Data: real vector $\boldsymbol{x} = (x_1, \ldots, x_n)$, the value of sample $\boldsymbol{X}$
- From sample to infer property of Population
- Statistic: a function of sample $\boldsymbol{X}$

$$\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$$

Important Concepts

Difference between an estimate and an estimator

- An estimator is a function of the sample
- An estimator is a function of the random variables $X_1, \ldots, X_n$.
- An estimate is the realized value of an estimator (a number) that is obtained when a sample is actually taken
- An estimate is a function of the realized values $x_1, \ldots, x_n$.

Nonparametric Models

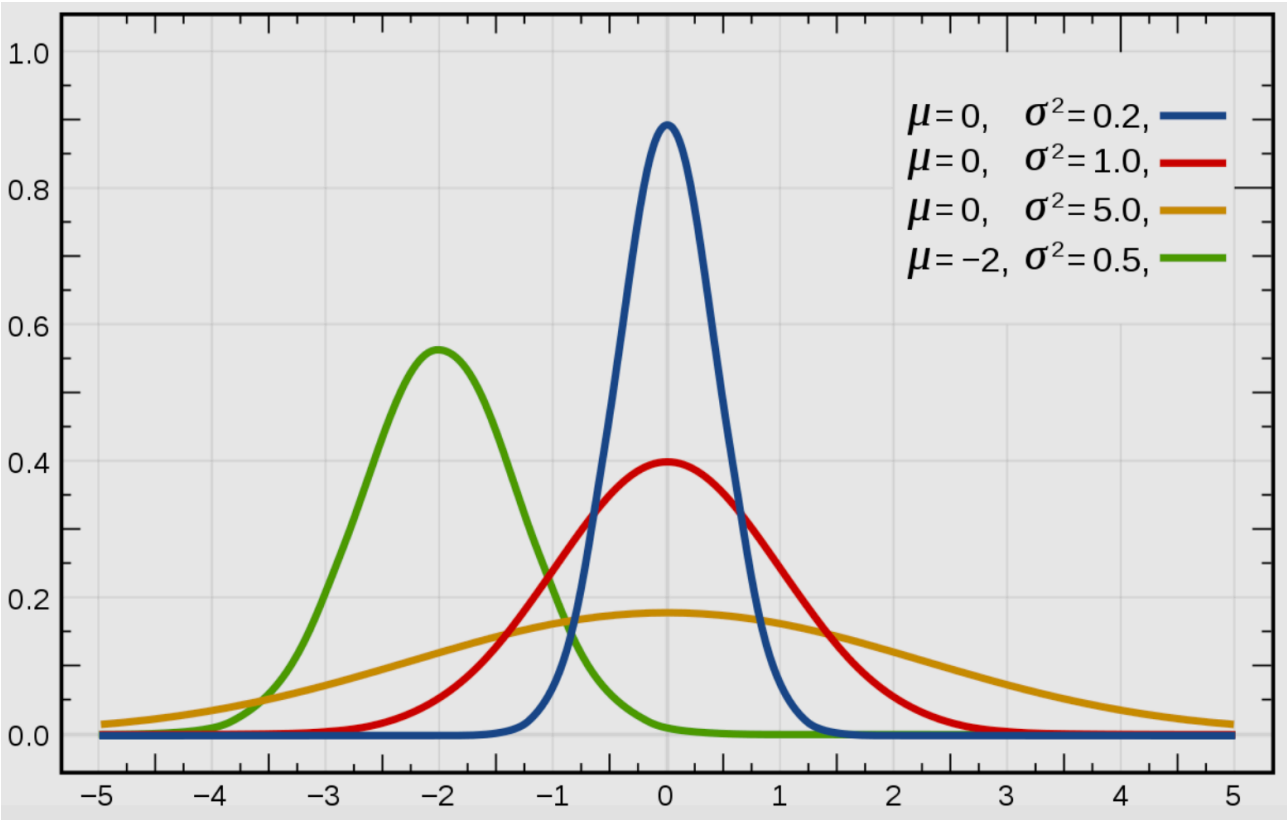
- A nonparametric model is a statistical model that cannot be parameterized by a finite number of parameters.
- Example: (nonparametric estimation of the CDF) Let $X_1, \ldots, X_n$ be independent observations from a CDF F. Our task is to estimate F.

Our Focus: Parameterized Statistical Inference

- Given a parametric distribution model (a family of PMFs or PDFs) $\mathcal{F} = \{p(x;\theta) : \theta \in \mathcal{R}\}$
- θ is an unknown parameter in a parameter space $\mathcal{R}$
- Now given (random)sample from such model: $\boldsymbol{X} = (X_1, \ldots, X_n)$
- How to make parameterized statistical inference?

Example: Parameterized Distribution Model

$$\left\{ p(x; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2 \right), \mu \in \mathbb{R}, \sigma > 0 \right\}.$$

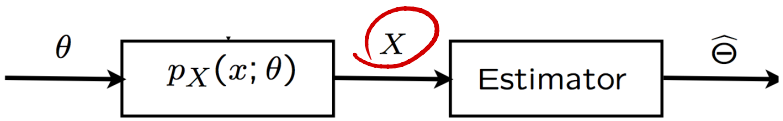


Parameterized Statistical Inference: Bayesian versus Frequentist

- Difference relates to the nature of the unknown parameter θ
- Treated as a random variable Θ with prior (known) distribution: **Bayesian approach**
- Treated as an unknown constant θ: **frequentist approach**

Statistical Inference: Frequentist Perspective

- Classical statistics: unknown constant θ



- Hypothesis testing: $H_0 : \theta = 1/2$ versus $H_1 : \theta = 3/4$
- Composite hypotheses: $H_0 : \theta = 1/2$ versus $H_1 : \theta \neq 1/2$
- Estimation: design an **estimator** $\widehat{\Theta}$, to "keep estimation **error** $\widehat{\Theta} - \theta$ small"

Estimation Method: Maximum Likelihood Estimation (MLE)

- We observe a particular data $\boldsymbol{x} = (x_1, \ldots, x_n)$,
- Likelihood: the probability (or probability density) of seeing data $\boldsymbol{x}$ under different values of parameter θ, i.e., $p(\boldsymbol{x}; \theta)$
- A **maximum likelihood estimate** (MLE) is a value of the parameter θ that maximizes the likelihood $p(\boldsymbol{x}; \theta)$ over all possible values:

$$\hat{\theta} = \arg\max_{\theta} p(\boldsymbol{x}; \theta)$$

MLE under Independent Case

- Random Sample: $\{X_i\}$ are i.i.d., we have

$$\log[p(\boldsymbol{x};\theta)] = \log \prod_{i=1}^{n} p(x_i;\theta) = \sum_{i=1}^{n} \log[p(x_i;\theta)]$$

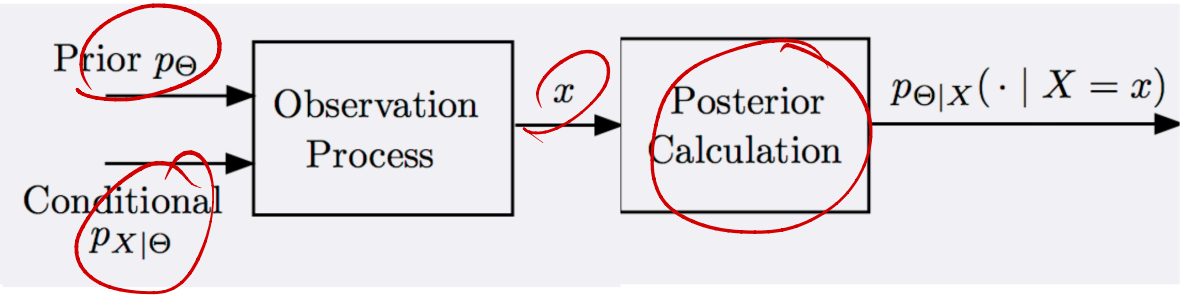
- Thus a **maximum likelihood estimate** (MLE) under independent case is shown as follows:

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} p(\boldsymbol{x};\theta) = \arg\max_{\theta} \log[p(\boldsymbol{x};\theta)] \\ &= \arg\max_{\theta} \sum_{i=1}^{n} \log[p(x_i;\theta)]\end{aligned}$$

Statistical Inference: Bayesian Perspective

- Unknown Θ
  - treated as a random variable
  - prior distribution p_Θ or f_Θ
- Observation X
  - observation model $p_{X|\Theta}$ or $f_{X|\Theta}$
- Use appropriate version of the Bayes rule to find $p_{\Theta|X}(\cdot \mid X = x)$ or $f_{\Theta|X}(\cdot \mid X = x)$

- Where does the prior come from?
  - symmetry
  - known range
  - earlier studies
  - subjective or arbitrary

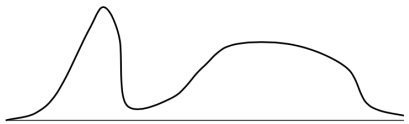


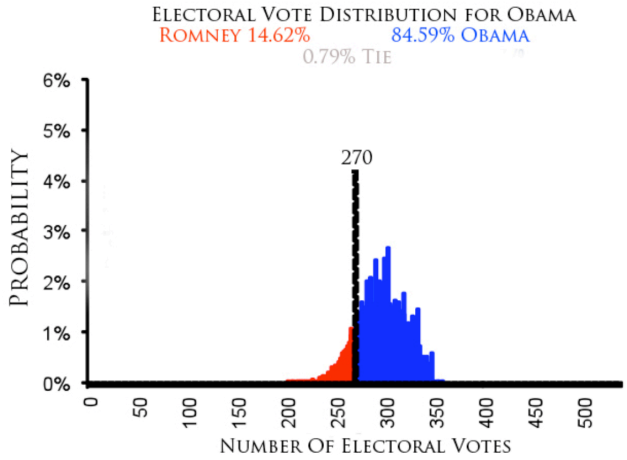
The Output of Bayesian Statistical Inference

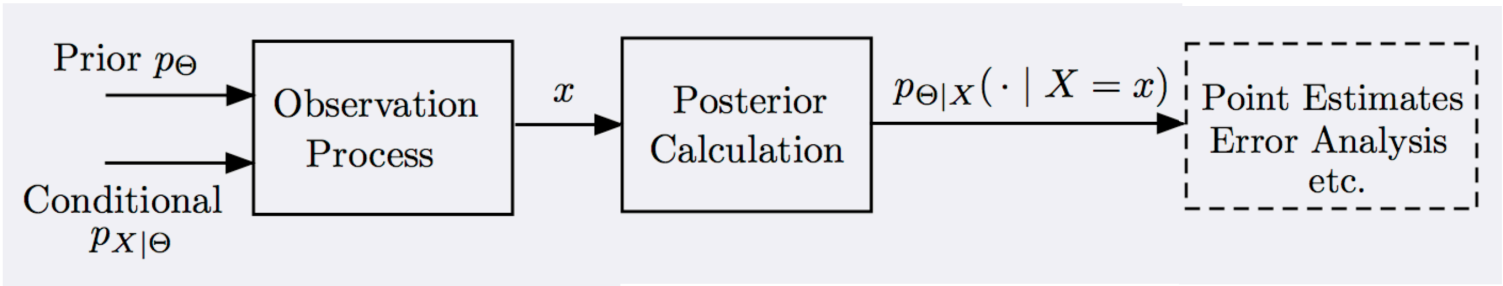
The complete answer is a posterior distribution:
PMF $p_{\Theta|X}(\cdot \mid x)$ or PDF $f_{\Theta|X}(\cdot \mid x)$

$p_{\Theta|X}(\cdot \mid x)$

$f_{\Theta|X}(\cdot \mid x)$







General Bayes' Rule for Bayesian Statistical Inference

The Four Versions of Bayes' Rule

- Θ discrete, X discrete:

$$p_{\Theta|X}(\theta \,|\, x) = \frac{p_\Theta(\theta) p_{X|\Theta}(x \,|\, \theta)}{\sum_{\theta'} p_\Theta(\theta') p_{X|\Theta}(x \,|\, \theta')}.$$

$\propto P_{\Theta}(\theta) \cdot P_{X|\Theta}(x|\theta)$

- Θ discrete, X continuous:

$$p_{\Theta|X}(\theta \,|\, x) = \frac{p_\Theta(\theta) f_{X|\Theta}(x \,|\, \theta)}{\sum_{\theta'} p_\Theta(\theta') f_{X|\Theta}(x \,|\, \theta')}.$$

- Θ continuous, X discrete:

$$f_{\Theta|X}(\theta \,|\, x) = \frac{f_\Theta(\theta) p_{X|\Theta}(x \,|\, \theta)}{\int f_\Theta(\theta') p_{X|\Theta}(x \,|\, \theta')\, d\theta'}.$$

- Θ continuous, X continuous:

$$f_{\Theta|X}(\theta \,|\, x) = \frac{f_\Theta(\theta) f_{X|\Theta}(x \,|\, \theta)}{\int f_\Theta(\theta') f_{X|\Theta}(x \,|\, \theta')\, d\theta'}.$$

Bayes' Rule: Bayesian Perspective

Theorem

A represents cognition, B represents data, then

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

- $P(A)$: *Prior Cognition*
- $P(B)$: *Evidence*
- $P(B|A)$: *Likelihood (info of data given prior cognition)*
- $P(A|B)$: *Posterior Cognition*

Estimation Method I

- The Maximum A Posteriori Probability (MAP)
- Given the observation value x, the MAP rule selects a value $\hat{\theta}$ that maximizes the posterior probability(probability density) $p_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$:

$$\hat{\theta} = \arg\max_{\theta} p_{\Theta|\boldsymbol{X}}(\theta|\boldsymbol{x})$$

- Equivalently,

$$\hat{\theta} = \arg\max_{\theta} p_{\Theta}(\theta) p_{\boldsymbol{X}|\Theta}(\boldsymbol{x}|\theta)$$

Estimation Method II

- Posterior Mean: Given the observation data $\boldsymbol{x}$, the estimate of Θ is $\hat{\theta} = E[\Theta|\boldsymbol{X} = \boldsymbol{x}]$
- Now given the observation sample $\boldsymbol{X}$, the estimator of Θ is $\hat{\Theta} = E[\Theta|\boldsymbol{X}]$

Outline

1. Stochastic Laws in Random Worlds
2. Statistical Inference
3. Conjugate Prior: A Weapon of Bayesian
4. Conditional Expectation
5. Monte Carlo Method: Another Weapon of Bayesian
6. Sampling: Random Variable Generation
7. Sampling: Random Vector Generation
8. Monte Carlo Integration
9. Performance Analysis of Monte Carlo Integration
10. References

Conjugate Prior

- Before Monte Carlo, posterior calculation is hard
- Conjugate Prior: reduce the computing complexity of posterior distribution
- Loosely speaking, a prior distribution is conjugate to the likelihood model if both the prior and posterior distribution stay in the same distribution family.

Story: Beta-Binomial Conjugacy

We have a coin that lands Heads with probability p , but we don't know what p is. Our goal is to infer the value of p after observing the outcomes of n tosses of the coin. The larger that n is, the more accurately we should be able to estimate p.

Bayesian Inference

- Treats all unknown quantities as random variables.
- In the Bayesian approach, we would treat the unknown probability p as a random variable and give p a distribution.
- This is called a **prior distribution**, and it reflects our uncertainty about the true value of p before observing the coin tosses.
- After the experiment is performed and the data are gathered, the prior distribution is updated using Bayes' rule; this yields the **posterior distribution**, which reflects our new beliefs about p.
- Now we adopt the beta distribution as the prior distribution.

