Notes of Probability

Author: Wenxiao Yang

Institute: Haas School of Business, University of California Berkeley

Date: 2023

All models are wrong, but some are useful.

Contents

Chapter 1 Metric Spaces Foundations (Lec 01 @ ECON 240A)

1.1 Probability Space

Definition 1.1 (Probability Space)

A **probability space** is triple $(S, \mathbb{B}, P)$, where S is a sample space, $\mathbb{B}$ is a $\sigma-$algebra of events, and P is a probability function.

♣

1.1.1 Sample Space

Definition 1.2 (Sample Space)

The **sample space** S of an experiment is the list of all possible outcomes of the experiment. Elements $s \in S$ are called elementary outcomes.

♣

Example 1.1 **Coin Tossing:** $S = \{H, T\}$.

1.1.2 $\sigma-$Algebra

Definition 1.3 ($\sigma-$Algebra of Events)

A subset (of S) $B \subseteq S$ is called an **event**.

A collection $\mathbb{B}$ of events is a $\sigma-$**algebra** if and only if

1. $\emptyset \in \mathbb{B}$;
2. $B \in \mathbb{B} \Rightarrow B^C \in \mathbb{B}$;
3. $B_1, B_2, ... \in \mathbb{B} \Rightarrow \cup_{i=1}^{\infty} B_i \in \mathbb{B}$.

♣

Note $S = \emptyset^C \in \mathbb{B}$.

Example 1.2 **Coin Tossing:** $S = \{H, T\}$ has $\sigma-$algebra $\mathbb{B} = 2^S = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$ (which is called the "discrete $\sigma-$algebra").

Standard $\sigma-$algebra when $S = \mathbb{R}$:

Definition 1.4 (Borel $\sigma-$algebra)

The **Borel** $\sigma-$algebra on $\mathbb{R}$, $\mathcal{B}(\mathbb{R})$, is the smallest $\sigma-$algebra contains all open sets.

♣

1.1.3 Probability Function

Definition 1.5 (Probability Function)

A function $P : \mathbb{B} \to [0, 1]$ is a **probability function** if and only if

1. $P(S) = 1$;
2. If $B_1, B_2, ... \in \mathbb{B}$ are disjoint event, then $P(\cup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} P(B_i)$.

♣

Note

1. $\mathbb{B}$ *is the domain of* P.
2. $S = \emptyset^C \in \mathbb{B}$ *because* $\mathbb{B}$ *is a* $\sigma-$*algebra.*
3. $P(\emptyset) = 0$ *because* $\emptyset$ *and* S *are disjoint,* $P(\emptyset) + P(S) = P(\emptyset \cup S) = P(S)$.

Example 1.3 **Coin Tossing:** $S = \{H, T\}$ has $\sigma-$algebra $\mathbb{B} = 2^S = \{\emptyset, \{H\}, \{T\}, \{H, T\}\}$ (which is called the "discrete $\sigma-$algebra").

$$P(\emptyset) = 0, P(\{H\}) = P(\{T\}) = \frac{1}{2}, P(\{H, T\}) = 1$$

Note ***Specifying*** $B\&P()$ ***in general:***

1. *We don't specify* $\mathbb{B}$ *unless we have to.*
2. *It suffices to specify* $P()$ *on a subset of* $\mathbb{B}$. *(In the example, we can only specify* $P(\{H\})$ *and* $P(\{T\})$*).*

Note ***Generalization:***

If S *is (at most) countable, then* $\mathbb{B} = 2^S$ *is a standard choice. It suffices to specify* $P(\{s\}), s \in S$.

1.2 Random Variables

1.2.1 Random Variable

Definition 1.6 (Random Variable)

Let $(S, \mathbb{B}, P)$ be a probability space. A **random variable** is a function $X : S \to \mathbb{R}$, which is "Borel measurable".

♣

Definition 1.7 (Probability Function of a Random Variable)

Any random variable X induces a probability space $(S_X, \mathbb{B}_X, P_X)$, where $S_X = \mathbb{R}$, $\mathbb{B}_X = \mathcal{B}(\mathbb{R})$, and $P_X = P \circ X^{-1}$, that is, for all $B \in \mathcal{B}(\mathbb{R})$,

$$P_X(B) = P \circ X^{-1}(B) \triangleq P(\{s \in S : X(s) \in B\})$$

Convention: We work with (S_X, B_X, P_X) and focus on specifying P_X.

♣

1.2.2 Represeting / Specifying P_X: Cumulative Distribution Function

Definition 1.8 (Cumulative Distribution Function)

The **cumulative distribution function** (cdf) of X is the function $F_X : \mathbb{R} \to [0,1]$ given by, for all $x \in \mathbb{R}$,

$$F_X(x) \triangleq P_X((-\infty, x]) = P(\{s \in S : X(s) \leq x\})$$

♣

Theorem 1.1 (Conditions of being a CDF)

A function $F_X : \mathbb{R} \to [0,1]$ is cdf if and only if

(i). $\lim_{x\to-\infty} F_X(x) = 0$ *and* $\lim_{x\to\infty} F_X(x) = 1$*;*

(ii). F_X *is non-decreasing;*

(iii). F_X *is right-continuous.*

♡

1.2.3 Correspondence Theorem; CDF $\Leftrightarrow$ Probability Function

Theorem 1.2 (Correspondence Theorem)

Let X&Y be random variables and let the associated probability functions and cdfs be P_X&P_Y and F_X&F_Y. Then,

$$F_X = F_Y \text{ if and only if } P_X = P_Y$$

♡

So, we can specify P_X (on $\mathcal{B}(\mathbb{R})$) by specifying F_X (on $\mathbb{R}$).

Example 1.4 Uniform Distribution A random variable X has a (standard) uniform distribution, $X \sim U[0,1]$, if and only if its cdf is given by

$$F_X(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x \leq 1 \\ 1, & x > 1 \end{cases}$$

1.2.4 Discrete / Continuous Random Variable

Definition 1.9 (Discrete / Continuous Random Variable)

Let X be a random variable with cdf F_X.

- X is **discrete** if and only if $\exists f_X : \mathbb{R} \to [0,1]$ such that

$$F_X(x) = \sum_{t \leq x} f_X(x), \forall x \in \mathbb{R}$$

in which case f_X is called the **probability mass function** (pmf) of X.

- X is **continuous** if and only if $\exists f_X : \mathbb{R} \to [0,1]$ such that

$$F_X(x) = \int_{-\infty}^{x} f_X(t)dt, \forall x \in \mathbb{R}$$

in which case f_X is called the **probability density function** (pdf) of X.

Proposition 1.1

- *F_X is a step function if and only if X is discrete.*

  If X is discrete, then $f_X(x) = P_X(\{x\}), \forall x \in \mathbb{R}$.

- *F_X is absolutely continuous if and only if X is continuous.*

  If X is continuous, then $P_X(\{x\}) = 0, \forall x \in \mathbb{R}$.

  If X is continuous, then $f_X(x) = \frac{dF_X(x)}{dx}$ "almost everywhere".

Note ***Useful Facts:***

1. *A function $f : \mathbb{R} \to [0,1]$ is a pmf if and only if $\sum_{x\in\mathbb{R}} f(x) = 1$.*
2. *A function $f : \mathbb{R} \to \mathbb{R}$ is a pdf if and only if $\int_{-\infty}^{\infty} f(x)dx = 1$.*

Example 1.5 Normal X has a standard normal distribution, $X \sim N(0,1)$ if and only if it is continuous with pdf $f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}x^2}, x \in \mathbb{R}$.

1.3 Expectation, Moment, Mean, Central Moment, Variance

1.3.1 Expectation

Definition 1.10 (Expecetd Values)

Let X be a discrete/continuous random variable with pmf/pdf f_X and let $g : \mathbb{R} \to \mathbb{R}$ be a function. The **expected value** of $g(x)$ is

$$\mathbb{E}[g(x)] = \begin{cases} \sum_{x\in\mathbb{R}} g(x) f_X(x) & \text{if } X \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x) f_X(x)dx & \text{if } X \text{ is continuous} \end{cases}$$

Note

- *Tacit Assumptions: $\sum_{x\in\mathbb{R}} |g(x)| f_X(x) < \infty$ and $\int_{-\infty}^{\infty} |g(x)| f_X(x)dx < \infty$.*
- *Alternative Formulation (for continuous): $\int_{-\infty}^{\infty} g(x)dF_X(x)$, where $F_X(x)$ is the cdf of X.*

1.3.2 Moment, Mean, Central Moment, Variance

Definition 1.11 (Moment, Mean, Central Moment, Variance)

1. The k^{th} **moment** of X is $\mu'_k = \mathbb{E}(X^k), k \in \mathbb{N}$.
2. The **mean** of X is $\mu = \mathbb{E}(X) = \mu'_1$.
3. The k^{th} **central moment** of X is $\mu_k = \mathbb{E}[(X-\mu)^k], k \in \mathbb{N}$.
4. The **variance** of X is $\text{Var}(X) = \sigma^2 = \mathbb{E}[(X-\mu)^2] = \mu_2$.

♣

Example 1.6 $N(0,1)$ Suppose $X \sim N(0,1)$. We claim

- $\mathbb{E}(X) = 0$, $\text{Var}(X) = 1$, which can be easily shown by the fact $\frac{df_X(x)}{dx} = -xf_X(x)$.

Example 1.7 Cauchy If X is continuous with pdf,

$$f_X(x) = \frac{1}{\pi}\frac{1}{1+x^2}, x \in \mathbb{R}$$

Then, the mean doesn't exist

$$\int_{\infty}^{\infty} |x| f_X(x) dx = \infty$$

1.3.3 $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$, $\mathbb{E}(aX+b) = a\mathbb{E}(X) + b$, $\text{Var}(aX+b) = a^2\text{Var}(X)$

Proposition 1.2 (General Facts)

(i). $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \leq \mathbb{E}(X^2)$.

(ii). $\mathbb{E}(aX+b) = a\mathbb{E}(X) + b$.

(iii). $\text{Var}(aX+b) = a^2\text{Var}(X)$.

(iv). $\mathbb{E}[g(X)] = g[\mathbb{E}(X)]$ *when* $g(X) = aX + b$ *(linear). General form can be found in Jensen's inequality (Theorem 3.1).*

♠

1.4 Univariate Distribution

1.4.1 Bernoulli Distribution

Definition 1.12 (Bernoulli Distribution)

A random variable X is called **Bernoulli** distribution with parameter $p \in [0,1]$, $X \sim \text{Ber}(p)$, if and only if it is discrete with pmf

$$f(x \mid p) = \begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases}$$

♣

- Mean: $\mathbb{E}(X) = \sum_{x\in\mathbb{R}} xf(x \mid p) = p$.
- Variance: $\text{Var}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 = p - p^2 = p(1-p)$.

1.4.2 Normal Distribution

Definition 1.13 (Normal Distribution)

A random X has a **normal distribution** (with mean μ and σ^2), $X \sim N(\mu, \sigma^2)$, if and only if it is continuous with pdf

$$f(x \mid \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}},\ x \in \mathbb{R}$$

♣

- If $Z \sim N(0,1)$, then $X = \sigma Z + \mu \sim N(\mu, \sigma^2)$.

1.4.3 Uniform Distribution

Definition 1.14 (Uniform Distribution)

X has a **uniform** distribution (on $[a,b]$), $X \sim U[a,b]$, if and only if it is continuous with pdf

$$f(x \mid a, b) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a,b] \\ 0, & \text{otherwise} \end{cases}$$

♣

- $\mathbb{E}(X) = \frac{a+b}{2}$.
- If $Z \sim U[0,1]$, then $X = a + (b-a)Z \sim U[a,b]$.

1.5 Exponential Families

Definition 1.15 (Exponential Family)

A family $\{f(\cdot \mid \theta) : \theta \in \Theta\}$ (Θ is a set of parameters) of pmfs/pdfs is an **exponential family** if and only if

$$f(x \mid \epsilon) = h(x)c(\epsilon)e^{\sum_{j=1}^{k} t_j(x)w_j(\epsilon)}, \forall x \in \mathbb{R}, \epsilon \in \Theta$$

for some $k \in \mathbb{N}$, $h : \mathbb{R} \to \mathbb{R}_+$, $c : \Theta \to \mathbb{R}_{++} \triangleq (0, \infty)$, $t_j : \mathbb{R} \to \mathbb{R}$, $w_j : \Theta \to \mathbb{R}$ $\forall j \in \{1, \cdots, k\}$.

♣

Example 1.8 $N(\mu, \sigma^2)$ Suppose $X \sim N(\mu, \sigma^2)$ for some $\mu \in \mathbb{R}$ and some $\sigma^2 > 0$. Then, $\theta = (\mu, \sigma^2)$ and

$\Theta = \mathbb{R} \times \mathbb{R}_{++}$. The pdf can be written as

$$\begin{aligned} f(x \mid \mu, \sigma^2) &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x^2+\mu^2-2\mu x)} \\ &= \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu^2}{2\sigma^2}} e^{\frac{\mu}{\sigma^2}x - \frac{1}{2\sigma^2}x^2} \end{aligned}$$

We can have $h(x) = 1, c(\mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{\mu^2}{2\sigma^2}}, t_1(x) = x, w_1(\mu, \sigma^2) = \frac{\mu}{\sigma^2}, t_2(x) = x^2, w_2(\mu, \sigma^2) = -\frac{1}{2\sigma^2}$.

Example 1.9 Ber(p) Suppose $X \sim \text{Ber}(p)$ for some $p \in (0, 1)$. Then $\theta = p$ and $\Theta = (0, 1) \subseteq \mathbb{R}$. The pdf can be written as

$$\begin{aligned} f(x \mid p) &= \begin{cases} p, & \text{if } x = 1 \\ 1-p, & \text{if } x = 0 \\ 0, & \text{otherwise} \end{cases} \\ &= \begin{cases} p^x(1-p)^{1-x}, & \text{if } x \in \{0, 1\} \\ 0, & \text{otherwise} \end{cases} \\ &= p^x(1-p)^{1-x}\mathbf{1}(x \in \{0, 1\}) \\ &= \mathbf{1}(x \in \{0, 1\})(1-p)\left(\frac{p}{1-p}\right)^x \\ &= \mathbf{1}(x \in \{0, 1\})(1-p)e^{\left[\ln\left(\frac{p}{1-p}\right)x\right]} \end{aligned}$$

we can have $h(x) = \mathbf{1}(x \in \{0, 1\}), c(p) = 1 - p, t(x) = x, w(p) = \ln\left(\frac{p}{1-p}\right)$.

Note *For any $r > 0$, $r^x = e^{(\ln r)x}$*

Proposition 1.3

1. *If $f(\cdot \mid \epsilon)$ is pdf, then $1 = \int_{-\infty}^{\infty} f(x \mid \epsilon)dx = c(\epsilon)\int_{-\infty}^{\infty} h(x)e^{\sum_{j=1}^k t_j(x)w_j(\epsilon)}dx$. So,*
$$\int_{-\infty}^{\infty} h(x)e^{\sum_{j=1}^k t_j(x)w_j(\epsilon)}dx = c(\epsilon)^{-1}$$
2. *The support of $f(\cdot \mid \epsilon)$ does not depend on ϵ:*
$$\text{supp}(f(\cdot \mid \epsilon)) = \{x : f(x \mid \epsilon) > 0\} = \{x : h(x) > 0\}$$

Example 1.10 Uniform distribution cannot be written as an exponential family form $U[a, b]$ has pdfs,

$$f(x \mid a, b) = \begin{cases} \frac{1}{b-a}, & \text{if } x \in [a, b] \\ 0, & \text{otherwise} \end{cases}$$

which cannot be written as an exponential family form.

1.6 Multiple Random Variables

Definition 1.16 (n-dimensional Random Vector)

An **n-dimensional random vector** ($n \in \mathbb{N}$) is a vector $X = (X_1, ..., X_n)'$ ($'$ means the transpose), where $X_1, ..., X_n$ are random variables. (defined on the same probability space).

♣

1.6.1 Bivariate Random Vector

We consider the most important special case ($n = 2$).

Definition 1.17 (Bivariate Random Vector)

A **bivariate random vector** is a vector $(X, Y)'$, where X and Y are random variables.

♣

Definition 1.18 (Joint CDF)

The **joint cumulative distribution function (cdf)** of (X, Y) is the function $F_{X,Y} : \mathbb{R}^2 \to [0, 1]$ given by

$$F_{X,Y}(x, y) = P(X \leq x, Y \leq y), \forall (x, y) \in \mathbb{R}^2$$

♣

Proposition 1.4

The correspondence theorem generalizes to $\mathbb{R}^2$ ($\mathbb{R}^n$): $\exists$ necessary and sufficient conditions for $F_{X,Y}$ to be a joint cdf.

♠

Definition 1.19 (Discrete / Continuous Random Vector)

Let (X, Y) be a bivariate random vector with joint cdf $F_{X,Y}$.

- (X, Y) is **discrete** if and only if $\exists f_{X,Y} : \mathbb{R}^2 \to [0, 1]$ such that

$$F_{X,Y}(x, y) = \sum_{s \leq x} \sum_{t \leq y} f_{X,Y}(s, t), \ \forall (x, y) \in \mathbb{R}^2$$

  in which case $f_{X,Y}$ is the joint pmf of $(X, Y)'$.

- (X, Y) is **continuous** if and only if $\exists f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_+$ such that

$$F_{X,Y}(x, y) = \int_{-\infty}^{x} \int_{-\infty}^{y} f_{X,Y}(s, t) ds dt, \ \forall (x, y) \in \mathbb{R}^2$$

  in which case $f_{X,Y}$ is the joint pdf of $(X, Y)'$.

♣

Proposition 1.5

Some useful facts:

(i). A function $f_{X,Y} : \mathbb{R}^2 \to [0, 1]$ is a joint pmf if and only if $\sum_{x \in \mathbb{R}} \sum_{y \in \mathbb{R}} f_{X,Y}(x, y) = 1$

(ii). A function $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}_+$ is a joint pdf if and only if $\int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{X,Y}(x, y) dx dy = 1$

♠

Definition 1.20 (Expected Value)

Let (X, Y) be a discrete/continuous random vector with pmf/pdf $f_{X,Y}$ and let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function.