Beta Distribution

$a=1, b=1$;

$Unif(0,1)$

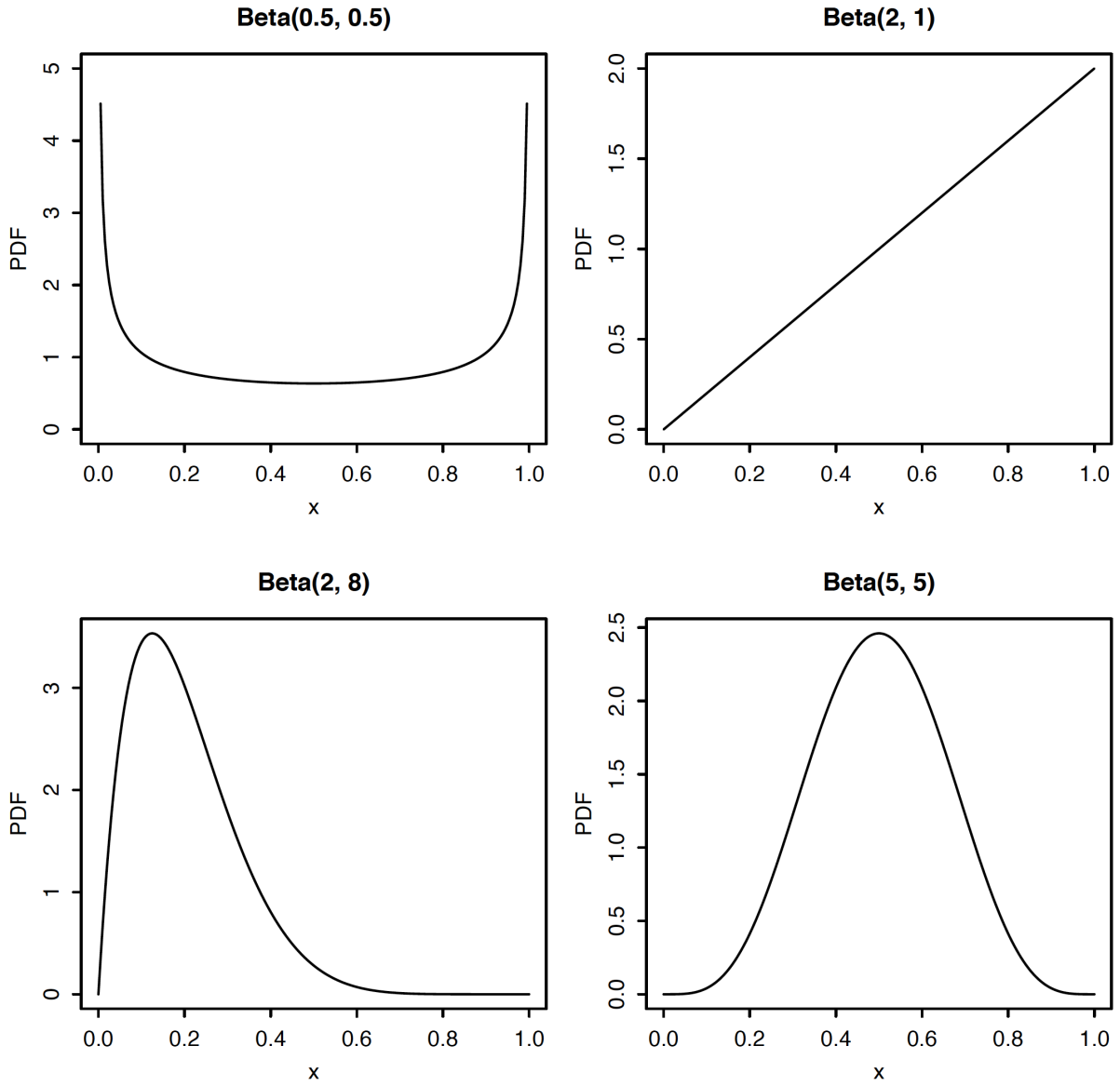
Definition

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

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

where the constant $\beta(a,b)$ is chosen to make the PDF integrate to 1. It is denoted as $X \sim \mathrm{Beta}(a,b)$. Beta distribution is a generalization of uniform distribution.

PDF of Beta Distribution



Story: Beta-Binomial Conjugacy

① P: r.v. Prior. Beta

X: # of heads out of n tosses

$X \mid P=p \sim \mathrm{Bin}(n, p)$

Likelihood

② $f_P(p \mid X=k) = \dfrac{\Pr(X=k \mid P=p)}{\Pr(X=k)} \cdot f_P(p)$

$= \dfrac{\binom{n}{k} p^k (1-p)^{n-k}}{\Pr(X=k)} \cdot \dfrac{1}{B(a,b)} p^{a-1}(1-p)^{b-1}$

$\Pr(X=k) = \int_0^1 \Pr(X=k \mid P=p) f(p)\, dp$

③ $f_P(p \mid X=k)$: a function of p

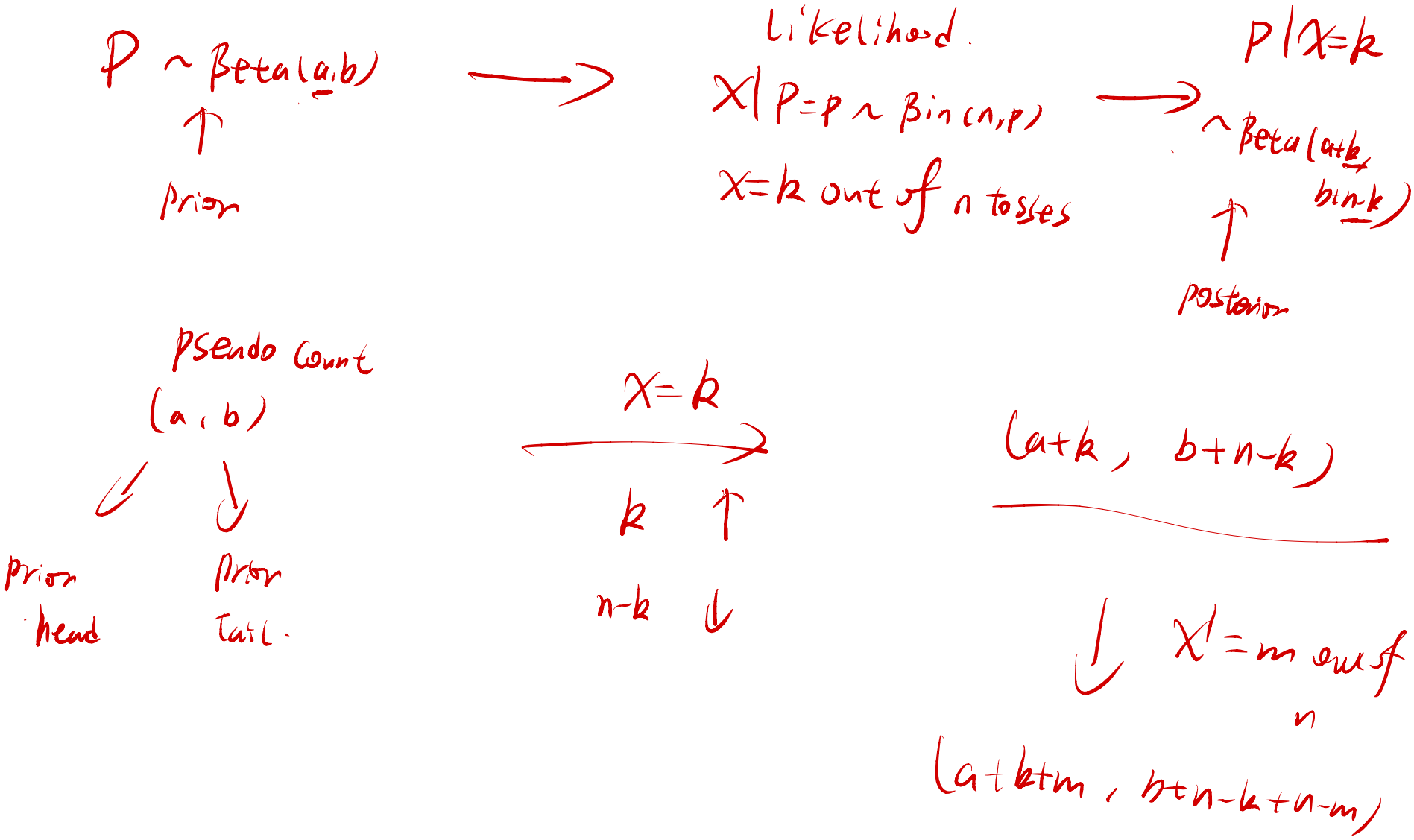
$f_P(p \mid X=k) \propto p^k (1-p)^{n-k} \cdot p^{a-1}(1-p)^{b-1}$

$0<p<1$

$= p^{k+a-1}(1-p)^{n-k+b-1}$

$\sim \mathrm{Beta}(k+a, n-k+b)$

Story: Beta-Binomial Conjugacy



Story: Beta-Binomial Conjugacy

Story: Beta-Binomial Conjugacy

- Furthermore, notice the very simple formula for updating the distribution of p.
- We just add the number of observed successes, k, to the first parameter of the Beta distribution.
- We also add the number of observed failures, $n - k$, to the second parameter of the Beta distribution.
- So a and b have a concrete interpretation in this context:
  - a as the number of prior successes in earlier experiments
  - b as the number of prior failures in earlier experiments
  - a, b: pseudo counts

Mean vs. Bayesian(Posterior) Average

- Infer the value of p (probability of coin lands heads)
- Observed k heads out of n tosses of the coin
- Mean: $\frac{k}{n}$
- Bayesian(Posterior) Average: $E(p|X=k) = \frac{a+k}{a+b+n}$
- Suppose the prior distribution is Unif(0,1): $a = 1, b = 1$
- Bayesian(Posterior) Average: $\frac{k+1}{n+2}$
- When $k = n$, we have: 1 (mean) vs. $\frac{n+1}{n+2}$ (Bayesian average)

Story: Beta-Binomial Conjugacy

If we have a Beta prior distribution on p and data that are conditionally Binomial given p, then when going from prior to posterior, we don't leave the family of Beta distributions. We say that **the Beta distribution is the conjugate prior of the Binomial distribution**.

Likelihood Model: Discrete

Sample Space	Sampling Dist.	Conjugate Prior	Posterior
$\mathcal{X} = \{0, 1\}$	Bernoulli(θ)	Beta(α, β)	Beta$(\alpha + n\overline{X}, \beta + n(1 - \overline{X}))$
$\mathcal{X} = \mathbb{Z}_+$	Poisson(λ)	Gamma(α, β)	Gamma$(\alpha + n\overline{X}, \beta + n)$
$\mathcal{X} = \mathbb{Z}_{++}$	Geometric(θ)	Gamma(α, β)	Gamma$(\alpha + n, \beta + n\overline{X})$
$\mathcal{X} = \mathbb{H}_K$	Multinomial(θ)	Dirichlet(α)	Dirichlet$(\alpha + n\overline{X})$

Likelihood Model: Continuous

Sampling Dist.	Conjugate Prior	Posterior
Uniform(θ)	Pareto(ν_0, k)	Pareto $\left(\max\{\nu_0, X_{(n)}\}, n+k\right)$
Exponential(θ)	Gamma(α, β)	Gamma$(\alpha+n, \beta+n\overline{X})$
$N(\mu, \sigma^2)$, known σ^2	$N(\mu_0, \sigma_0^2)$	$N\left(\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}\left(\frac{\mu_0}{\sigma_0^2}+\frac{n\overline{X}}{\sigma^2}\right), \left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}\right)$
$N(\mu, \sigma^2)$, known μ	InvGamma(α, β)	InvGamma $\left(\alpha+\frac{n}{2}, \beta+\frac{n}{2}\overline{(X-\mu)^2}\right)$
$N(\mu, \sigma^2)$, known μ	ScaledInv-$\chi^2(\nu_0, \sigma_0^2)$	ScaledInv-$\chi^2\left(\nu_0+n, \frac{\nu_0\sigma_0^2}{\nu_0+n}+\frac{n\overline{(X-\mu)^2}}{\nu_0+n}\right)$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, known $\boldsymbol{\Sigma}$	$N(\boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$	$N\left(\mathbf{K}\left(\boldsymbol{\Sigma}_0^{-1}\mu_0+n\boldsymbol{\Sigma}^{-1}\overline{X}\right), \mathbf{K}\right)$, $\mathbf{K}=\left(\boldsymbol{\Sigma}_0^{-1}+n\boldsymbol{\Sigma}^{-1}\right)^{-1}$
$N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, known $\boldsymbol{\mu}$	InvWishart$(\nu_0, \mathbf{S}_0)$	InvWishart$(\nu_0+n, \mathbf{S}_0+n\overline{\mathbf{S}})$, $\overline{\mathbf{S}}$ sample covariance

Outline

1. Stochastic Laws in Random Worlds
2. Statistical Inference
3. Conjugate Prior: A Weapon of Bayesian
4. Conditional Expectation
5. Monte Carlo Method: Another Weapon of Bayesian
6. Sampling: Random Variable Generation
7. Sampling: Random Vector Generation
8. Monte Carlo Integration
9. Performance Analysis of Monte Carlo Integration
10. References

Motivation

- Conditional expectation is a powerful tool for calculating expectations: first-step analysis
- Conditional expectation allows us to predict or estimate unknowns based on whatever evidence is currently available.
- Conditional Expectation given an event: $E(Y|A)$
- Conditional Expectation given a random variable: $E(Y|X)$

Conditional Expectation Given An Event

Definition

Let A be an event with positive probability. If Y is a discrete r.v., then the *conditional expectation of Y given A* is

$$E(Y|A) = \sum_y yP(Y = y|A),$$

where the sum is over the support of Y. If Y is a continuous r.v. with PDF f, then

$$E(Y|A) = \int_{-\infty}^{\infty} yf(y|A)\,dy,$$

where the conditional PDF $f(y|A)$ is defined as the derivative of the conditional CDF $F(y|A) = P(Y \leq y|A)$, and can also be computed by a hybrid version of Bayes' rule:

$$f(y|A) = \frac{P(A|Y = y)\,f(y)}{P(A)}.$$

Intuition for $E(Y|A)$

$E(Y) \approx \frac{1}{n}(Y_1 + \cdots + Y_n)$

Principle

$E(Y|A)$ is approximately the average of Y in a large number of simulation runs in which A occurred.

Life Expectancy

$$E[T] = 70;$$

$$E[T \mid T \geq 20]$$

Law of Total Expectation

Theorem

Let $A_1, \ldots, A_n$ be a partition of a sample space, with $P(A_i) > 0$ for all i, and let Y be a random variable on this sample space. Then

$$E(Y) = \sum_{i=1}^{n} E(Y|A_i)\, P(A_i)\,.$$

Conditional Expectation Given An R.V.

Definition

Let $g(x) = E(Y|X = x)$. Then the *conditional expectation of Y given X*, denoted $E(Y|X)$, is defined to be the random variable $g(X)$. In other words, if after doing the experiment X crystallizes into x, then $E(Y|X)$ crystallizes into $g(x)$.

Remark

- $E(Y|X)$ is a function of X, and it is a random variable.
- It makes sense to computer $E(E(Y|X))$ and $Var(E(Y|X))$.

Property I: Dropping What's Independent

$g(x) = E[Y|X=x] = E[Y]$, $\forall x$.

$\Rightarrow g(x) = E[Y]$

$\Rightarrow E[Y|X] = g(X) = E[Y]$

Theorem

If X and Y are independent, then $E(Y|X) = E(Y)$.