The **expected value** of $g(X, Y)$ is

$$\mathbb{E}[g(X,Y)] = \begin{cases} \sum_{x\in\mathbb{R}} \sum_{y\in\mathbb{R}} g(x,y) f_{X,Y}(x,y), & \text{if } (X,Y) \text{ is discrete} \\ \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} g(x,y) f_{X,Y}(x,y) dx dy, & \text{if } (X,Y) \text{ is continuous} \end{cases}$$

♣

1.6.2 Marginal CDF, PMF, PDF

Definition 1.21 (Marginal cdf, pmf, pdf)

Let (X, Y) be a bivariate random vector with joint cdf $F_{X,Y}$. Then,

(i). X is a random variable with **marginal cdf** F_X given by

$$F_X(x) = \lim_{y\to\infty} F_{X,Y}(x,y)$$

(ii). X is discrete with **marginal pmf** $f_X(x)$ given by

$$f_X(x) = \sum_{y\in\mathbb{R}} f_{X,Y}(x,y)$$

(iii). X is continuous with **marginal pdf** $f_X(x)$ given by

$$f_X(x) = \int_{-\infty}^{\infty} f_{X,Y}(x,y) dy$$

♣

1.6.3 Independent Random Variables

Definition 1.22 (Independent Random Variables)

Two random variables X, Y with joint cdf $F_{X,Y}$ and marginal cdfs F_X, F_Y are **independent** (denoted by $X \perp Y$) if and only if

$$F_{X,Y}(x,y) = F_X(x) F_y(y), \ \forall (x,y) \in \mathbb{R}^2$$

♣

Proposition 1.6

If $(X, Y)'$ is a discrete/continuous random vector with joint pmf/pdf $f_{X,Y}$ and marginal pmf/pdf f_X, f_Y, then X, Y are independent ($X \perp Y$) if and only if

$$f_{X,Y}(x,y) = f_X(x) f_y(y), \ \forall (x,y) \in \mathbb{R}^2$$

♠

1.6.4 Conditional PMF, PDF, Expected Value, Mean, Variance

Definition 1.23 (Conditional pmf/pdf)

Let $(X, Y)'$ be a discrete/continuous random vector with joint pmf/pdf $f_{X,Y}$ and marginal pmf/pdf f_Y of Y. For any $y \in \mathbb{R}$, a conditional pmf/pdf of X given $Y = y$ is any function $f_{X|Y}(\cdot \mid y) : \mathbb{R} \to \mathbb{R}_+$ such that

$$f_{X,Y}(x, y) = f_{X|Y}(x \mid y) \cdot f_Y(y), \ \forall x \in \mathbb{R}$$

♣

Proposition 1.7

Suppose $(X, Y)'$ is discrete, $f_{X|Y}(\cdot \mid y)$ exists for every $y \in \mathbb{R}$ and is interpretable if $f_Y(y) > 0$.

♠

Definition 1.24 (Conditional Expected Value)

Let $(X, Y)'$ be a discrete/continuous random vector and let $g : \mathbb{R}^2 \to \mathbb{R}$ be a function. For any $y \in \mathbb{R}$, a **conditional expected value** of $g(X, Y)$ given $Y = y$ is

$$\mathbb{E}_{X|Y}[g(X, Y) \mid y] = \begin{cases} \sum_{x \in \mathbb{R}} g(x, y) f_{X|Y}(x \mid y), & \text{if } (X, Y) \text{ is discrete} \\ \int_{-\infty}^{\infty} g(x, y) f_{X|Y}(x \mid y) dx, & \text{if } (X, Y) \text{ is continuous} \end{cases}$$

♣

Definition 1.25 (Conditional Mean/Variance)

- The **conditional mean** of X given $Y = y$ is

$$\mathbb{E}_{X|Y}(X \mid y)$$

- The **conditional variance** of X given $Y = y$ is

$$\begin{aligned} \text{Var}_{X|Y}(X \mid y) &= \mathbb{E}_{X|Y}[(X - \mathbb{E}_{X|Y}(X \mid Y))^2 \mid y] \\ &= \mathbb{E}_{X|Y}(X^2 \mid y) - (\mathbb{E}_{X|Y}(X \mid y))^2 \end{aligned}$$

♣

1.6.5 Law of Iterated Expectation: $\mathbb{E}_X(X) = \mathbb{E}_Y[\mathbb{E}_{X|Y}(X \mid Y)]$

Theorem 1.3 (Law of Iterated Expectation: Adam's Law)

Suppose $(X, Y)'$ is continuous and suppose that $\mathbb{E}_X(X)$ exists (i.e., $\mathbb{E}_X(|X|) < \infty$). Then,

$$\mathbb{E}_X(X) = \mathbb{E}_Y[\mathbb{E}_{X|Y}(X \mid Y)]$$

♡

Proof 1.1

$$\begin{aligned}\mathbb{E}_X(X) &= \mathbb{E}_{X,Y}(X)\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X,Y}(x,y)dxdy\\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} x f_{X|Y}(x\mid y) f_Y(y)dydx\\ &= \int_{-\infty}^{\infty}\left\{\int_{-\infty}^{\infty} x f_{X|Y}(x\mid y)dx\right\} f_Y(y)dy\\ &= \mathbb{E}_Y[\mathbb{E}_{X|Y}(X\mid Y)]\end{aligned}$$

Example 1.11 Cauchy Distribution Suppose $X = \frac{Z_1}{Z_2}\mathbf{1}_{\{Z_2\neq 0\}}$, where $Z_1 \sim N(0,1)$, $Z_2 \sim N(0,1)$, and $Z_1 \perp Z_2$.

Let $Y = Z_2$. If $Y = 0$, then $X = 0$; If $Y = y \neq 0$, then $X \mid Y = y \sim N(0, \frac{1}{y^2})$.

$$\mathbb{E}_{X|Y}(X|y) = 0, \forall y \in \mathbb{R}$$

$$\Rightarrow \mathbb{E}_Y[\mathbb{E}_{X|Y}(X|y)] = 0$$

We can't conclude that $\mathbb{E}_X(X) = \mathbb{E}_Y[\mathbb{E}_{X|Y}(X|y)] = 0$ because we need to check the existence of $\mathbb{E}_X(X)$ firstly.

1.6.6 Law of Total Variance: $\mathrm{Var}_X(X) = \mathbb{E}_Y[\mathrm{Var}_{X|Y}(X \mid Y)] + \mathrm{Var}_Y[\mathbb{E}_{X|Y}(X \mid Y)]$

Theorem 1.4 (Law of Total Variance: Conditional Variance Identity)

If $(X,Y)'$ is a bivariate random vector, then

$$\mathrm{Var}_X(X) = \mathbb{E}_Y[\mathrm{Var}_{X|Y}(X \mid Y)] + \mathrm{Var}_Y[\mathbb{E}_{X|Y}(X \mid Y)]$$

(provided that the expectations exist)

♡

1.6.7 Covariance

Definition 1.26 (Covariance)

The **covariance** of X and Y is

$$\mathrm{Cov}(X,Y) = \mathbb{E}[(X-\mu_X)(Y-\mu_Y)]$$

where $\mu_X = \mathbb{E}[X]$ and $\mu_Y = \mathbb{E}[Y]$.

♣

Proposition 1.8

Covariance has the following properties:

(i). $\mathrm{Cov}(X,X) = \mathrm{Var}(X)$*;*

(ii). $\text{Cov}(X, Y) = \mathbb{E}(XY) - \mu_X \mu_Y$*;*

(iii). $\text{Cov}(X, Y) = \text{Cov}(Y, X)$*;*

(iv). If X *and* Y *are independent* ($X \perp Y$)*, then* $\text{Cov}(X, Y) = 0$*;*

(v). $\text{Var}(aX + bY + c) = \text{Var}(aX + aY) = a^2\text{Var}(X) + b^2\text{Var}(Y) + 2ab\text{Cov}(X, Y)$

Chapter 2 Statistics

2.1 Random Sampling

Definition 2.1 (Random Sample)

A **random sample** is a collection $X_1, ..., X_n$ of random variables that are (mutually) independent and identical marginal distributions.

$X_1, ..., X_n$ are called "independent and identically distributed". The notation is $X_i \sim i.i.d.$

♣

Definition 2.2 (Statistic)

If $X_1, ..., X_n$ is a random sample and $T : \mathbb{R}^n \to \mathbb{R}^k$ (for some $k \in \mathbb{N}$), then $T(X_1, ..., X_n)$ is called a **statistic**.

♣

2.1.1 Sample Mean and Sample Variance

Definition 2.3 (Sample Mean and Sample Variance)

1. The **sample mean** is $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$;
2. The **sample variance** is $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n-1}(\sum_{i=1}^{n} X_i^2 - n\bar{X}^2)$

♣

Note *We use "$X_i \sim$ i.i.d(μ, σ^2)" to denote a random sample from a distribution with mean μ and variance σ^2.*

Theorem 2.1 ($\mathbb{E}(\bar{X})$, Var$(\bar{X})$, $\mathbb{E}(S^2)$)

Suppose $X_1, ..., X_n$ is a random sample from a distribution with mean μ and variance σ^2 (denoted by $X_i \sim$ i.i.d(μ, σ^2)). Then,

(a). $\mathbb{E}(\bar{X}) = \mu$;

(b). Var$(\bar{X}) = \frac{\sigma^2}{n}$;

(c). $\mathbb{E}(S^2) = \sigma^2$.

♡

2.1.2 Distributional Properties

Theorem 2.2

If $X_i \sim$ i.i.d. $N(\mu, \sigma^2)$, then

(a). $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$

(b). $\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$

(c). $\bar{X} \perp S^2$

♡

Theorem 2.3 ("Asymptotics")

If $X_i \sim$ i.i.d. (μ, σ^2) *and if* n *is "large", then*

(a). $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ *(converges in distribution) by CLT 3.4;*

(b). $S^2 = \sigma^2$ *by LLN;*

♡

2.1.3 Order Statistics

Definition 2.4 (Order Statistics)

If $X_1, ..., X_n$ is a random sample, then the **characteristics** are the sample values placed in ascending order. Notation:

$$X_{(1)} \leq X_{(2)} \leq ... \leq X_{(n)}$$

♣

Proposition 2.1 (Distribution of $X_n = \max_{i=1,...,n} X_i$)

If $X_1, ..., X_n$ *is a random sample form a distribution with cdf* F *(denoted by "*$X_i \sim$ i.i.d. F*"), then*

$$F_{X_{(n)}}(x) = P(X_{(n)} \leq x) = F^n(x)$$

♠

2.2 Basic Statistics

In statistics, we define **data** be a vector $x = (x_1, ..., x_n)'$ of numbers.

Assumption ***[Fundamental Assumption]*** x *is the realization of a random vector* $X = (X_1, ..., X_n)'$.

Objective: Using x to give (data-based) answers to questions about the distribution of X.

Probability vs. Statistics:

- Probability: Distribution known, outcome unknown;
- Statistics: Distribution unknown, outcome known.

Setting: $X_1, ..., X_n$ is a random sample from a discrete/continuous distribution with pmf/pdf $f(\cdot \mid \theta)$, where $\theta \in \Theta$ is unknown.

Types of Statistical Inference:

- Point estimation $\Rightarrow$ "What is θ?";
- Hypothesis testing $\Rightarrow$ "Is $\theta = \theta_0$?";
- Interval estimation $\Rightarrow$ "Which values of θ are 'plausible'?".

Example 2.1 Examples of Statistical Models

(1). $x_i \sim$ i.i.d. Bernoulli(p), where p is unknown.

(2). $x_i \sim$ i.i.d. $U(0, \theta)$, where $\theta > 0$ is unknown.

(3). $x_i \sim$ i.i.d. $N(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown.

2.3 Point Estimation

Suppose $X_1, ..., X_n$ is a random sample from a discrete/continuous distribution with pmf/pdf $f(\cdot \mid \theta)$, where $\theta \in \Theta$ is unknown.

Definition 2.5 (Point Estimator)

A **point estimator** (of θ) is a function of $(X_1, ..., X_n)$.

Notation: $\hat{\theta} = \hat{\theta}(X_1, ..., X_n)$.

♣

Agenda

(1). Constructing point estimators

- Method of moments;
- Maximum likelihood.

(2). Comparing estimators

- Pairwise comparisons;
- Finding 'optimal' estimators.

2.3.1 Method of Moments (MM)

Definition 2.6 (Method of Moments in $\mathbb{R}^1$)

Suppose $\Theta \subseteq \mathbb{R}^1$. A **method of moments** estimator $\hat{\theta}_{MM}$ solves

$$\mu(\hat{\theta}_{MM}) = \bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

where $\mu : \Theta \to \mathbb{R}$ is given by

$$\mu(\theta) = \begin{cases} \sum_{x \in \mathbb{R}} x f(x \mid \theta), & \text{if } X_i \text{ are discrete} \\ \int_{-\infty}^{\infty} x f(x \mid \theta) dx, & \text{if } X_i \text{ are continuous} \end{cases}$$

♣

Remark Existence of $\mu(\cdot)$ is assumed; Existence (and uniqueness) of $\hat{\theta}_{MM}$ is assumed.

Example 2.2

1. Suppose $X_i \sim$ i.i.d. Ber(p) where $p \in [0, 1]$ is unknown. The moment function is

$$\mu(p) = p$$

Then, the estimator is

$$\hat{p}_{MM} = \mu(\hat{p}_{MM}) = \bar{X}$$

Remark $\hat{p}_{MM} = \bar{X}$ is the 'best' estimator of p.

2. Suppose $X_i \sim$ i.i.d.$U(0,\theta)$ where $\theta > 0$ is unknown.

   Remark Non-regular statistical model: parameter dependent support, where supp$X = [0,\theta]$.

   The moment function is

$$\mu(\theta) = \frac{\theta}{2}$$

   Then, the estimator is

$$\hat{\theta}_{MM} = 2\mu(\hat{\theta}_{MM}) = 2\bar{X}$$

   Remark $\hat{\theta}_{MM}$ is not a very good estimator of θ. Concern $X_i > \hat{\theta}_{MM}$ could happen. So, $\max\{\hat{\theta}_{MM}, X_{(n)}\}$ can be better.

Definition 2.7 (Method of Moments in $\mathbb{R}^k$)

Suppose $\Theta \subseteq \mathbb{R}^k$. A **method of moments** estimator $\hat{\theta}_{MM}$ solves

$$\mu'_j(\hat{\theta}_{MM}) = \frac{1}{n}\sum_{i=1}^{n} X_i^j, \quad (j = 1, ..., k)$$

where $\mu'_j : \Theta \to \mathbb{R}$ is given by

$$\mu'_j(\theta) = \begin{cases} \sum_{x\in\mathbb{R}} x^j f(x \mid \theta), & \text{if } X_i \text{ are discrete} \\ \int_{-\infty}^{\infty} x^j f(x \mid \theta) dx, & \text{if } X_i \text{ are continuous} \end{cases}$$

♣

Example 2.3

Suppose $X_i \sim$ i.i.d.$N(\mu, \sigma^2)$ where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown. The moment function is

$$\mu'_1(\mu, \sigma^2) = \mu$$

$$\mu'_2(\mu, \sigma^2) = \mu^2 + \sigma^2$$

Then, the estimator is

$$\mu'_1(\hat{\mu}_{MM}, \hat{\sigma}^2_{MM}) = \hat{\mu}_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

$$\mu'_2(\hat{\mu}_{MM}, \hat{\sigma}^2_{MM}) = \hat{\mu}_{MM} + \hat{\sigma}^2_{MM} = \frac{1}{n}\sum_{i=1}^{n} X_i^2$$

$$\Rightarrow \hat{\mu}_{MM} = \bar{X}$$

$$\hat{\sigma}^2_{MM} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X})^2$$

Remark $\bar{X}$ is the 'best' estimator of μ; An alternative better estimator of σ^2 is $\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$.

2.3.2 Maximum Likelihood (ML)

Definition 2.8 (Maximum Likelihood)

A **maximum likelihood estimator** $\hat{\theta}_{ML}$ solves

$$L(\hat{\theta}_{ML} \mid X_1, ..., X_n) = \max_{\theta \in \Theta} L(\theta \mid X_1, ..., X_n)$$

where $L(\cdot \mid X_1, ..., X_n) : \Theta \to \mathbb{R}_+$ is given by

$$L(\theta \mid X_1, ..., X_n) = \prod_{i=1}^{n} f(X_i \mid \theta),\ \theta \in \Theta$$

♣

Remark $L(\cdot \mid X_1, ..., X_n)$ is called the likelihood function.

Definition 2.9 (Log-Likelihood)

The **log-likelihood** function is

$$l(\theta \mid X_1, ..., X_n) = \log L(\theta \mid X_1, ..., X_n) = \sum_{i=1}^{n} \log f(X_i \mid \theta),\ \theta \in \Theta$$

♣

Example 2.4

1. Suppose $X_i \sim$ i.i.d. Ber(p) where $p \in [0, 1]$ is unknown. The marginal pmf is

$$f(x \mid p) = \begin{cases} p, & x = 1 \\ 1 - p, & x = 0 \\ 0, & \text{otherwise} \end{cases} = p^x (1-p)^{1-x} \mathbf{1}_{\{x \in \{0,1\}\}}$$

Then, the likelihood function is

$$L(p \mid X_1, ..., X_n) = \prod_{i=1}^{n} \left\{ p^{X_i} (1-p)^{1-X_i} \underbrace{\mathbf{1}_{\{X_i \in \{0,1\}\}}}_{=1} \right\}$$
$$= p^{\sum_{i=1}^n X_i} (1-p)^{n - \sum_{i=1}^n X_i},\ p \in [0, 1]$$

and the log-likelihood function is

$$l(p \mid X_1, ..., X_n) = (\sum_{i=1}^{n} X_i) \log p + (n - \sum_{i=1}^{n} X_i) \log(1-p),\ p \in (0, 1)$$

Maximization:

(a). Suppose $0 < \sum_{i=1}^n X_i < n$, we can give the first-order condition:

$$\frac{\partial l(p \mid X_1, ..., X_n)}{\partial p}\Big|_{p=\hat{p}_{ML}} = \frac{\sum_{i=1}^n X_i}{\hat{p}_{ML}} - \frac{n - \sum_{i=1}^n X_i}{n - \hat{p}_{ML}} = 0$$
$$\Rightarrow \hat{p}_{ML} = \frac{\sum_{i=1}^n X_i}{n} = \bar{X}$$

(b). Suppose $\sum_{i=1}^n X_i = 0$, then

$$l(p \mid X_1, ..., X_n) = n \log(1-p),\ p \in [0, 1) \Rightarrow \hat{p}_{ML} = 0$$

(c). Suppose $\sum_{i=1}^{n} X_i = n$, then

$$l(p \mid X_1, ..., X_n) = n \log p,\ p \in (0, 1] \Rightarrow \hat{p}_{ML} = 1$$

All in all,

$$\hat{p}_{ML} = \bar{X}$$

Remark $\hat{p}_{ML} = \bar{X} = \hat{p}_{MM}$ is the 'best' estimator of p.