Property II: Taking Out What's Known

1°. $g(x) = E[h(x)Y \mid X=x]$

$= E[h(x)Y \mid X=x] = h(x)E[Y \mid X=x]$

2°. $g(X) = h(X) \cdot E[Y|X]$

Theorem

For any function h,

$$E(h(X)\,Y|X) = h(X)\,E(Y|X)$$

Property III: Linearity

$$g(x) = E[Y_1 + Y_2 | X = x] = E[Y_1 | X = x] + E[Y_2 | X = x]$$

$$\Rightarrow g(X) = E[Y_1 | X] + E(Y_2 | X)$$

Theorem

$E(Y_1 + Y_2|X) = E(Y_1|X) + E(Y_2|X)$.

Adam's Law

W.L.O.G. X and Y are both discrete r.v.s.

1°: $g(X) = E[Y|X]$:

$g(x) = E[Y|X=x] = \sum_y y \cdot P(Y=y|X=x)$

2°. LHS: $E[E[Y|X]] = E[g(X)]$

$= \sum_x g(x) \cdot P(X=x)$

Theorem (Law of Iterated Expectation)

For any r.v.s X and Y,

$$\underbrace{E(E(Y|X))}_{\text{LHS}} = \underbrace{E(Y)}_{\text{RHS}}.$$

$= \sum_x \left[\sum_y y \cdot P(Y=y|X=x) \right] \cdot P(X=x)$

$= \sum_y y \left[\sum_x P(Y=y|X=x) \cdot P(X=x) \right] = \sum_y y \cdot \left[\sum_x P(Y=y, X=x) \right]$

$= \sum_y y \cdot P(Y=y) = E[Y]$

Proof

Adam's Law with Extra Conditioning

Adam's Law: $E[E[Y|X]] = E[Y]$

$\hat{E}[\hat{E}(Y|X)] = \hat{E}(Y)$

$\hat{E} = E(\cdot|Z)$

$\hat{E}(Y) = E(Y|Z)$

Theorem (Law of Iterated Expectation with Extra Conditioning)

For any r.v.s X, Y, Z, we have

$$E(E(Y|X,Z)|Z) = E(Y|Z)$$

$$E(E(X|Z,Y)|Y) = E(X|Y)$$

$\hat{E}(\hat{E}(Y|X)) = \hat{E}(E[Y|X,Z]) = \hat{E}[E[Y|X,Z]|Z]$

$\hat{E} = E(\cdot|Y)$

Conditional Variance

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

$\hat{E}(\cdot) = E(\cdot|X)$

$Var(Y|X) = \hat{E}[(Y - \hat{E}(Y))^2]$

Definition

The *conditional variance of Y given X* is

$$\mathrm{Var}(Y|X) = E\left((Y - E(Y|X))^2 |X\right).$$

This is equivalent to

$Var(Y) = E(Y^2) - E^2(Y)$

$$\mathrm{Var}(Y|X) = E(Y^2|X) - (E(Y|X))^2.$$

$Var(Y|X) = \hat{E}(Y^2) - \hat{E}^2(Y)$

Eve's law

Theorem (Law of Total Variance)

For any r.v.s X and Y,

$$\operatorname{Var}(Y) = E(\operatorname{Var}(Y|X)) + \operatorname{Var}(E(Y|X)).$$

The ordering of E's and Var's on the right-hand side spells EVVE, whence the name Eve's law. Eve's law is also known as the law of total variance or the variance decomposition formula.

Proof

① $g(X) = E[Y|X] \Rightarrow E[g(X)] = E[E(Y|X)] = E[Y]$

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

$= E[E[Y^2|X] - g^2(X)] = E[E[Y^2|X]] - E[g^2(X)]$

$= E[Y^2] - E[g^2(X)]$

③ $Var[E(Y|X)] = Var[g(X)] = E[g^2(X)] - E^2[g(X)]$

$= E[g^2(X)] - E^2(Y)$ ④

② + ③ $E[Var(Y|X)] + Var[E(Y|X)] = E[Y^2] - E^2(Y)$

$= Var(Y)$

Prediction & Estimation

- Estimate Y from the observed value X
- Choose the estimator (inference function) $g(\cdot)$ to minimize the expected error $E(c(Y, g(X)))$
- $c(Y, \hat{Y})$ is the cost of guessing $\hat{Y}$ when the actually value is Y.
- When $c(Y, \hat{Y}) = ||Y - \hat{Y}||^2$, the best guess is called "the least square estimate (LSE)" estimate of Y given X.
- Further, if the function $g(\cdot)$ is restricted to be linear, i.e., of the form $a + bX$, it is called "the Linear Least Square Estimate (LLSE)" estimate of Y given X.
- Further, if the function $g(\cdot)$ can be arbitrary, it is called "the Minimum Mean Square Estimate (MMSE)" estimate of Y given X.

Linear Least Square Estimate

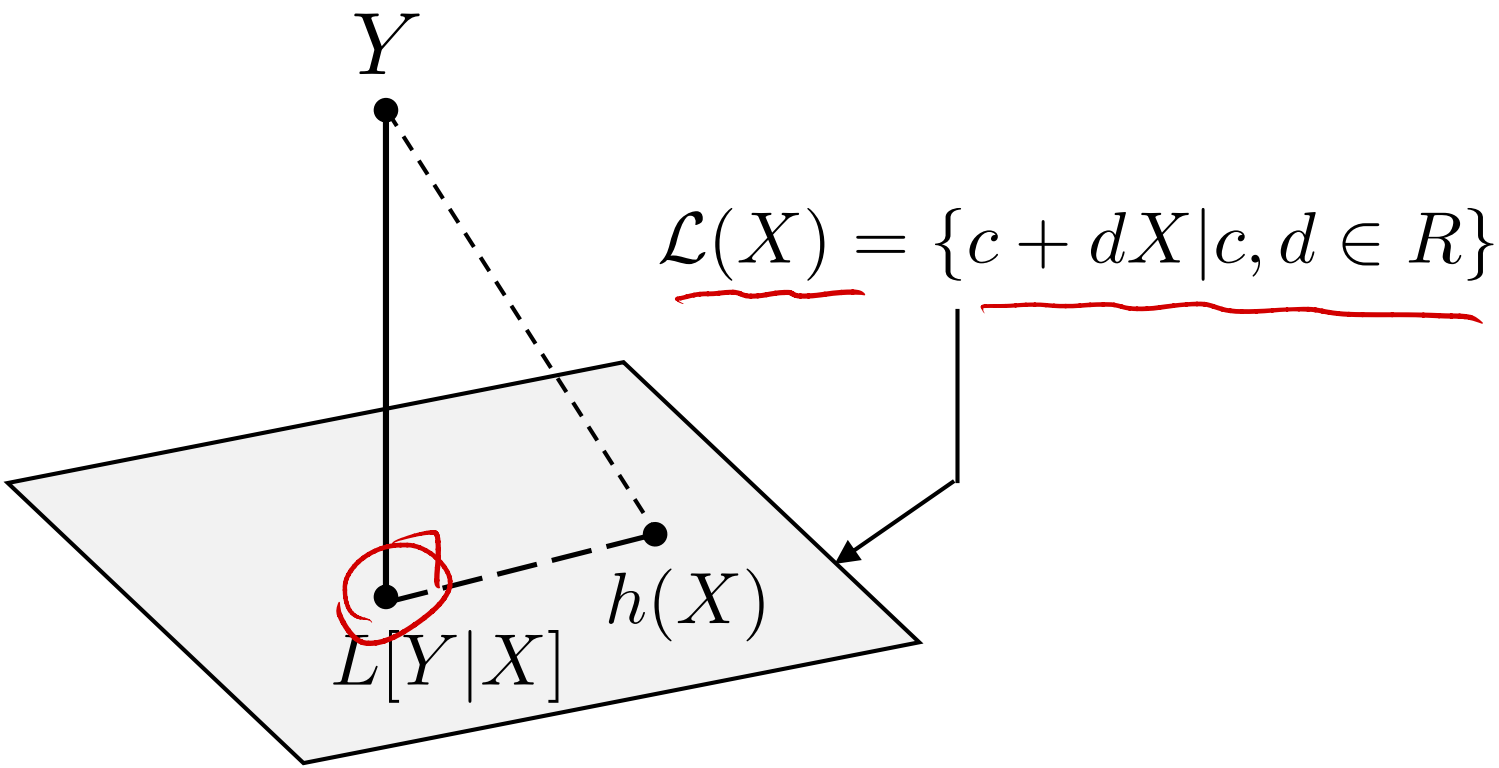
$\min_{a,b} \; E[(Y-a-bX)^2]$

Theorem

The Linear Least Square Estimate (LLSE) of Y given X, denoted by $L[Y|X]$, is the linear function $a+bX$ that minimizes $E[(Y-a-bX)^2]$. In fact,

$$L[Y|X] = E(Y) + \frac{Cov(X,Y)}{Var(X)}(X - E(X))$$

LLSE: Projection Perspective



Minimum Mean Square Error Estimator

LLSE $\min_{\hat{Y}} E[(Y-\hat{Y})^2]$, $\hat{Y}=a+bX$

$\Rightarrow \hat{Y}^* = L[Y|X]$

MMSE $\min_{\hat{Y}} E[(Y-\hat{Y})^2]$, $\hat{Y}=g(X)$

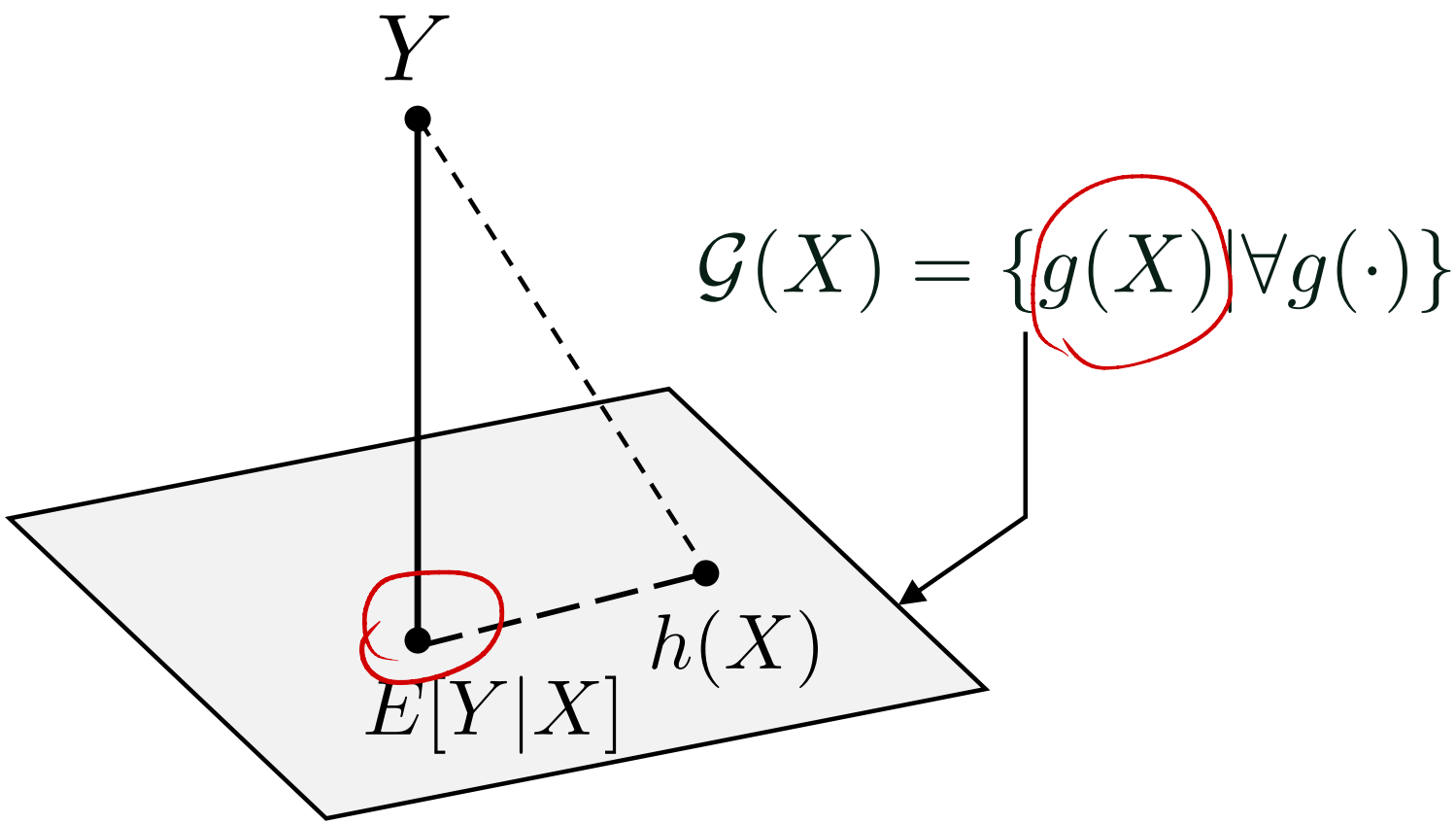
$\Rightarrow \hat{Y}^* = E[Y|X]$

Theorem

The MMSE of Y given X is given by

$$g(X) = E[Y|X]$$

MMSE: Projection Perspective



Projection Interpretation

$E[X^2] < \infty$ ① inner product

$\cos\theta = \frac{\langle X, Y\rangle}{\|X\| \cdot \|Y\|}$ $\langle X, Y\rangle = E[X \cdot Y]$

$dist(X, Y) = \sqrt{\langle X-Y, X-Y\rangle}$ $\|X\| = \sqrt{\langle X, X\rangle}$

$= \sqrt{E(X-Y)^2}$ $= \sqrt{E[X^2]}$

$X \perp Y$

Theorem ② X and Y are orthogonal if $\langle X, Y\rangle = 0$ $\Leftrightarrow E[X \cdot Y] = 0$

For any function h, the r.v. $Y - E(Y|X)$ is uncorrelated with $h(X)$. Equivalently,

$$E((Y - E(Y|X))h(X)) = 0.$$

$(Y - E[Y|X]) \perp h(X)$

(This is equivalent since $E(Y - E(Y|X)) = 0$, by linearity and Adam's law.)

$= E(Y) - E[E(Y|X)] = E(Y) - E(Y) = 0$

③ $Cov(X, Y) = E(X \cdot Y) - E(X) \cdot E(Y)$

If $E(X) = 0$ or $E(Y) = 0$ or both, $Cov(X, Y) = E(X \cdot Y)$

$\Rightarrow$ Uncorrelated $\Leftrightarrow$ Orthogonal

Proof

$$\mathbb{E}[(Y - \mathbb{E}[Y|X]) \cdot h(X)]$$

$$= \mathbb{E}[Y h(X) - \mathbb{E}[Y|X] \cdot h(X)]$$

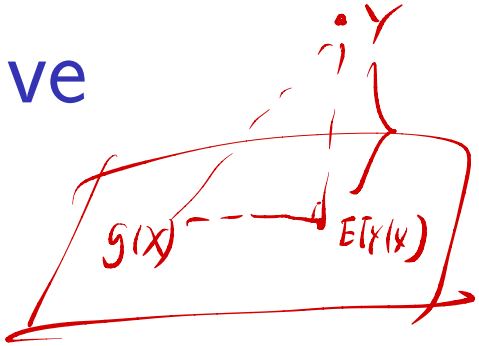
$$= \mathbb{E}[Y h(X)] - \mathbb{E}[\mathbb{E}[Y|X] \cdot h(X)]$$

$$- \mathbb{E}[\mathbb{E}[Y h(X)|X]]$$

$$- \mathbb{E}[Y h(X)]$$

$$= 0$$

Prediction Perspective



$$E[(Y-E[Y|X])^2] \le E[(Y-g(X))^2] \quad \forall\, g(\cdot)$$

- Predict or estimate the future observations or unknown parameters based on data
- $E(Y|X)$ is our **best predictor** of Y based on X.
- Best means it is the function of X with the lowest **mean squared error** (expected squared difference between Y and prediction of Y).