Chapter 3 Basis

3.1 Covariance and Variance

(1) $\sigma_{XY} = Cov(X,Y) = \mathbb{E}\left[(X - \mathbb{E}X)(Y - \mathbb{E}Y)\right] = \mathbb{E}(XY) - \mathbb{E}X\mathbb{E}Y$

(2) $Cov(aX + b, Y) = aCov(X,Y)$

(3) $Cov(X + Y, W) = Cov(X,W) + Cov(Y,W)$

(4) $Cov(aX + bY, cX + dY) = ac\,Var(X) + (ad + bc)Cov(X,Y) + bd\,Var(Y)$

(5) $Var(aX + bY) = a^2\,Var(X) + 2ab\,Cov(X,Y) + b^2\,Var(Y)$

3.2 Conditional Expectation and Variance

(1) Conditional variance:

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

(2) **Law of Total Expectation:**

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

(3) **Law of Iterated Expectation (Adam's Law):**

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

(4) **Adam's Law with extra conditioning:**

$$\mathbb{E}\left(\mathbb{E}(Y|X,Z)|Z\right) = \mathbb{E}(Y|Z)$$

(5) **Law of Total Variance:**

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

3.3 Gambler's Ruin

Suppose a gambler at each round either wins a dollar or loses a dollar with probability $\frac{1}{2}$ each. Suppose the gambler starts at k dollars. He stops when either he reaches his goal of N dollars or he goes bankrupt and loses all his money.

Let A be the event that the gambler is ruined.

$$P(A|x=k) = \frac{1}{2}P(A|x=k-1) + \frac{1}{2}P(A|x=k+1)$$

$$\Rightarrow p_k - p_{k-1} = p_{k+1} - p_k$$

According to the setting, $p_0 = 1, p_N = 0$, then $p_k = \frac{N-k}{N}$.

3.4 Moment Generating Function (MGF)

Definition 3.1 (Moment Generating Function (MGF))

Let X be a random variable. The moment generating function (mgf) of X, denoted by $M_X(t)$:

$$M_X(t) = \mathbb{E}[e^{tX}]$$
$$= \mathbb{E}\left[1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \cdots\right] = \sum_{n=0}^{\infty} \frac{\mathbb{E}[X^n]t^n}{n!}$$

♣

Let X be a random variable. The moment generating function (mgf) of X, denoted by $M_X(t)$:

$$\underline{M_X(t) = \mathbb{E}[e^{tX}]} = \mathbb{E}\left[1 + tX + \frac{(tX)^2}{2!} + \frac{(tX)^3}{3!} + \cdots\right] = \sum_{n=0}^{\infty} \frac{\mathbb{E}[X^n]t^n}{n!}$$

We can find $\frac{\partial M_X(t)}{\partial t} = \mathbb{E}\left[Xe^{tX}\right]$, then $\frac{\partial M_X(0)}{\partial t} = \mathbb{E}\left[X\right]$. More generally, we can find that

$$\frac{\partial^n M_X(0)}{\partial t^n} = \mathbb{E}\left[X^n\right], n = 1, 2, ...$$

Why MGF is useful?

(1) If X, Y are independent, then

$$M_{X+Y}(t) = \mathbb{E}[e^{(X+Y)t}] = \mathbb{E}[e^t X]\mathbb{E}[e^t Y] = M_X(t)M_Y(t)$$

(2) Unique random variable (RV) $\Leftrightarrow$ unique MGF

3.5 Inequality

3.5.1 Cauchy-Schwarz inequality: $|\mathbb{E}XY| \leq \sqrt{\mathbb{E}X^2 \cdot \mathbb{E}Y^2}$

For any r.vs. X and Y with finite variance: $|\mathbb{E}XY| \leq \sqrt{\mathbb{E}X^2 \cdot \mathbb{E}Y^2}$

Example 3.1Second Moment Method X is a non-negative r.v. We want to find an upper bound on $P(X = 0)$.

Because X is non-negative, $X = X \cdot \mathbf{I}_{X>0} = \begin{cases} X, & X > 0 \\ 0, & X = 0 \end{cases}$. Hence,

$$\mathbb{E}X = \mathbb{E}X \cdot \mathbf{I}_{X>0} \leq \sqrt{\mathbb{E}X^2 \cdot \mathbf{I}_{X>0}^2} = \sqrt{\mathbb{E}X^2}\sqrt{P(X>0)}$$
$$\Rightarrow P(X>0) \geq \frac{(\mathbb{E}X)^2}{\mathbb{E}X^2} \Rightarrow P(X=0) = 1 - P(X>0) \leq \frac{Var(X)}{\mathbb{E}X^2}$$

3.5.2 Jensen's Inequality: convex $g \Rightarrow \mathbb{E}(g(X)) \geq g(\mathbb{E}(X))$

Theorem 3.1 (Jensen's Inequality)

If g is convex $\mathbb{E}(g(X)) \geq g(\mathbb{E}X)$; If g is concave $\mathbb{E}(g(X)) \leq g(\mathbb{E}X)$. ♡

3.5.3 Markov's Inequality: $P(|X| \geq a) \leq \frac{\mathbb{E}|X|}{a}$

For any r.v. X and a constant $a > 0$. $P(|X| \geq a) \leq \frac{\mathbb{E}|X|}{a}$

Proof 3.1

$Y = \frac{|X|}{a}$, $Y \geq \mathbb{I}_{Y \geq 1} \Rightarrow \mathbb{E}Y \geq P(Y \geq 1) \Rightarrow \frac{\mathbb{E}|X|}{a} \geq P(|X| \geq a)$

Note: Markov's Inequality can also be written as $P(X \geq a) \leq \frac{\mathbb{E}X}{a}$, $a > 0$, X is non-negative r.v.

3.5.4 Chebychev's inequality: $P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$

Let X be any r.v. with mean μ, variance $\sigma^2 < \infty$. Then for $a > 0$, $P(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$.

Proof 3.2

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

3.5.5 Chernoff Inequality: $P(X \geq a) \leq \frac{\mathbb{E}e^{tX}}{e^{ta}}$

For any r.v. X and constant $a > 0, t > 0$. $P(X \geq a) \leq \frac{\mathbb{E}e^{tX}}{e^{ta}}$

Proof 3.3

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

3.6 Law of Large Numbers (LLN)

Describe the behavior of the sample mean of i.i.d. as the sample size grows.

$x_1, x_2, \ldots, x_n$ i.i.d. with some distribution. $\mu < \infty, \sigma^2 < \infty, \bar{x} = \frac{1}{n}(x_1 + x_2 + \cdots + x_n)$.

3.6.1 Weak Law of Large Numbers (wLLN)

Theorem 3.2 (Weak Law of Large Numbers (wLLN))

The weak law of large numbers (also called Khinchin's law) states that the sample average converges in probability towards the expected value.

$$\overline{X}_n \xrightarrow{P} \mu \qquad \text{when } n \to \infty.$$

That is, for any positive number ε,

$$\lim_{n\to\infty} \Pr\big(\,|\overline{X}_n - \mu| < \varepsilon\,\big) = 1.$$

♡

Proof 3.4

Prove by Chebychev's inequality.

$$P(|\bar{x} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \quad (Var\bar{x} = \frac{\sigma^2}{n})$$

$$\lim_{n\to\infty} \frac{\sigma^2}{n\varepsilon^2} = 0$$

$$\Rightarrow \lim_{n\to\infty} P(|\bar{x} - \mu| > \varepsilon) \text{ also converges to } 0.$$

3.6.2 Strong Law of Large Numbers (sLLN)

Theorem 3.3 (Strong Law of Large Numbers (sLLN))

With probability 1 (wp1) or almost surely (as).

$$\overline{X}_n \xrightarrow{a.s.} \mu \qquad \text{when } n \to \infty.$$

That is,

$$\Pr\Big(\lim_{n\to\infty} \overline{X}_n = \mu\Big) = 1.$$

♡

3.6.3 Differences between convergence in probability (wLLN) and wp1(a.s.) (sLLN)

a) Weak Law of Large Numbers (wLLN)

$$P(|\bar{x} - \mu| \geq \varepsilon) \to 0 \text{ as } n \to +\infty,\ \forall \varepsilon > 0$$

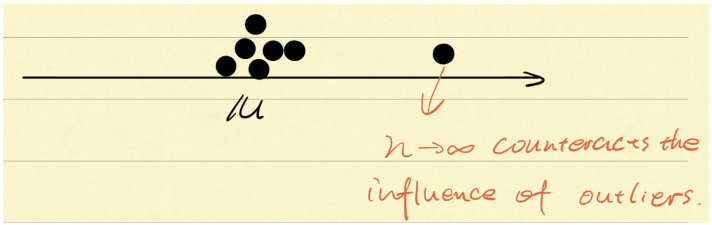

Figure 3.1: convergence in probability

b) Strong Law of Large Numbers (sLLN)

$$P(|\bar{x} - \mu| \geq \varepsilon \text{ as } n \to +\infty) = 0, \ \forall \varepsilon > 0$$

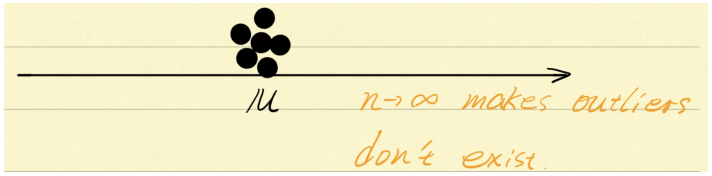

Figure 3.2: wp1(a.s.)

3.7 Central Limit Theorem (CLT)

Theorem 3.4 (Central Limit Theorem (CLT))

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{D} N(0,1) \text{ when } n \to \infty$$

Z converges in distribution to $N(0,1)$ as $n \to \infty$

(converges in distribution: $P(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq a) \to \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-\frac{x^2}{2}} dx$)

♡

Proof 3.5

Prove the situation of $\mu = 0, \sigma^2 = 1$, we can use linear transformations to get other situations.