Proof

1°. $\hat{Y}$: estimator of Y based on X. $(\hat{Y}=g(X))$

$$E[(Y-\hat{Y})^2] = E[(Y-g(X))^2] \quad \because Y-g(X) = \underbrace{Y-E[Y|X]}_{A} + \underbrace{E[Y|X]-g(X)}_{B}$$

2°. $E[(Y-g(X))^2] = E[(A+B)^2] = E[A^2] + 2E[AB] + E[B^2]$

$$E[AB] = E[(Y-E[Y|X])\underbrace{(E[Y|X]-g(X))}_{\phi(X)}] = 0$$

$$= E[(Y-E[Y|X])^2] + E[(E[Y|X]-g(X))^2] \geq 0$$

$$\geq E[(Y-E[Y|X])^2]$$

$$\text{If } g(X) \overset{(a.s.)}{=} E[Y|X]$$

Proof

Proof

Orthogonality Property of MMSE

Theorem

(a) For any function $\phi(\cdot)$, one has

$$E[(Y - E[Y|X])\phi(X)] = 0$$

(b) Moreover, if the function $g(X)$ is such that

$$E[(Y - g(X))\phi(X)] = 0, \forall \phi(\cdot).$$

then $g(X) = E(Y|X)$

Proof

$$\mathbb{E}\left[(g(X)-\mathbb{E}[Y|X])^2\right] = \mathbb{E}\left[(g(X)-\mathbb{E}[Y|X])(g(X)-\mathbb{E}[Y|X])\right]$$

$$g(X)-Y+Y-\mathbb{E}[Y|X]$$

$$= \mathbb{E}\left[(g(X)-\mathbb{E}[Y|X])(g(X)-Y)\right] \quad \phi(X) \quad 0 \quad (*2)$$

$$+ \mathbb{E}\left[(g(X)-\mathbb{E}[Y|X])(Y-\mathbb{E}[Y|X])\right]$$

$$\phi(X) \quad 0 \quad (*1)$$

$$= 0$$

$$\Rightarrow \quad g(X) = \mathbb{E}[Y|X]$$

Proof

MMSE for Jointly Gaussian(Normal) Random Variables

X, Y joint $\swarrow \Leftrightarrow a_1X + a_2Y \sim$ Normal

1^o. $Y - L[Y|X] \perp X$.

2^o. $E[Y - L[Y|X]] = 0$

3^o. Orthogonal $\Leftrightarrow$ uncorrelated.

Theorem

Let X, Y be jointly Gaussian random variables. Then

$$E[Y|X] = L[Y|X] = E(Y) + \frac{Cov(X,Y)}{Var(X)}(X - E(X)).$$

4^o. $Y - L[Y|X]$ and X are joint Normal $\surd$

$a_3(Y - L[Y|X]) + a_4X = a_3(Y - a_5X - a_6) + a_4X$

$= a_3Y + (a_4 - a_3a_5)X + a_3a_6$

Proof

5°. $Y - L[Y|X]$ and X are independent.

$\forall \phi(\cdot)$. $Y - L[Y|X]$ and $\phi(X)$ are independent.

6°. $Y - L[Y|X]$ and $\phi(X)$ are uncorrelated.

$E[Y - L[Y|X]] = 0$

7°. $Y - L[Y|X] \perp \phi(X)$, $\forall \phi(\cdot)$

↓ projector.

8°. $L[Y|X] = \text{MMSE}$

Proof

Example: Biased Coin Problem

$X \sim \text{Beta}(a,b)$, $E(X) = \frac{a}{a+b}$.

1°. MMSE. $E[\Theta|X]$: $E[\Theta|X=k]$; $\Theta \sim \text{Unif}(0,1) = \text{Beta}(1,1)$

# of heads $X|\Theta=\theta \sim \text{Bin}(n,\theta)$.

By Beta-Binomial conjugacy, $\Theta|X=k \sim \text{Beta}(1+k, 1+n-k)$

We wish to estimate the probability of landing heads, denoted by θ, of a biased coin. We model θ as the value of a random variable Θ with a known prior PDF $f_\Theta \sim \mathit{Unif}(0,1)$. We consider n independent tosses and let X be the number of heads observed. Find the MMSE $E(\Theta|X)$ and LLSE $L(\Theta|X)$.

$$E[\Theta|X=k] = \frac{k+1}{1+k+1+n-k} = \frac{k+1}{n+2}$$

$$\Rightarrow E[\Theta|X] = \frac{X+1}{n+2}.$$

Solution

2°. LLSE $L[\theta|X] = E[\theta] + \frac{Cov(\theta, X)}{Var(X)}(X - E(X))$

$\theta \sim Unif(0,1) \Rightarrow E[\theta] = \frac{1}{2}$, $Var(\theta) = \frac{1}{12}$, $E(\theta^2) = \frac{1}{3}$.

$X|\theta = \theta \sim Bin(n, \theta) \Rightarrow E[X|\theta = \theta] = n\theta \Rightarrow E[X|\theta] = n\theta$.

$Var[X|\theta = \theta] = n\theta(1-\theta) \Rightarrow Var[X|\theta] = n\theta(1-\theta)$

$\Rightarrow E[X] \overset{Adam}{=} E[E[X|\theta]] = E[n\theta] = nE[\theta] = \frac{1}{2}n$

$Var(X) \overset{EVe}{=} E[Var(X|\theta)] + Var[E[X|\theta]]$

$= E[n\theta(1-\theta)] + Var[n\theta]$

$= n[E(\theta) - E[\theta^2]] + n^2 Var(\theta)$

$= n[\frac{1}{2} - \frac{1}{3}] + n^2 \cdot \frac{1}{12}$

$= \frac{n}{12}(n+2)$

Solution

$$\mathrm{Cov}(X,\Theta) = E[\Theta X] - E[\Theta]\cdot E[X]$$

$$= E[E[\Theta X|\Theta]] - \frac{1}{2}\cdot\frac{n}{2}$$

$$= E[\Theta E[X|\Theta]] - \frac{n}{4}$$

$$= E[\Theta\cdot n\Theta] - \frac{n}{4}$$

$$= nE[\Theta^2] - \frac{n}{4}$$

$$= n\cdot\frac{1}{3} - \frac{n}{4} = \frac{1}{12}n$$

$\Rightarrow$ LLSE

$$L[\Theta|X] = E[\Theta] + \frac{\mathrm{Cov}(\Theta,X)}{\mathrm{Var}(X)}(X-E(X))$$

$$= \frac{1}{2} + \frac{\frac{1}{12}n}{\frac{n}{12}(n+2)}\left(X-\frac{n}{2}\right)$$

$$= \frac{X+1}{n+2} = E[\Theta|X]$$

Solution

Remark: Statistical Learning Perspective

$E[Y|X]$ $E[Y|X=x]$

- In general, MMSE is a highly nonlinear function. Regression function.
- Adoption of various approximation methods leads to various learning methods
  - Linear regression
  - Logistic regression
  - Polynomial regression
  - Regression with Spline functions
  - Neural network

Outline

1. Stochastic Laws in Random Worlds
2. Statistical Inference
3. Conjugate Prior: A Weapon of Bayesian
4. Conditional Expectation
5. Monte Carlo Method: Another Weapon of Bayesian
6. Sampling: Random Variable Generation
7. Sampling: Random Vector Generation
8. Monte Carlo Integration
9. Performance Analysis of Monte Carlo Integration
10. References

Motivation I

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

- Simulations using Monte Carlo Methods
- 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
- 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...

History

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

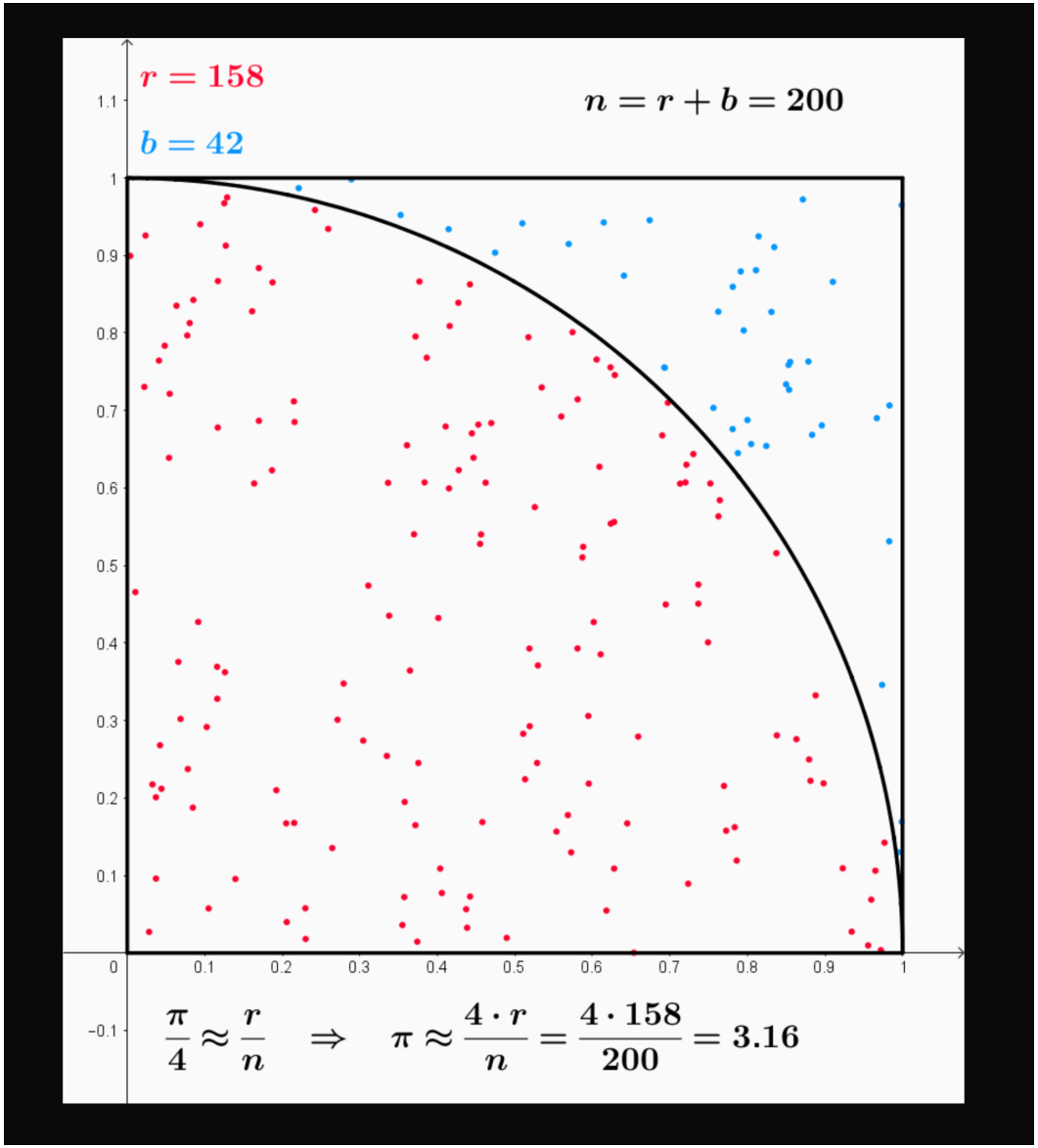
$$c = E_\pi(h(x)) = \int f(x)h(x)dx.$$

- Optimizing a target function to find its maxima or minima

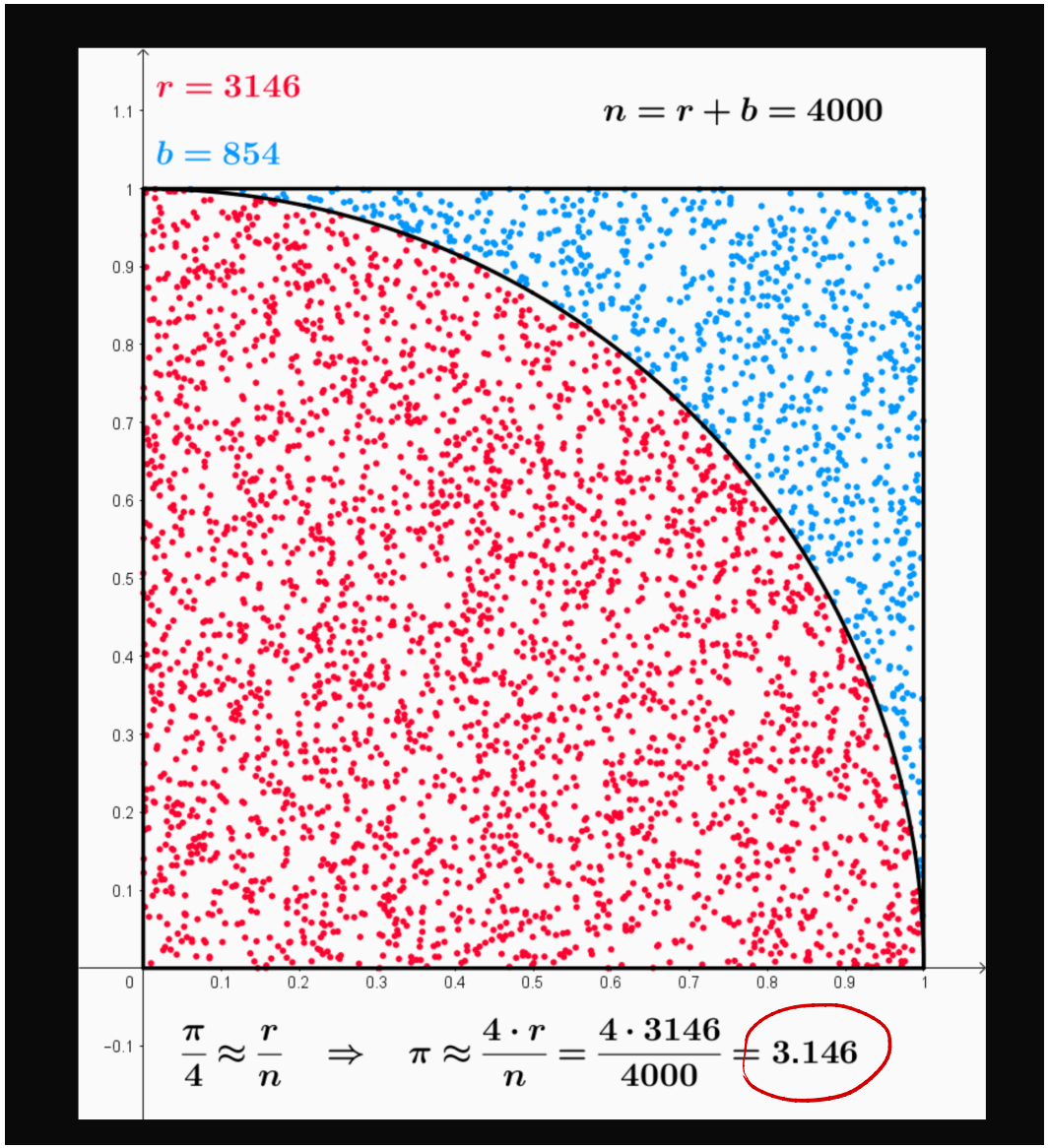
Two Main Categories

- Sequential Monte Carlo(Our Focus in this Lecture): preserves & propagates a population of examples by sequential sampling and importance reweighing, often in a low dimensional state space.
- Markov Chain Monte Carlo(Next Lecture): simulate a Markov chain to explore the state space with a stationary probability designed to converge to a given target probability