Moment-generating function(MGF) of X_i: $M_0(t) = E(e^{tX_i})$.

$$M_0(0) = 1, M_0'(0) = EX_i = 0, M_0''(0) = EX_i^2 = 1$$

Moment-generating function(MGF) of $\sqrt{n}\overline{X}$:

$$\begin{aligned} M_1(t) &= Ee^{t\sqrt{n}\overline{X}} = Ee^{t\frac{\sum_{i=1}^{n} X_i}{\sqrt{n}}} \\ &= Ee^{t\frac{X_1}{\sqrt{n}}} \cdot Ee^{t\frac{X_2}{\sqrt{n}}} \cdots Ee^{t\frac{X_n}{\sqrt{n}}} \\ &= [M_0(\frac{t}{\sqrt{n}})]^n \end{aligned}$$

$$\begin{aligned}
\lim_{n\to\infty} \log M_1(t) &= \lim_{n\to\infty} n \log M_0(\frac{t}{\sqrt{n}}) \\
&(\textit{let } y = \frac{1}{\sqrt{n}}) \\
&= \lim_{y=0} \frac{\log M_0(yt)}{y^2} \\
&(\textit{L'Hôpital's rule}) \\
&= \lim_{y=0} \frac{tM_0'(yt)}{2yM_0(yt)} \\
&(\textit{L'Hôpital's rule}) \\
&= \lim_{y=0} \frac{t^2 M_0''(yt)}{2M_0(yt) + 2ytM'(yt)} \\
&= \frac{t^2}{2}
\end{aligned}$$

As we know the Moment-generating function(MGF) of $Z \sim N(0,1)$ *is* $M_Z(t) = \frac{t^2}{2}$.

Hence, $M_1(t) = M_Z(t)$ *i.e.* $\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{D} N(0,1)$ *as* $n \to \infty$

Chapter 4 Distribution

4.1 Discrete

4.1.1 Bernoulli Distribution – $\text{Bernoulli}(\pi)$: an event happens with probability π

Assume n independent binary (taking values 0 or 1) observations arising from independent and identical trials: $y_1, y_2, \ldots, y_n$ such that: $P(Y_i = 1) = \pi$ and $P(Y_i = 0) = 1 - \pi$.

Random variables Y_i are normally called **Bernoulli** trials, $Y_i \sim \text{Bernoulli}(\pi)$.

$$\mathbb{E}(Y_i) = \pi, Var(Y_i) = \pi(1-\pi)$$

4.1.2 Binomial distribution – $bin(n, \pi)$: n independent Bernoulli distributions

The random variable $Y = \sum_{i=1}^{n} Y_i$ has the Binomial distribution with index n and parameter π denoted as $Y \sim \text{bin}(n, \pi)$. Mass probability function for Y :

$$P(y) = \begin{pmatrix} n \\ y \end{pmatrix} \pi^y (1-\pi)^{n-y}$$

with $\begin{pmatrix} n \\ y \end{pmatrix} = \frac{n!}{y!(n-y)!}$ and $y = 0, 1, 2, \ldots, n$

(1) Mean and Variance: $\mathbb{E}(Y) = \mu = n\pi, \text{Var}(Y) = \sigma^2 = n\pi(1-\pi)$

(2) Skewness: $\mathbb{E}\frac{(Y-\mu)^3}{\sigma^3} = \frac{1-2\pi}{\sqrt{n\pi(1-\pi)}}$

(3) If the independence assumption is violated, the Binomial distribution does not apply.

(4) Normal approximation: $\frac{Y-n\pi}{\sqrt{n\pi(1-\pi)}} \xrightarrow{d} N(0,1), \quad n \to \infty.$

4.1.3 Multinomial Distribution

Assume n independent trials have outcomes in $c > 2$ categories. Let $y_{ij} = 1$ if trial i has outcome in category j; otherwise $y_{ij} = 0$. For example, if $c = 5$, a possible outcome of a trail is $(0, 1, 0, 0, 0)$. The binary vector $\vec{y_i} = (y_{i1}, y_{i2}, ..., y_{ic})$ represents a multinomial trial.

Note that $\sum_j y_{ij} = 1$ whereas $\sum_i y_{ij} = n_j$ is the number of outcomes for category j. Also note that y_{ic} is redundant because it is dependent on the remaining outcomes: $y_{ic} = 1 - \sum_{j=1}^{c-1} y_{ij}$.

The vector of counts $(n_1, n_2, ..., n_c)$ has a multinomial distribution, with mass probability function:

$$p(n_1, n_2, ..., n_{c-1}) = \frac{n!}{n_1! n_2! ... n_c!} \pi_1^{n_1} \pi_2^{n_2} ... \pi_c^{n_c}$$

where $\pi_j = \Pr(Y_{ij} = 1)$

1. Please note that the marginal distribution of each n_j is a binomial distribution.
2. The binomial distribution is a special case of the multinomial distribution when $c = 2$.
3. $\mathbb{E}(n_j) = n\pi_j$, $\text{Var}(n_j) = n\pi_j(1-\pi_j)$, $\text{Cov}(n_j, n_k) = -n\pi_j\pi_k$.
4. Entries of correlation matrix:

$$\rho(n_j, n_j) = 1, \forall j;\ \rho(n_j, n_k) = \frac{\text{Cov}(n_j, n_k)}{\sqrt{\text{Var}(n_j)\text{Var}(n_k)}} = -\sqrt{\frac{\pi_j\pi_k}{(1-\pi_j)(1-\pi_k)}}, \forall j \neq k$$

4.1.4 Poisson Distribution – $Pois(\lambda)$: an event happens k times within unit time

λ: frequency of the event, i.e., the average number of event happens within unit time.

$$\Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!},\ k = 0, 1, 2, 3...$$

$$E(X) = Var(X) = \lambda$$

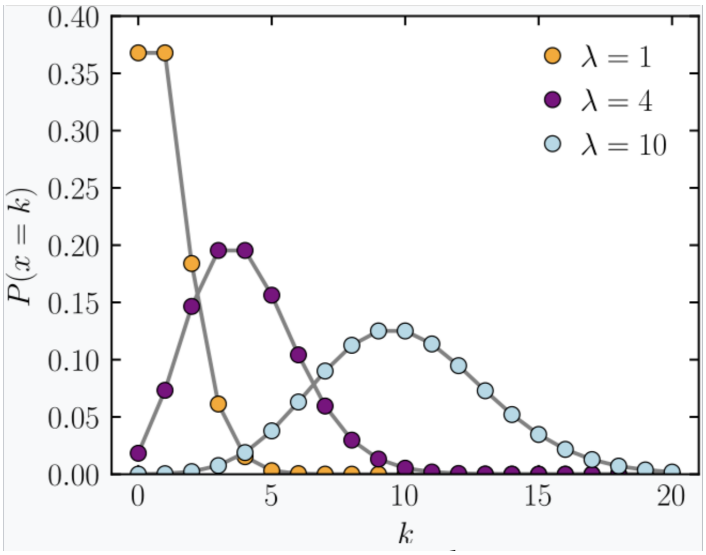

Figure 4.1: The Poisson pmf is unimodal with mode equal to the integer part of λ.

Its skewness (measure of the asymmetry of the probability distribution) is: $\mathbb{E}\frac{(X-\mu)^3}{\sigma^3} = \frac{1}{\sqrt{\lambda}}$. That is the higher the λ is, the less skew the distribution is.

Derivation process (an approximation to the binomial distribution when n is large and π is small, such that $\mu = n\pi$.)

Consider a unit time (the unit is divided into n equal subparts, $n \to \infty$), there is an event may occur with every subpart, the number of the event happens should follow binomial distribution $B(n, p)$. where $n \to \infty, p \to 0$; $\lambda = n \cdot p$ is the expected number of events in this period of time.

The probability the number of the event happens:

$$\begin{aligned}
\Pr(X = k) &= \lim_{n\to\infty} \binom{n}{k} (\frac{\lambda}{n})^k (1 - \frac{\lambda}{n})^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{(n-k)!k!} (\frac{\lambda}{n})^k (1 - \frac{\lambda}{n})^n (1 - \frac{\lambda}{n})^{-k} \\
&= \lim_{n\to\infty} \frac{n!}{(n-k)!k!} (\frac{\lambda}{n})^k e^{-\lambda} \\
&= \frac{\lambda^k e^{-\lambda}}{k!} \lim_{n\to\infty} \frac{n!}{(n-k)!n^k} \\
&= \frac{\lambda^k e^{-\lambda}}{k!} \lim_{n\to\infty} \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-k+1}{n} \\
&= \frac{\lambda^k e^{-\lambda}}{k!}
\end{aligned}$$

Sums of independent Poisson random variables are Poisson random variables

$X \sim Pois(\lambda_1), Y \sim Pois(\lambda_2)$ are two independent Poisson random variables, then $Z = X + Y$ also follow Poisson distribution, and the paramenter is the sum of X's and Y's

$$Z \sim Pois(\lambda_1 + \lambda_2)$$

Then,

$$Z = X_1 + X_2 + \cdots + X_n \sim Pois(\lambda_1 + \lambda_2 + \cdots + \lambda_n)$$

4.1.5 Connection between Poisson and multinomial distribution

Consider a sum of independent Poisson random variables Y_i with parameters λ_i. $\sum_i Y_i$ has a Poisson distribution with parameter $\lambda = \sum_i \lambda_i$. If $\sum_i Y_i = n$ and n is fixed, the random variables $Y_i \mid n$ are no longer independent nor have a Poisson distribution.