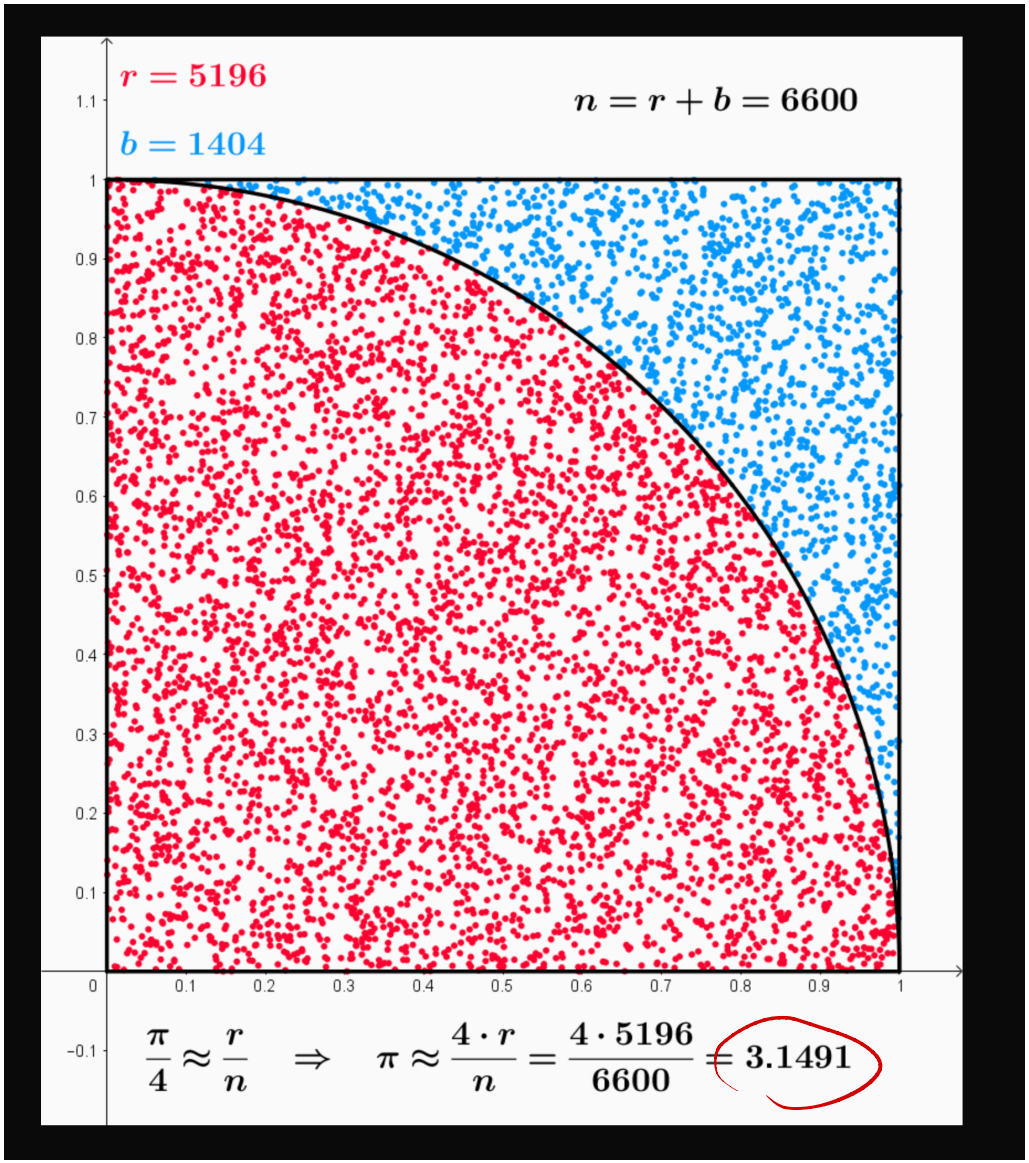
Classical Example: Estimation of π



Classical Example: Estimation of π



Classical Example: Estimation of π

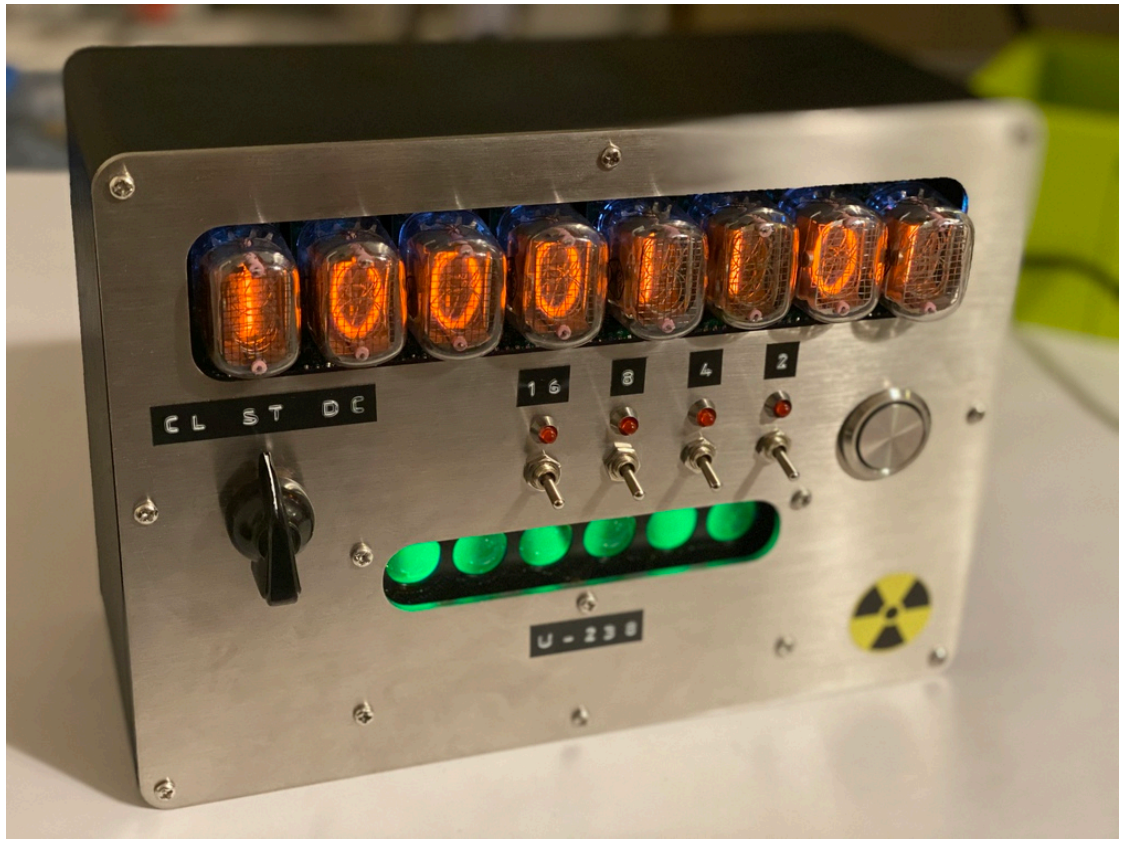


Outline

1. Stochastic Laws in Random Worlds
2. Statistical Inference
3. Conjugate Prior: A Weapon of Bayesian
4. Conditional Expectation
5. Monte Carlo Method: Another Weapon of Bayesian
6. Sampling: Random Variable Generation
7. Sampling: Random Vector Generation
8. Monte Carlo Integration
9. Performance Analysis of Monte Carlo Integration
10. References

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

$0 \le F(x) \le 1$

$P(F^{-1}(U) \le x) = P(U \le F(x)) = F(x)$

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.

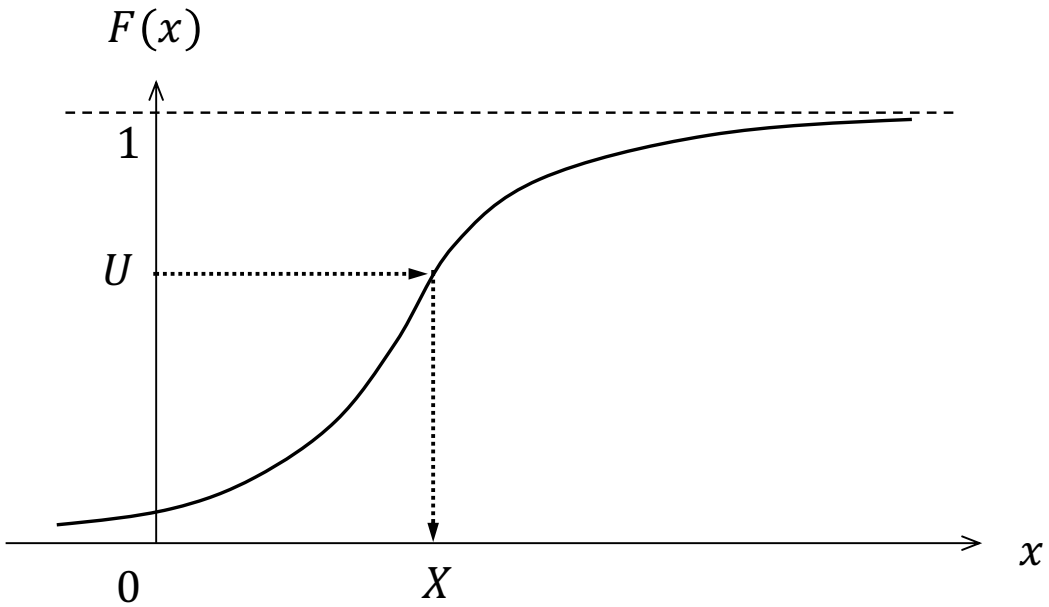
1. *Let $U \sim \mathrm{Unif}(0,1)$ and $X = F^{-1}(U)$. Then X is an r.v. with CDF F.*
2. *Let X be an r.v. with CDF F. Then $F(X) \sim \mathrm{Unif}(0,1)$.*

Algorithm Inverse-Transform Method: PDF Case

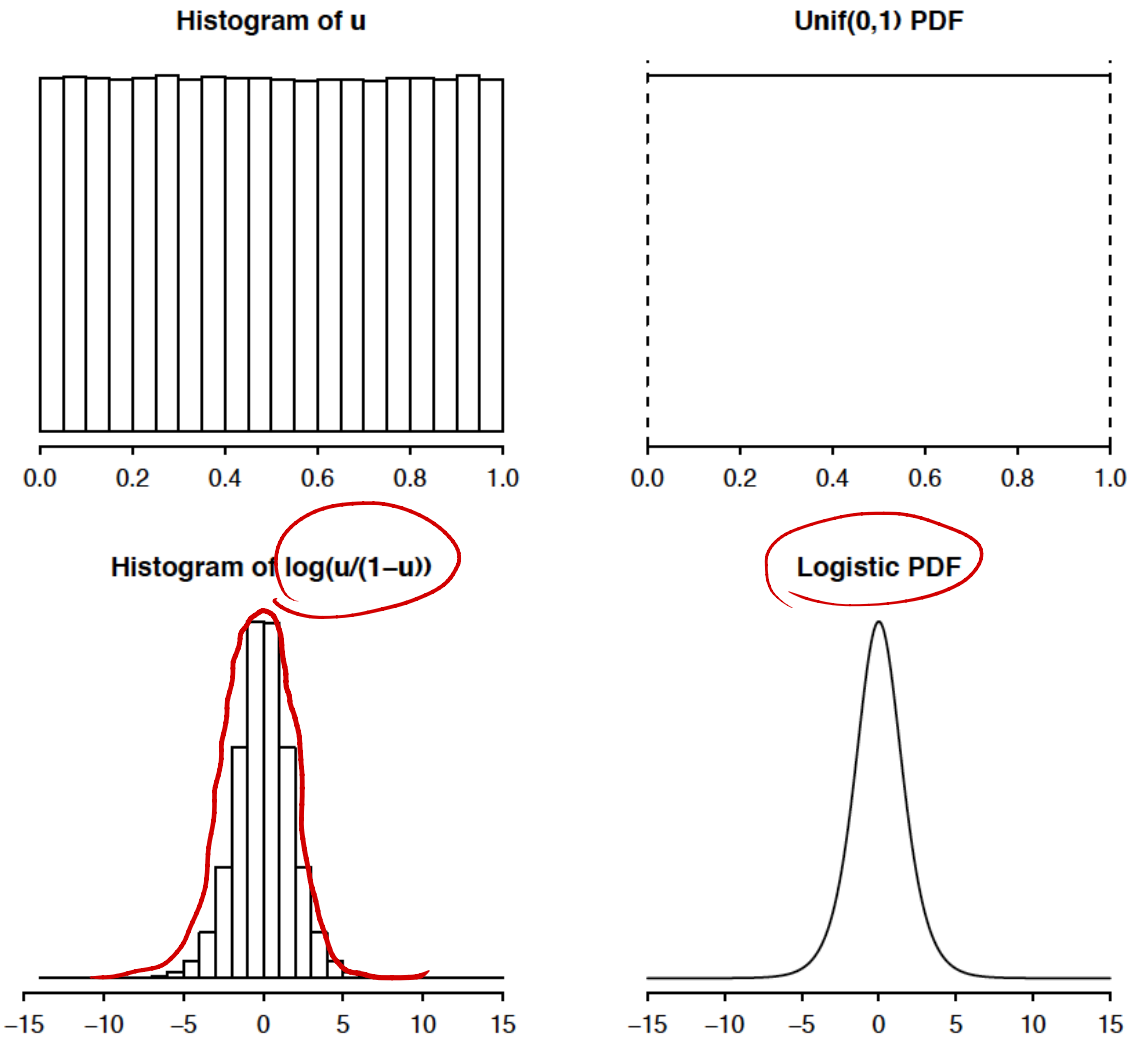
input: Cumulative distribution function F.

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

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

$T \sim \mathrm{Expo}(1)$

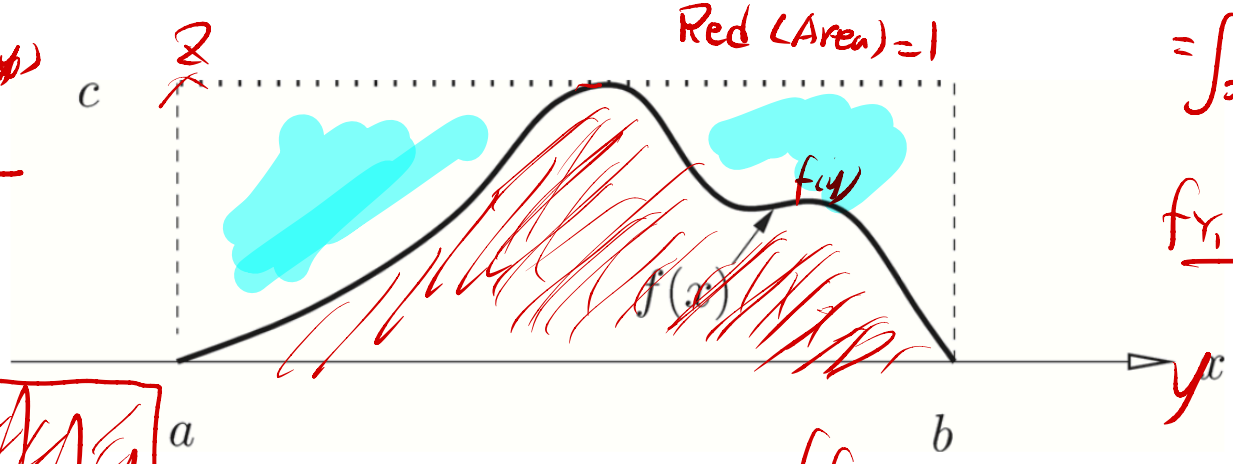
$F_T(t) = 1 - e^{-t}, \quad t > 0$

$F_T^{-1}(u) = -\ln(1-u)$ $\quad 0 < u < 1$

$U \sim \mathrm{unif}(0,1)$

$1-U \sim \mathrm{unif}(0,1)$

Acceptance-Rejection Method



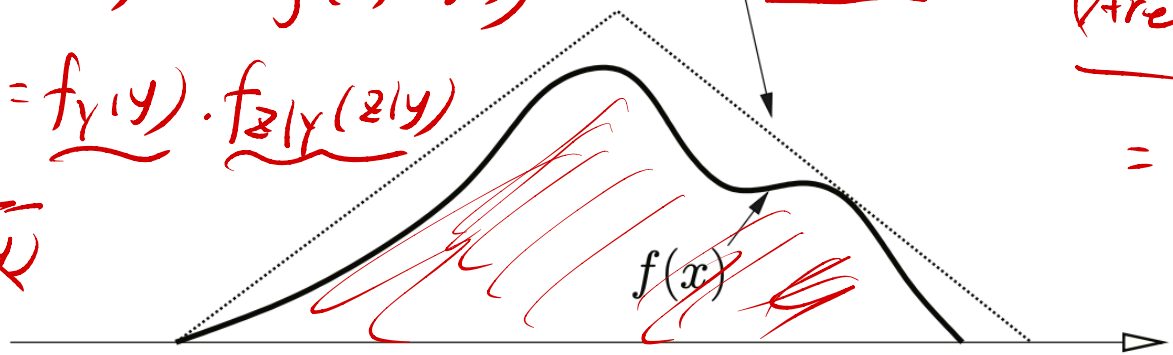
Algorithm Acceptance-Rejection Algorithm

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.

Acceptance-Rejection Method



Algorithm Acceptance-Rejection Algorithm

Step 1: Generate $Y \sim g$.

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

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

Acceptance-Rejection Method

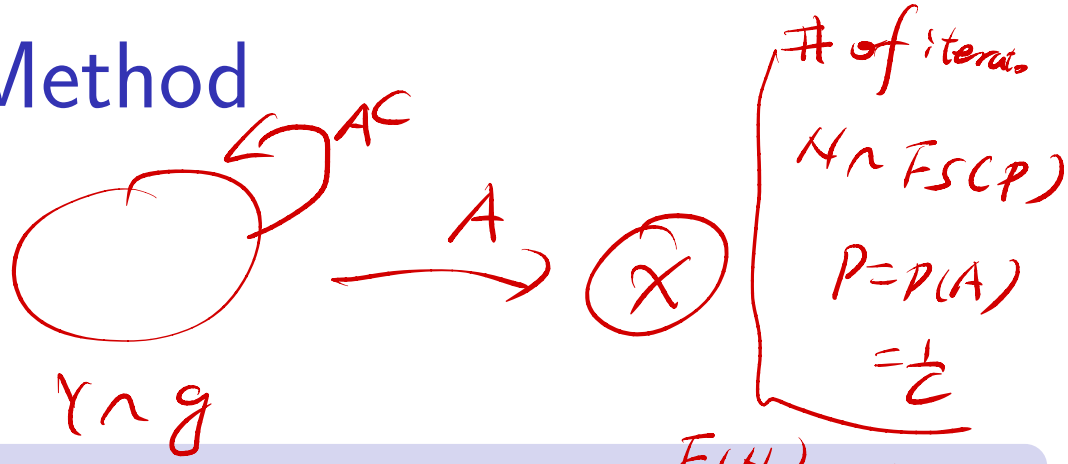
- Suppose one can generate samples (relatively easily) from PDF g
- How can random samples be simulated from PDF f?

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

Acceptance-Rejection Method



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) $c \geq 1$

Proof

(i) event $A = "U \le \frac{f(Y)}{c \cdot g(Y)}"$. $U \sim \mathrm{unif}(0,1)$

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

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

$= P\left(U \le \frac{f(y)}{c \cdot g(y)}\right) = \frac{f(y)}{c \cdot g(y)}$ (@ 0,1)

$c \ge \sup_y \frac{f(y)}{g(y)}$

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

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

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

$f_X(y) = f(y)$

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

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

  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: $unif(0,1)$, $g(x) = 1$, $0<x<1$.

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

$$\Rightarrow c \geq \frac{f(y^*)}{g(y^*)} = \frac{135}{64} > 1, \Rightarrow \text{Choose } c = \frac{135}{64}.$$

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

Step 1: Generate $Y \sim unif(0,1)$

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

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

otherwise reject Y, Go back to Step 1.

Solution

Outline

1. Stochastic Laws in Random Worlds
2. Statistical Inference
3. Conjugate Prior: A Weapon of Bayesian
4. Conditional Expectation
5. Monte Carlo Method: Another Weapon of Bayesian
6. Sampling: Random Variable Generation
7. Sampling: Random Vector Generation
8. Monte Carlo Integration
9. Performance Analysis of Monte Carlo Integration
10. References

Change of Variables

Theorem

Let $\mathbf{X} = (X_1, ..., X_n)$ *be a continuous random vector with joint PDF* $f_{\mathbf{X}}(x)$*, and let* $\mathbf{Y} = g(\mathbf{X})$ *where* g *is an invertible function from* $\mathbb{R}^n$ *to* $\mathbb{R}^n$*. Let* $y = g(\mathbf{x})$ *and suppose that all the partial derivatives* $\frac{\partial x_i}{\partial y_j}$ *exists and are continuous, so we can form the* ***Jacobian matrix***

$$\frac{\partial \mathbf{x}}{\partial \mathbf{y}} = \begin{pmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_1}{\partial y_2} & \cdots & \frac{\partial x_1}{\partial y_n} \\ \vdots & \vdots & & \vdots \\ \frac{\partial x_n}{\partial y_1} & \frac{\partial x_n}{\partial y_2} & \cdots & \frac{\partial x_n}{\partial y_n} \end{pmatrix}$$

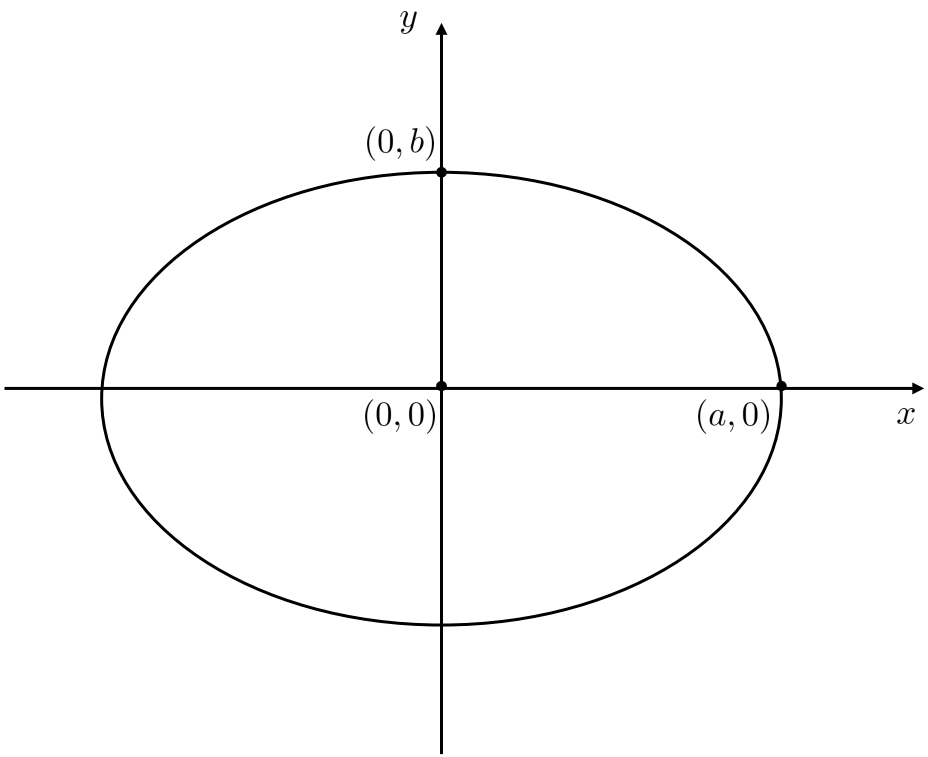
Also assume that the determinant of the Jacobian matrix is never 0. Then the joint PDF of $\mathbf{Y}$ *is*

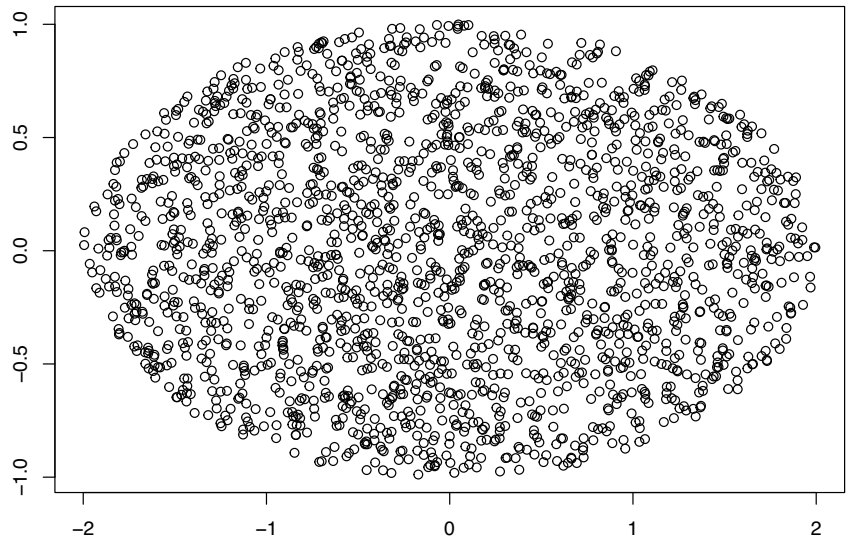
$$f_{\mathbf{Y}}(y) = f_{\mathbf{X}}(x) \left| \frac{\partial \mathbf{x}}{\partial \mathbf{y}} \right|$$

Example: Generate Uniform Distribution over An Ellipse

- Ellipse:

$$E_2(a,b) = \left\{ (x,y) \in \mathbb{R}^2 : \left(\frac{x}{a}\right)^2 + \left(\frac{y}{b}\right)^2 \leq 1 \right\}.$$





Solution

① object PDF $f_{X,Y}(x,y) = \frac{1}{\pi ab}$, $\forall x, y \in E_2(a,b)$

$x = \rho a \cos\theta$

$y = \rho b \sin\theta$.

$\rho \in [0,1]$, $\theta \in [0, 2\pi]$

② Jacobi Matrix $\begin{bmatrix} \frac{\partial x}{\partial \rho} & \frac{\partial x}{\partial \theta} \\ \frac{\partial y}{\partial \rho} & \frac{\partial y}{\partial \theta} \end{bmatrix} = \begin{bmatrix} a\cos\theta & -\rho a \sin\theta \\ b\sin\theta & \rho b \cos\theta \end{bmatrix}$

$J = \det(\quad) = \rho ab$.

$\Rightarrow f_{R,\Theta}(\rho,\theta) = f_{X,Y}(x,y) \cdot |J| = \frac{1}{\pi ab} \cdot \rho \cdot a \cdot b = \frac{\rho}{\pi}$.

$\Rightarrow f_R(\rho) = \int_0^{2\pi} \frac{\rho}{\pi} d\theta = 2\rho$, $0 \le \rho \le 1$ $\Rightarrow$ CDF $F_R(\rho) = \rho^2$, $0 \le \rho \le 1$.

$\Rightarrow F_R^{-1}(z) = \sqrt{z}$, $0 \le z \le 1$

$f_\Theta(\theta) = \int_0^1 \frac{\rho}{\pi} d\rho = \frac{1}{2\pi}$, $0 \le \theta \le 2\pi$. $\Rightarrow \Theta \sim unif(0, 2\pi) = 2\pi\, unif(0,1)$

$f_{R,\Theta}(\rho,\theta) = f_R(\theta) f_\Theta(\theta)$, R and Θ are independent.

Solution

③

$X = \rho a \cos\theta, \quad Y = \rho b \sin\theta$

U_1, U_2 independent $\mathrm{Unif}(0,1)$

$X \leftarrow a\sqrt{U_1}\cos(2\pi U_2)$

$Y \leftarrow b\sqrt{U_1}\sin(2\pi U_2)$

(X, Y)

Solution

Change of Variables

$A: n \times n$ matrix

$\det(A^T A) = \det^2(A)$

Theorem

Let $\mathbf{X} = (X_1, ..., X_n)$ *be a continuous random vector with joint PDF* $f_{\mathbf{X}}(x)$*, and let* $\mathbf{Y} = g(\mathbf{X})$ *where* g *is a function from* $\mathbb{R}^n$ *to* $\mathbb{R}^m$*. Let* $y = g(\mathbf{x})$ *and we have the* ***Jacobian matrix*** $\frac{\partial \mathbf{x}}{\partial \mathbf{y}}$*. The corresponding Gram matrix is*

$$\mathbf{G} = \left(\frac{\partial \mathbf{x}}{\partial \mathbf{y}}\right)^T \frac{\partial \mathbf{x}}{\partial \mathbf{y}}.$$

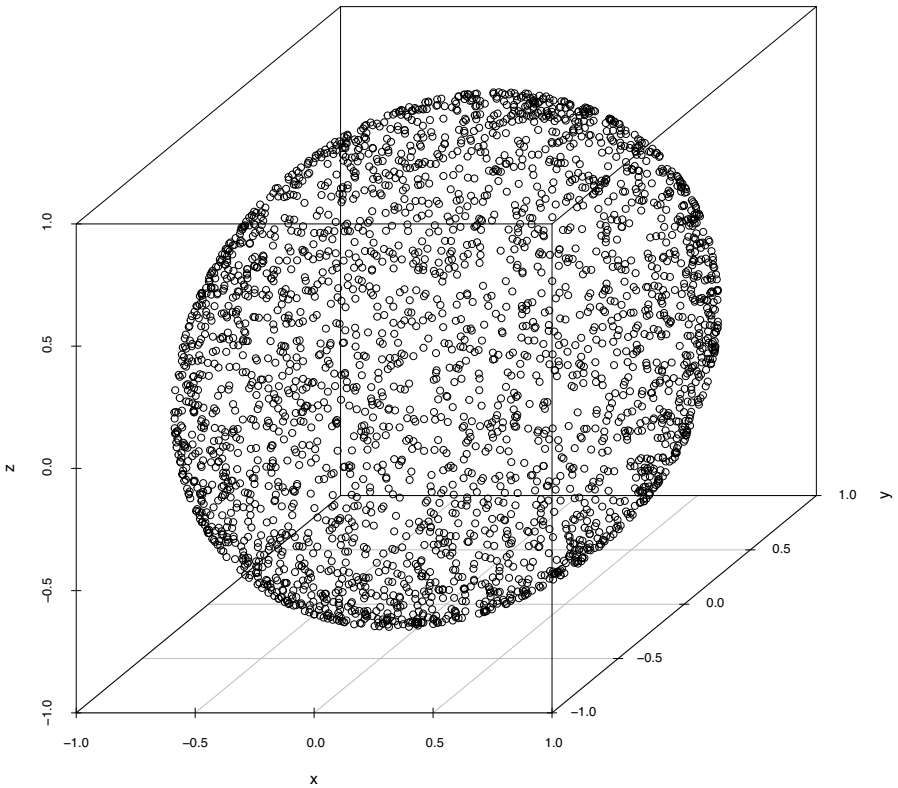
Then the joint PDF of $\mathbf{Y}$ *is*

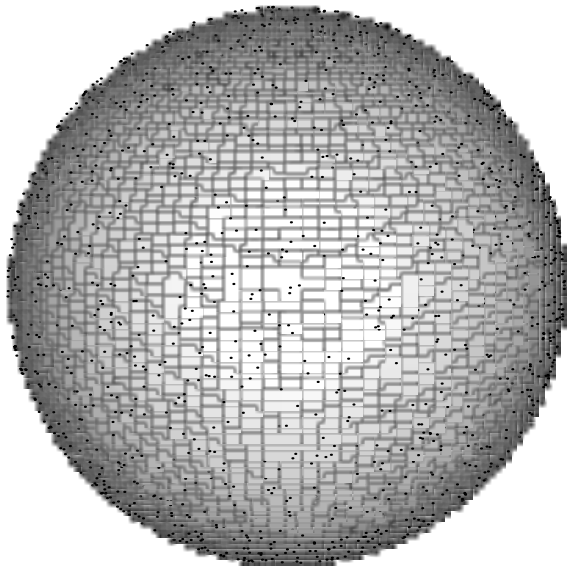
$$f_{\mathbf{Y}}(y) = f_{\mathbf{X}}(x)\sqrt{det(\boldsymbol{G})}$$

Example: Generate Uniform Distribution over A Sphere Surface

- Sphere Surface:

$$S_2(r) = \left\{(x, y, z) \in \mathbb{R}^3 : x^2 + y^2 + z^2 = r^2\right\}.$$





Solution

① $f_{X,Y,Z}(x,y,z) = \frac{1}{4\pi r^2}$, $\forall (x,y,z) \in S_2(r)$

$x = r\sin\theta\cos\phi$, $\theta \in [0,\pi]$

$y = r\sin\theta\sin\phi$, $\phi \in [0,2\pi]$.