For a c number of Poisson random variables, we can calculate the joint probability distribution of a set of counts $\{n_i\}$ conditioned on $\sum_i Y_i = n$ as:

$$\begin{aligned}
P(Y_1 = n_1, Y_2 = n_2, ..., Y_c = n_c \mid \sum_i Y_i = n) &= \frac{P(Y_1 = n_1, Y_2 = n_2, ..., Y_c = n_c)}{P(\sum_i Y_i = n)} \\
&= \frac{\prod_{i=1}^c \frac{e^{-\lambda_i} \lambda_i^{n_i}}{n_i!}}{\frac{e^{-\sum_i \lambda_i} (\sum_i \lambda_i)^n}{n!}} = \frac{n!}{n_1! n_2! ... n_c!} \pi_1^{n_1} \pi_2^{n_2} ... \pi_c^{n_c}
\end{aligned}$$

where $\pi_i = \frac{\mu_i}{\sum_i \mu_i}$. This results in a multinomial $(n, \{\pi_i\})$ distribution.

4.1.6 Geometric distribution: $P(X = k) = (1-p)^{k-1}p$

The geometric distribution gives the probability that the first occurrence of success requires k independent trials, each with success probability p. It is a discrete form of exponential distribution.

$$P(X = k) = (1-p)^{k-1}p,\ P(X \leq k) = 1 - (1-p)^k$$
$$\mathbb{E}(X) = \frac{1}{p},\ Var(X) = \frac{1-p}{p^2}$$

4.2 Continuous

4.2.1 Exponential distribution $Exp(\lambda)$: interval between to independent identical event / the first time a event happened

λ: frequency of the event.

X follows exponential distribution with parameter λ or β:

$$X \sim \text{Exp}(\lambda) \text{ or } X \sim \text{Exp}(\beta)$$

They are equivlanet, the only difference is $\beta = \frac{1}{\lambda}$.

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

$$f(x;\beta) = \begin{cases} \frac{1}{\beta} e^{-\frac{1}{\beta}x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

c.d.f is:

$$F(x;\lambda) = \begin{cases} 1-e^{-\lambda x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

Note that $\lambda > 0$ is the frequcy of the occurance of the event; $\beta > 0$ is the probability of the event happens in each second. The range of exponential distribution is $[0, \infty)$.

$$\mathbb{E}(X) = \frac{1}{\lambda};\ Var(X) = \frac{1}{\lambda^2}$$

Memorylessness: $\Pr(T > s+t \mid T > s) = \Pr(T > t)$

$$\begin{aligned}\Pr(T > s+t \mid T > s) &= \frac{\Pr(T > s+t \text{ and } T > s)}{\Pr(T > s)} \\ &= \frac{\Pr(T > s+t)}{\Pr(T > s)} \\ &= \frac{e^{-\lambda(s+t)}}{e^{-\lambda s}} \\ &= e^{-\lambda t} \\ &= \Pr(T > t)\end{aligned}$$

Derivation process:

Consider a unit time (the unit is divided into n equal subparts, $n \to \infty$), there is an event may occur with every subpart, the number of the event happens should follow binomial distribution $B(n,p)$. where $n \to \infty, p \to 0$; $\lambda = n \cdot p$ is the expected number of events in this period of time. (the same as Poisson)

CDF:

$$1 - F(x;\lambda) = \lim_{n\to\infty}(1 - \frac{\lambda}{n})^{nx} = e^{-\lambda x} \Rightarrow F(x;\lambda) = 1 - e^{-\lambda x}$$

PDF:

$$f(x;\lambda) = \frac{\partial F(x;\lambda)}{\partial x} = \lambda e^{-\lambda x}$$

4.2.2 Gaussian/Normal Distribution

$N(\mu, \sigma^2)$. p.d.f. $f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}(\frac{x-\mu}{\sigma})^2}$

Theorem 4.1

Suppose $X \sim N(\mu_1, \sigma_1^2)$ and $Y \sim N(\mu_2, \sigma_2^2)$ are independent, then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma^2)$. ♡

Proof 4.1

The MGF of X is

$$\begin{aligned}M_X(t) = \mathbb{E}e^{tx} &= \int_{-\infty}^{\infty} e^{tx} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-\mu_1)^2}{2\sigma_1^2}} dx \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{x^2 - 2(\mu_1 + \sigma_1^2 t)x + \mu_1^2}{2\sigma_1^2}} dx \\ &= e^{\frac{\sigma_1^4 t^2 + 2\mu_1\sigma^2 t}{2\sigma_1^2}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}\sigma_1} e^{-\frac{(x-(\mu_1+\sigma_1^2 t))^2}{2\sigma_1^2}} dx \\ &= e^{t\mu_1 + \frac{1}{2}\sigma_1^2 t^2}\end{aligned}$$

Then, the MGF of $X+Y$ is

$$M_{X+Y}(t) = \mathbb{E}e^{t(X+Y)} = \mathbb{E}e^{tX}\mathbb{E}e^{tY} = e^{t(\mu_1+\mu_2)+\frac{1}{2}(\sigma_1^2+\sigma_2^2)t^2} = M_Z(t)$$

where $Z \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma^2)$

4.2.3 Multivariate/Joint Gaussian/Normal Distribution (MVN)

A k-dimensional random vector $(X_1, X_2, ..., X_k)^T = \mathbf{X} \sim N(\mu, \Sigma)$.

p.d.f.

$$f_{\mathbf{X}}(x_1, \ldots, x_k) = \frac{\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)}{\sqrt{(2\pi)^k|\boldsymbol{\Sigma}|}}$$

A random vector is said to be k-variate normally distributed if every linear combination of its k components has a univariate normal distribution.

(1) μ is a k-dimensional **mean vector**:

$$\mu = \mathbb{E}[\mathbf{X}] = (\mathbb{E}[X_1], \mathbb{E}[X_2], ..., \mathbb{E}[X_k])^T$$

(2) Σ is a $k \times k$ **covariance matrix**

$$\Sigma_{i,j} = \mathrm{E}[(X_i - \mu_i)(X_j - \mu_j)] = \mathrm{Cov}[X_i, X_j]$$

(3) The inverse of Σ, $\boldsymbol{Q} = \Sigma^{-1}$ is **precision matrix**.

Theorem 4.2

MVN distribution is completely specified by knowing μ, Σ. ♡

Proof 4.2

MGF: $M_X(t_1, t_2, ..., t_k) = \mathbb{E}e^{\sum_{i=1}^k t_i x_i}$

Since any linear combination of X is also normal distribution, $\Omega = \sum_{i=1}^k t_i x_i$ follows normal distribution.

$$M_X(t_1, t_2, ..., t_k) = \mathbb{E}e^{\Omega} = e^{\mathbb{E}(\Omega)+\frac{1}{2}Var(\Omega)} = e^{\sum_{i=1}^k t_i\mathbb{E}(x_i)+\frac{1}{2}Var(\sum_{i=1}^k t_i x_i)}$$

Generally, "independence" is a **stronger** condition than "0 correlation" ($Cov = 0$).

Theorem 4.3

*For **MVN**, "independence" is **equivalent** to "0 correlation"* ♡

Proof 4.3

As we show $M_X(t_1, t_2, ..., t_k) = e^{\sum_{i=1}^k t_i \mathbb{E}(x_i) + \frac{1}{2} Var(\sum_{i=1}^k t_i x_i)}$. *If* $Cov(x_i, x_j) = 0, \forall i, j \in S$,

$$\begin{aligned} M_X(t_1, t_2, ..., t_k) &= e^{\sum_{i=1}^k t_i \mathbb{E}(x_i) + \frac{1}{2} Var(\sum_{i=1}^k t_i x_i)} \\ &= e^{\sum_{i=1}^i t_k i \mathbb{E}(x_i) + \frac{1}{2} \sum_{i=1}^i t_i^2 Var(x_i)} \\ &= \prod_{i=1}^k e^{t_i \mathbb{E}(x_i) + \frac{1}{2} t_i^2 Var(x_i)} \\ &= \prod_{i=1}^k M_{x_i}(t_i) \end{aligned}$$

Theorem 4.4

Independent $X = [X_1, X_2, ..., X_n] \sim$ *MVN and* $Y = [Y_1, Y_2, ..., Y_m] \sim$ *MVN, then* $W = [X_1, X_2, ..., X_n, Y_1, Y_2, ..., Y_m] \sim$ *MVN.*

♡

Theorem 4.5

Independent $X = [X_1, X_2, ..., X_n] \sim N(\mu_1, \Sigma_1)$ *and* $Y = [Y_1, Y_2, ..., Y_n] \sim N(\mu_2, \Sigma_2)$, *then* $X + Y \sim N(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$.

♡

4.3 Poisson process: A sequence of arrivals in continuous time with rate λ

4.3.1 Definition

$N(t) \sim Pois(\lambda t)$: Number of arrivals in length t follows Poisson distribution

$$N(t) \sim Pois(\lambda t)$$

$$\Pr(N(t) = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$

The number of arrivals in disjoint time intervals are independent.

4.3.2 T_j: time of j^{th} arrival

$T_1 > t$ is same as $N(t) = 0$: $P(T_1 > t) = P(N(t) = 0) = e^{-\lambda t}$

$\Rightarrow T_1 \sim Expo(\lambda) \Rightarrow T_j - T_{j-1} \sim Expo(\lambda); T_j \sim Gamma(j, \lambda)$

4.3.3 Theorem (Conditional counts): $N(t_1)|N(t_2) = n \sim Bin(n, \frac{t_1}{t_2})$

(We can interpret the theorem as: n points distribute uniformly in $(0, t_2]$, so the probability a point loctae within $(0, t_1]$ is $\frac{t_1}{t_2}$)

Chapter 5 Markov Chain

5.1 Definition

For discrete state space S, a Markov Chain is a stochastic process $X_0, X_1, X_2, ...$ such that

$$P(X_{n+1} = i | X_n = j, X_{n-1} = x_{n-1}, ..., X_0 = x_0) = P(X_{n+1} = i | X_n = j)$$

for all $n \in \mathbb{Z}$ and $x_0, x_1, ..., x_{n-1}, i, j \in S$.