$z = r\cos\theta$

$(x,y,z) \longrightarrow (\theta,\phi)$.

Jacobi matrix $M = \begin{bmatrix} \frac{\partial x}{\partial\theta} & \frac{\partial x}{\partial\phi} \\ \frac{\partial y}{\partial\theta} & \frac{\partial y}{\partial\phi} \\ \frac{\partial z}{\partial\theta} & \frac{\partial z}{\partial\phi} \end{bmatrix} = \begin{bmatrix} r\cos\theta\cos\phi & -r\sin\theta\sin\phi \\ r\cos\theta\sin\phi & r\sin\theta\cos\phi \\ -r\sin\theta & 0 \end{bmatrix}$

Gram matrix $G = M^T M = \begin{bmatrix} r^2 & 0 \\ 0 & r^2\sin^2\theta \end{bmatrix} \Rightarrow \det(G) = r^4\sin^2\theta$.

$\Rightarrow f_{\theta,\phi}(\theta,\phi) = f_{X,Y,Z}(x,y,z)\sqrt{\det(G)} = \frac{1}{4\pi r^2}\cdot r^2\sin\theta = \frac{1}{4\pi}\sin\theta$.

Solution

② $f_\Theta(\theta) = \int_0^{2\pi} f_{\Theta,\Phi}(\theta,\phi)\,d\phi = \frac{1}{2}\sin\theta$, $0 \le \theta \le \pi$.

$F_\Theta(\theta) = \int_0^\theta f_\Theta(s)\,ds = \frac{1-\cos\theta}{2}$, $0 \le \theta \le \pi$.

$f_\Phi(\phi) = \int_0^\pi f_{\Theta,\Phi}(\theta,\phi)\,d\theta = \frac{1}{2\pi}$, $0 \le \phi \le 2\pi$.

$\Rightarrow f_{\Theta,\Phi}(\theta,\phi) = f_\Theta(\theta)\cdot f_\Phi(\phi)$. $\Rightarrow \Theta$ and Φ are independent.

$\phi \sim \text{unif}(0,2\pi) = 2\pi\,\text{unif}(0,1)$

$F_\Theta^{-1}(s) = \arccos(1-2s)$, $0 \le s \le 1$,

If $\Theta = \arccos(1-2S) \Rightarrow \cos\theta = 1-2S \Rightarrow \sin^2\theta = 1-\cos^2\theta = 4S(1-S)$.

$\sin\theta \ge 0 \Rightarrow \sin\theta = 2\sqrt{S(1-S)}$

Solution

③ $X = r\sin\theta\cos\phi = r \cdot 2\sqrt{s(1-s)}\cos(2\pi s')$

$Y = r\sin\theta\sin\phi = r \cdot 2\sqrt{s(1-s)}\sin(2\pi s')$

$Z = r\cos\theta = r(1-2s)$

---

④ U_1, U_2 independently $\mathrm{Unif}(0,1)$

$X \leftarrow r \cdot 2\sqrt{U_1(1-U_1)} \cdot \cos(2\pi U_2)$

$Y \leftarrow r \cdot 2\sqrt{U_1(1-U_1)} \cdot \sin(2\pi U_2)$

$Z \leftarrow r(1-2U_1)$

Outline

1. Stochastic Laws in Random Worlds
2. Statistical Inference
3. Conjugate Prior: A Weapon of Bayesian
4. Conditional Expectation
5. Monte Carlo Method: Another Weapon of Bayesian
6. Sampling: Random Variable Generation
7. Sampling: Random Vector Generation
8. Monte Carlo Integration
9. Performance Analysis of Monte Carlo Integration
10. References

Monte Carlo Integration

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

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

- Now we have integration

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

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

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

- Monte Carlo Integration:

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

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 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

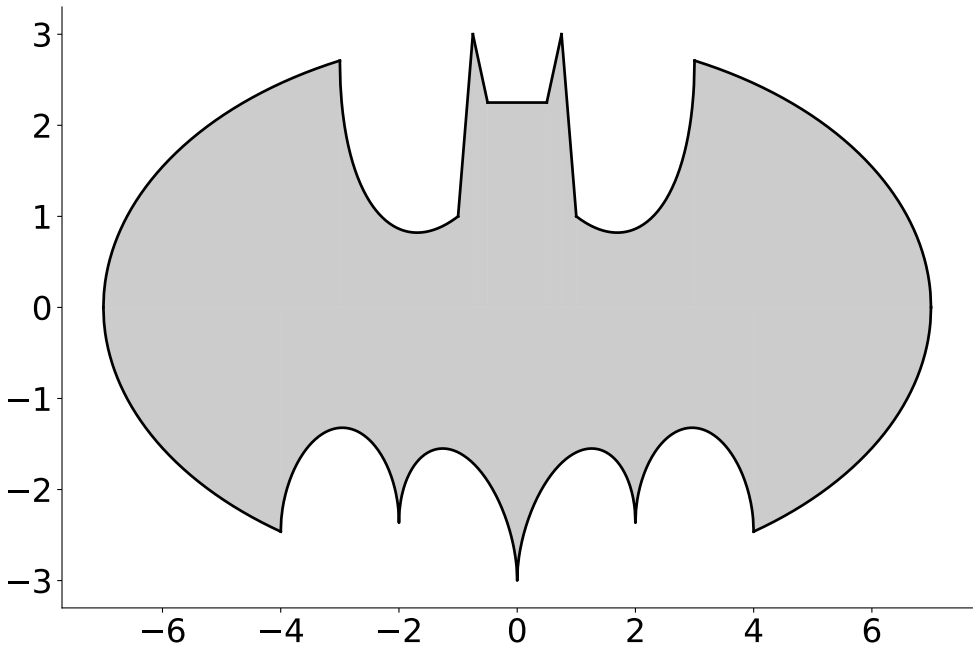
$$g(x)=\sqrt{x+\sqrt{x+\sqrt{x+\sqrt{x}}}}.$$

- $X_1,\ldots,X_n$: 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`



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

$c = P(Y>8) = E[I(Y>8)] \approx \frac{1}{n}\sum_{j=1}^{n} I(Y_j>8)$

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

$h(y) = I(y>8) = \begin{cases} 1 & \text{if } y>8 \\ 0 & \text{otherwise} \end{cases}$

Choose $g \sim N(8,1)$, $Y_1, \dots Y_n \sim g$ $\quad f \sim N(0,1)$

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

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

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

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

Solution

Outline

1. Stochastic Laws in Random Worlds
2. Statistical Inference
3. Conjugate Prior: A Weapon of Bayesian
4. Conditional Expectation
5. Monte Carlo Method: Another Weapon of Bayesian
6. Sampling: Random Variable Generation
7. Sampling: Random Vector Generation
8. Monte Carlo Integration
9. Performance Analysis of Monte Carlo Integration
10. References

Sample Mean: Recall

Definition

Let $X_1, ..., X_n$ 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.

Asymptotic Analysis: Strong Law of Large Numbers (SLLN)

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.

Asymptotic Analysis: 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).

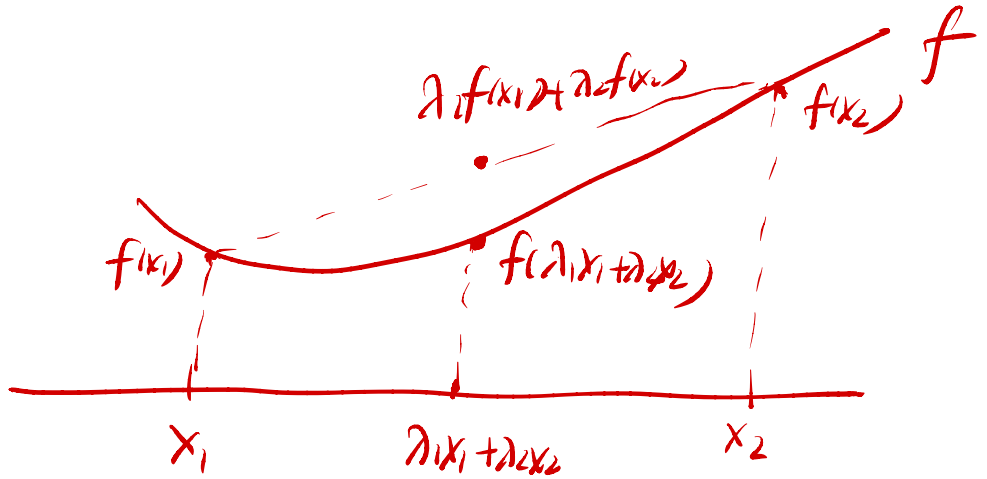
Finite Sample Analysis: Inequalities

Theorem (Cauchy-Schwarz Inequality)

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

1°. $g(y) \geq g(x) + g'(x)(y-x)$

$y \in X$

$x \in E[X]$

$\Rightarrow g(X) \geq g(E(X)) + g'(E(X))(X - E(X))$

$\Rightarrow E[g(X)] \geq g(E(X)) + g'(E(X)) \cdot E[X - EX]$

$= g(E(X))$ (since $E[X - EX] = 0$)

2°. Subgradient $g^0(\cdot)$

$g(y) \geq g(x) + g^0(x)(y-x)$

$y \in X$

$x \in E[X]$

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 $\Rightarrow \mathbb{E}[g(X)] \geq g[\mathbb{E}(X)]$, $g''(\cdot) \geq 0$

Concave $\Rightarrow \mathbb{E}[g(X)] \leq g[\mathbb{E}(X)]$, $g''(\cdot) \leq 0$

1° $g(x) = x^2$, $x \in \mathbb{R}$, Convex $\Rightarrow \mathbb{E}[X^2] \geq (\mathbb{E}(X))^2$ ✓

$\mathrm{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}^2(X) \geq 0$

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

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

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: $H(X) = \sum_{j=1}^{n} p_j \log_2 (1/p_j)$.
- 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.

$$\int p(x) \log_2 p(x) \, dx$$

Proof

① Construct a r.v. Y s.t.

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

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

$\forall p_1, p_2 \dots p_n$, $p_1 + \dots + p_n = 1$ $\Rightarrow \max_{p_1 \dots p_n} H(X) \le \log_2 n$

③ when $X \sim \text{unif}(\frac{1}{n})$, $p_1 = p_2 = \dots = p_n = \frac{1}{n}$, $H(X) = \sum_{j=1}^{n} \frac{1}{n} \log_2 n = \log_2 n$

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

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

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

Construct a r.v. Y s.t.

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

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

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

② $\log_2 x \leq x - 1$

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

$$= -\sum_{j=1}^{n} (r_j - p_j) = -\sum_{j=1}^{n} r_j + \sum_{j=1}^{n} p_j$$

$$= -1 + 1 = 0$$

Markov's Inequality

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

$a \uparrow$ prob $\downarrow$

Concentration Inequality

$\frac{1}{a}$

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

e^{-a}

e^{-a^2}

Theorem

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

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

Proof

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

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

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

② $E[I(Y \geq 1)] \leq E[Y]$

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

$\Downarrow$

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

$\Downarrow$

$P(|X| \geq a)$

$\leq$

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 $a > 0$,

$$P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}. \qquad O\left(\frac{1}{a^2}\right)$$

Proof

Chernoff's Inequality

$\forall t>0$,

$P(X \geq a) = P(tX \geq ta)$

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

$\overset{\text{Markov's inequality}}{\leq} \dfrac{E[e^{tX}]}{e^{ta}} = f(t)$

Theorem

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

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

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

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

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,

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

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

Proof

Hoeffding Lemma

$X \sim N(\mu, \sigma^2)$

$E[e^{tX}] = e^{\mu t + \frac{1}{2}\sigma^2 t^2}$

$\mu = 0 \Rightarrow e^{\frac{1}{2}\sigma^2 t^2}$

Lemma

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

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

Sub-Gaussian

Useful Analysis Tools

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

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

- 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 θ between a and x such that

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

Hoeffding Bound

(1) $Z = \frac{1}{n}\sum_{i=1}^{n} X_i - \mu = \frac{1}{n}\sum_{i=1}^{n} Z_i$

$Z_i = X_i - \mu = X_i - E[X_i] \Rightarrow E[Z_i] = 0, \ i = 1, 2, \ldots n.$

$Z_1, \ldots Z_n$ independent $\qquad E[Z] = 0.$

Theorem

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

$\{|Z| \ge \varepsilon\} \in \{Z \ge \varepsilon\} + \{Z \le -\varepsilon\}$

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

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

(2) $\forall \lambda > 0, \ P(Z \ge \varepsilon) = P(e^{\lambda Z} \ge e^{\lambda \varepsilon})$

$\le \frac{1}{e^{\lambda\varepsilon}} E[e^{\lambda Z}] = \frac{1}{e^{\lambda\varepsilon}} E[e^{\lambda \cdot \frac{1}{n}\sum_{i=1}^{n} Z_i}]$

Proof

$\forall \lambda > 0$; $P(\bar{X} \geq \varepsilon) \leq e^{-\lambda\varepsilon} \prod_{i=1}^{n} E[e^{\lambda \cdot \frac{1}{n} Z_i}]$

$\frac{1}{n} Z_i$: $E(\frac{1}{n} Z_i) = \frac{1}{n} E(Z_i) = 0$

$a \leq X_i \leq b \Rightarrow a - \mu \leq \underbrace{X_i - \mu}_{Z_i} \leq b - \mu$

$\Rightarrow \frac{1}{n}(a-\mu) \leq \frac{1}{n} Z_i \leq \frac{1}{n}(b-\mu)$

By Hoeffding's Lemma. $E[e^{\lambda \cdot \frac{1}{n} Z_i}] \leq e^{\frac{1}{8}\lambda^2 [\frac{1}{n}(b-\mu) - \frac{1}{n}(a-\mu)]^2}$

$= e^{\frac{1}{8}\lambda^2 \cdot \frac{(b-a)^2}{n^2}}$, $\forall i = 1, 2, \ldots, n$

(3) $P(\bar{X} \geq \varepsilon) \leq e^{-\lambda\varepsilon} \cdot \prod_{i=1}^{n} [e^{\frac{1}{8}\lambda^2 \frac{(b-a)^2}{n^2}}]$ $= e^{-\lambda\varepsilon} \cdot e^{\frac{1}{8}\lambda^2 \frac{(b-a)^2}{n}}$

$\leq e^{-\lambda\varepsilon + \frac{1}{8}\lambda^2 \cdot \frac{(b-a)^2}{n}} = e^{g(\lambda)}$ $\forall (\lambda > 0)$

$g(\lambda) = -\lambda\varepsilon + \frac{1}{8}\lambda^2 \frac{(b-a)^2}{n}$ $\Rightarrow g'(\lambda^*) = 0$; $g''(\lambda^*) > 0$.

$\Rightarrow \lambda^* = \frac{4n\varepsilon}{(b-a)^2}$, $\Rightarrow g(\lambda^*) = \frac{-2n\varepsilon^2}{(b-a)^2}$

Proof

$$P(Z \geq \varepsilon) \leq \inf_{\lambda > 0} e^{g(\lambda)} = e^{g(\lambda^*)} = e^{\frac{-2n\varepsilon^2}{(b-a)^2}}.$$

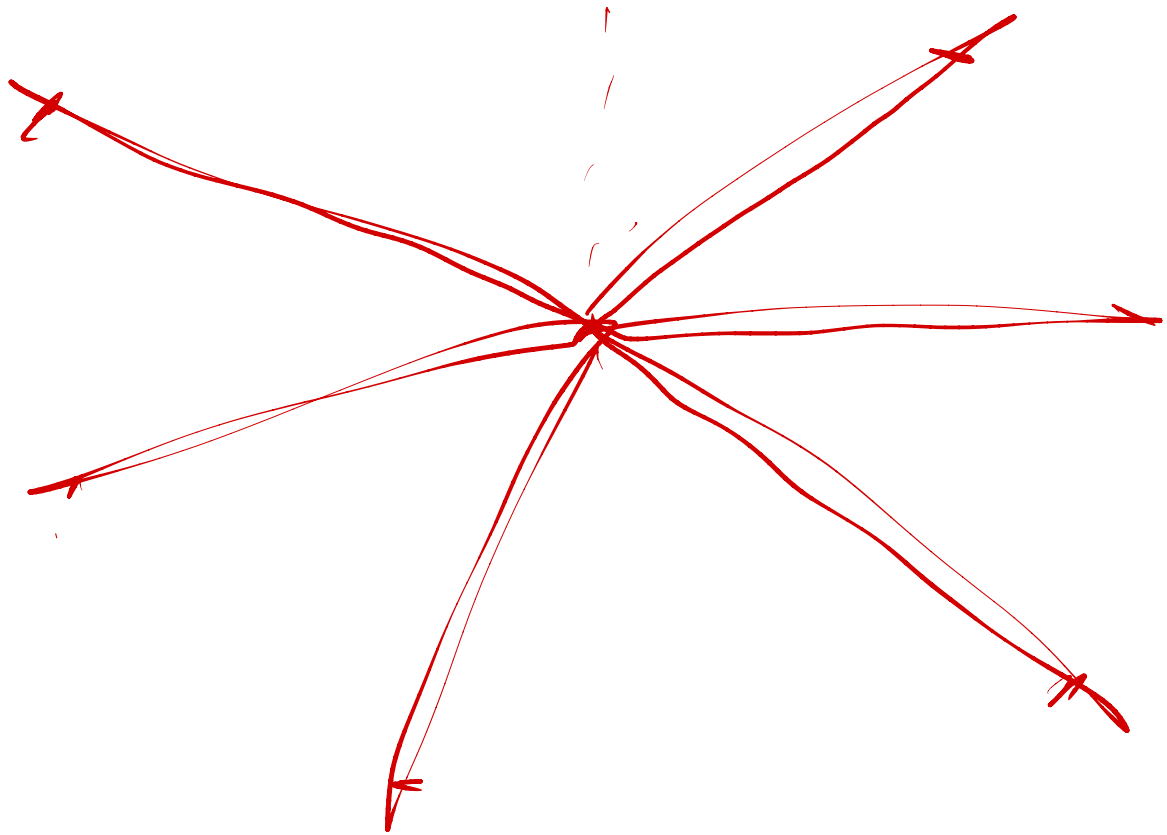
(4) $$P(Z \leq -\varepsilon) \leq e^{\frac{-2n\varepsilon^2}{(b-a)^2}}$$

$$\Rightarrow P\left(\left|\frac{1}{n}\sum_{i=1}^{n} x_i - \mu\right| \geq \varepsilon\right) = P(|Z| \geq \varepsilon)$$

$$= P(Z \geq \varepsilon) + P(Z \leq -\varepsilon) = 2e^{\frac{-2n\varepsilon^2}{(b-a)^2}}$$

$n \to \infty$ $\Rightarrow \frac{1}{n}\sum_{i=1}^{n} x_i \to \mu$

Proof



More General Hoeffding Bound

Theorem

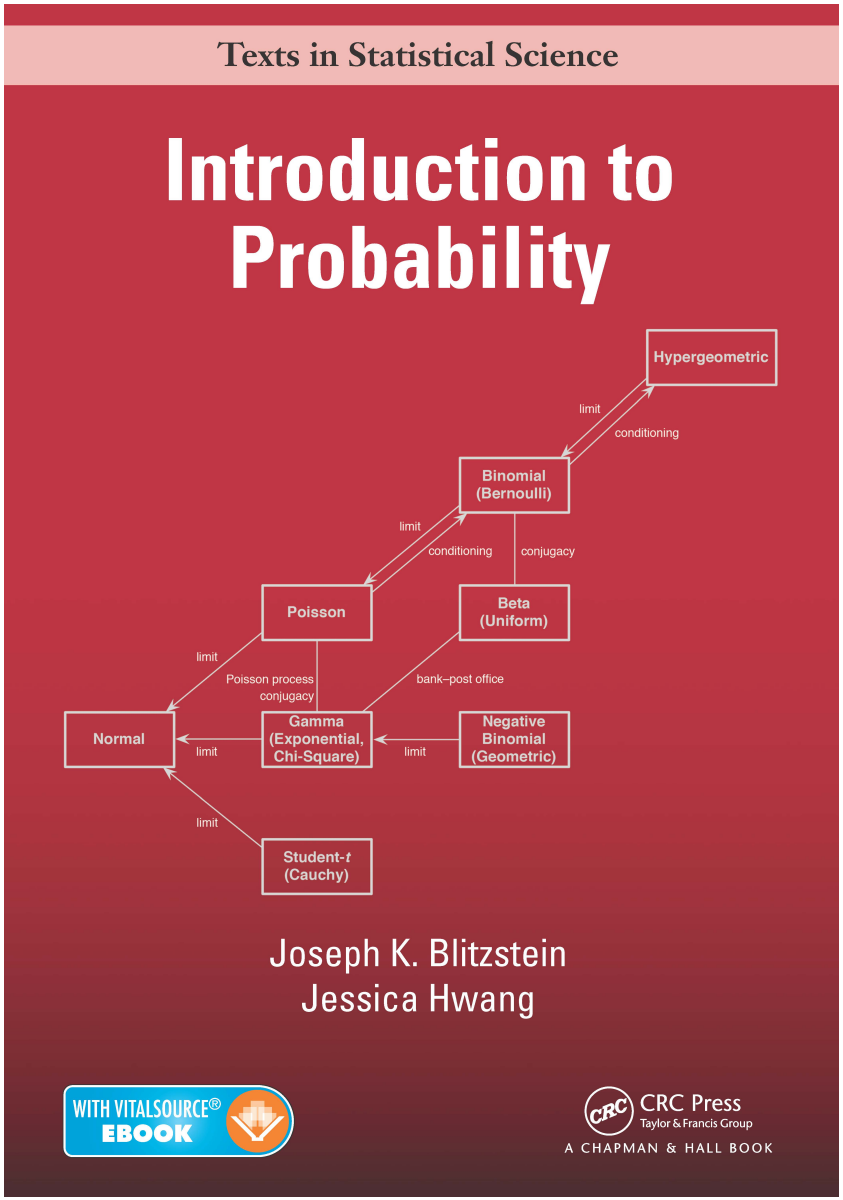
Let the random variables $X_1, X_2, \ldots, X_n$ be independent, with $a_k \leq X_k \leq b_k$ for each k, where a_k, b_k are constants. Let $S_n = \sum_{k=1}^{n} X_k$ and let $\mu = \mathbb{E}(S_n)$. Then for any $t \geq 0$,

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

Outline

1. Stochastic Laws in Random Worlds
2. Statistical Inference
3. Conjugate Prior: A Weapon of Bayesian
4. Conditional Expectation
5. Monte Carlo Method: Another Weapon of Bayesian
6. Sampling: Random Variable Generation
7. Sampling: Random Vector Generation
8. Monte Carlo Integration
9. Performance Analysis of Monte Carlo Integration
10. References

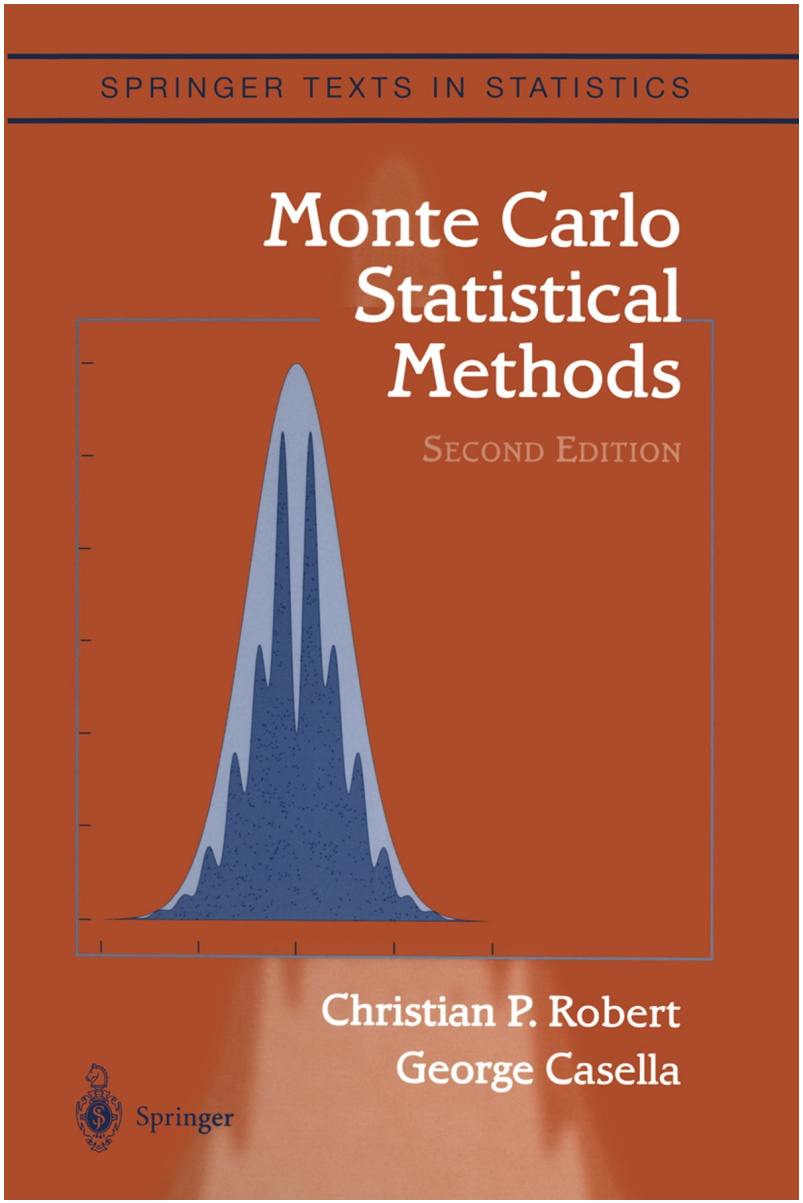
BH



Joseph K. Blitzstein & Jessica Hwang

- Introduction to Probability
- Chapman & Hall/CRC, 2014.
- Chapman & Hall/CRC, 2019.
- Chapters 8 & 9

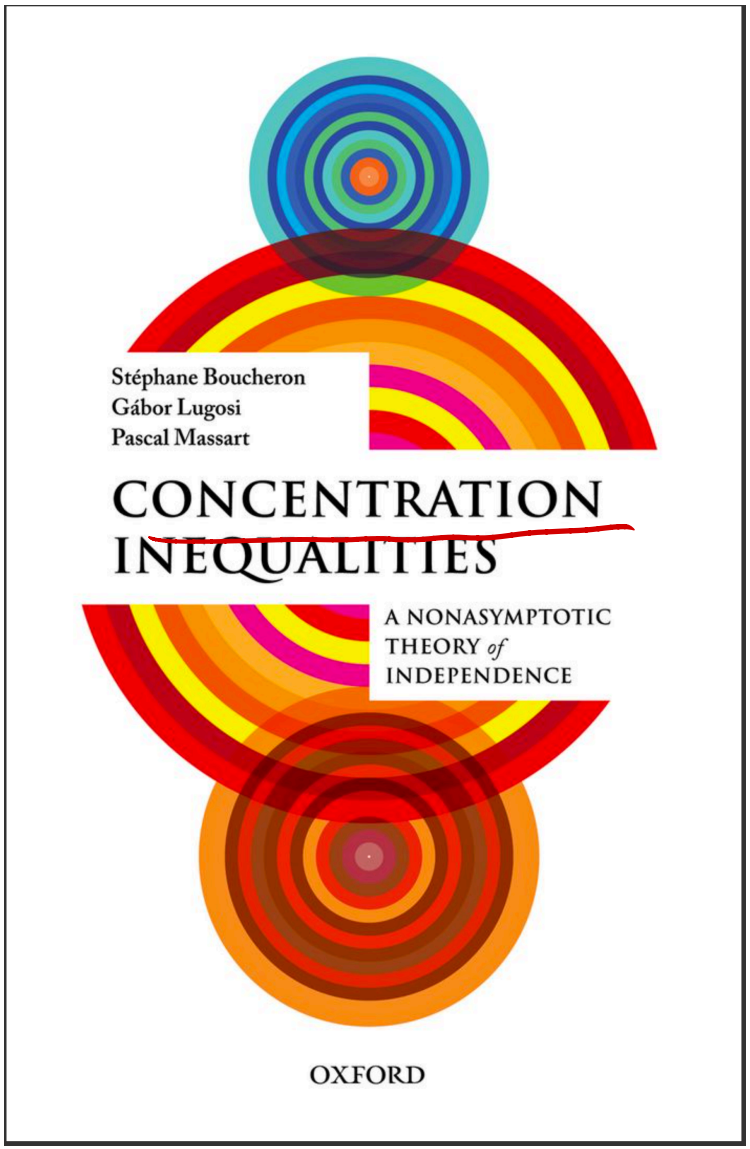
RC



Christian P. Robert & George Casella

- Monte Carlo Statistical Methods
- Springer, 2010

BLM



Stephane Boucheron & Gabor Lugosi & Pascal Massart

- Concentration Inequalities
- Oxford, 2013