A MC is called time homogeneous if $P(X_{n+1} = i | X_n = j) = P(X_1 | X_0 = j), \forall n \in \mathbb{Z}^+$ and $i, j \in S$ (we only consider time homogeneous MC).

The **transition probabilities** for a time homogeneous MC can be written down as a matrix P satisfying $P_{ij} = P(X_1 = j | X_0 = i)$. This matrix P satisfies two properties:

(1) $P_{ij} \geq 0$ for all $i, j \in S$.

(2) $\sum_{j \in S} P_{ij} = 1$ for all $i \in S$.

Any matrix satisfies the two properties is called a **stochastic matrix**.

5.2 Matrix Computations

Given a time homogeneous MC with initial distribution $X_0 \sim \alpha \in [0, 1]^{|S|}$ and transition matrix P.

Lemma 5.1 (Distribution of Entire Sequence)

$$P(X_0 = x_0, X_1 = x_1, ..., X_n = x_n) = P(X_0 = x_0) P_{x_0, x_1} P_{x_1, x_2} ... P_{x_{n-1}, x_n}$$

♡

Lemma 5.2 (Markov Property)

$$P(X_{t_n} = x_{t_n} | X_{t_{n-1}} = x_{t_{n-1}}, ..., X_{t_0} = x_{t_0}) = P(X_{t_n} = x_{t_n} | X_{t_{n-1}} = x_{t_{n-1}})$$

♡

Lemma 5.3 (Transition Probability after n states)

$$P(X_n = j | X_0 = i) = (P^n)_{ij}$$

♡

Proof 5.1

$P(X_2 = j | X_0 = i) = \sum_{k \in S} P(X_2 = j | X_1 = k) P(X_1 = k | X_0 = i) = \sum_{k \in S} P_{kj} P_{ik} = (P^2)_{ij}$.

Then prove by mathematical induction, $P(X_n = j | X_0 = i) = \sum_{k \in S} P(X_n = j | X_{n-1} = k) P(X_{n-1} = k | X_0 = i) = \cdots = (P^n)_{ij}$

5.2.1 Chapman Kolmogorov Equations (C-K Equations)

$P(X_{n+m} = j|X_0 = i) = (P^{m+n})_{ij} = \sum_{k\in S}(P^m)_{ik}(P^n)_{kj}$

$m-$step transition probabilities from state k to state j:

$$P(X_{n+m} = j|X_0 = i) = (P^{m+n})_{ij} = \sum_{k\in S}(P^m)_{ik}(P^n)_{kj} \tag{5.1}$$

Proof 5.2

$P(X_n = j|X_0 = i) = \sum_{k\in S} P(X_{n+m} = j|X_m = k)P(X_m = k|X_0 = i) = \sum_{k\in S} P(X_n = j|X_0 = k)P(X_m = k|X_0 = i)$

5.2.2 Marginal Distribution $P(X_n = j) = (\alpha P^n)_j$

Lemma 5.4 (Marginal Distribution)

Given initial distribution $X_0 \sim \alpha$ and transition matrix P. α is distribution vector $(1 \times |S|)$ with $\sum_{i\in S} \alpha_i = 1$.

$$P(X_n = j) = (\alpha P^n)_j$$

♡

Corollary 5.1 (Distribution of Subsequence)

$$P(X_{t_n} = x_{t_n}, X_{t_{n-1}} = x_{t_{n-1}}, ..., X_{t_0} = x_{t_0}) = (\alpha P^{t_0})_{x_{t_0}} P^{t_1-t_0}_{x_{t_0},x_{t_1}} P^{t_2-t_1}_{x_{t_1},x_{t_2}} \cdots P^{t_n-t_{n-1}}_{x_{t_{n-1}},x_{t_n}}$$

♡

5.3 States, Class

5.3.1 Irreducible, Reducible

- Accessible: j is accessible from i if $\exists n$ s.t. $P^n_{ij} > 0$.
- Communicate/Communication: i communicates j ($i \leftrightarrow j$) if j is accessible from i and i is accessible from j. (Reflexivity: $i \leftrightarrow i$; Symmetry: $i \leftrightarrow j \Rightarrow j \leftrightarrow i$; Transitivity: $i \leftrightarrow j$ and $j \leftrightarrow k \Rightarrow i \leftrightarrow k$.)
- (Communication) Class: if $i \leftrightarrow j$, then states i, j are said to be in the same (communication) class. (Since communication is an equivalence relation, the state space can be partitioned into equivalence classes, called *communication classes*.)
- Irreducible: A Markov Chain that has only one class is said to be irreducible.

5.3.2 Recurrent, Transient

- Recurrent State: State i is recurrent if $f_i = P(\text{ever re-enter state } i \text{ if started in state } i) = 1$. (the expected number of times it visits state i is $\sum_{n=0}^{\infty} P_{ii}^n = +\infty$). (A MC is irreducible if all states are recurrent)
- Transient State: State i is transient if $f_i = P(\text{ever re-enter state } i \text{ if started in state } i) < 1$. (the expected number of times it visits state i is $\sum_{n=0}^{\infty} P_{ii}^n < +\infty$; $P(\text{visits state } i \text{ exactly } n \text{ times}) = f_i^{n-1}(1-f_i)$; The expected number is $\sum_{n=0}^{\infty} f_i^n(1-f_i)n = n = 0^{\infty} P_{ii}^n < \infty$).
- Transient Class: A communicating class is called transient if starting from that class, with probability 1 the MC leaves that class and never returns. The states of such a class are called transient states.
- Recurrent Class: communicating class that is not transient.

Lemma 5.5

If i is recurrent, $i \leftrightarrow j \Rightarrow j$ is recurrent.

♡

Theorem 5.1

The states of a communication class are either all recurrent or all transient.

♡

Corollary 5.2

For a finite irreducible Markov chain, all states are recurrent.

♡

Canonical Decomposition

Definition 5.1

A set of states C is said to be closed if no state outside of C is accessible from any state in C. If C is closed, then

$$P_{ij} = 0,\ \forall i \in C, j \notin C$$

♣

Lemma 5.6

(1) A communication class is closed if it consists of all recurrent states. (2) A finite communication class is closed only if it consist of all recurrent states.

♡

Proof 5.3

(1): if not closed, $\exists i \in C, j \notin C, P_{ij} > 0$. i shouldn't be accessible from j since i, j are not in one class. There exists positive probability that starting from i then hit j and never hit i again, which contradicts to i is recurrent. (2):According to former corollary, a finite class's all states are recurrent.

5.4 Periodicity

Suppose P is the transition matrix for an irreducible MC. For a given state i, we define the set

$$J_i = \{n \geq 1 : P^n(i,i) > 0\}$$

J_i is the set of times when it is possible for the MC to come back to i starting from i at time 0. We define the **period** of a state i is

$$d(i) = gcd(J_i)$$

5.4.1 Lemma: all states in an irreducible MC have the same period

Lemma 5.7

For an irreducible MC, all states have the same period. ♡

Proof 5.4

Let d be a common divisor of J_i. Consider any other state j. We want to show d is also the common divisor of J_j.

Since the MC is irreducible, there exists m and n s.t. $P_{ij}^m > 0$ and $P_{ji}^n > 0$. Then $P_{ii}^{m+n} \geq P_{ij}^m P_{ji}^n > 0 \Rightarrow m+n \in J_i$. d should be a divisor of $m+n$.

For any $l \in J_j$, $P_{ii}^{m+n+l} \geq P_{ij}^m P_{jj}^l P_{ji}^n > 0 \Rightarrow m+n+l \in J_i$. d divides $m+n+l \Rightarrow d$ divides l. Since l can be any number in J_j, d is a common divisor of J_j.

5.4.2 Periodic, Aperiodic

A state is **aperiodic** if period equals 1, **periodic** otherwise.

A chain is **aperiodic** if all its states are aperiodic, **periodic** otherwise.

5.5 Regular Matrix

5.5.1 Regular matrix: $\exists n \geq 1$ s.t. $P^n > 0$

A matrix M is said to be positive if all the entries of M are positive. We write $M > 0$.

Definition 5.2 (Regular Transition Matrix)

A transition matrix P is said to be regular if some power of P is positive. That is, $P^n > 0$, for some $n \geq 1$. ♣

5.5.2 Lemma: Finite MC is Irreducible, Aperiodic $\Leftrightarrow$ has Regular transition matrix

Lemma 5.8

*A finite MC is **irreducible** and **aperiodic** is equivalent to the transition matrix P is **regular**.*

♡

We also call an MC is **ergodic** if it is **irreducible** and **aperiodic**.

5.6 Eigenvalues of a Stochastic Matrix: $\lambda = 1$ must exist and other $|\lambda| \leq 1$ (not equal when if regular matrix)

Lemma 5.9

A stochastic matrix $\boldsymbol{P}$ has an eigenvalue $\lambda^* = 1$. All other eigenvalues λ of $\boldsymbol{P}$ are such that $|\lambda| \leq 1$.

If $\boldsymbol{P}$ is a regular matrix, then the inequality is strict. That is, $|\lambda| < 1$ for all $\lambda \neq \lambda^*$.

♡

5.7 Long Run Behavior of Finite Markov Chains

As $n \to \infty$, P^n:

(1) Convergence. ($P^{n+1} = P^n$)

(2) Forgetting the initial states. (each row is identical)

5.7.1 Limiting Distribution

Definition 5.3

*A MC is said to have a **limiting distribution** λ if we have*

$$\lim_{n\to\infty} P_{ij}^n = \lambda_j, \ \forall i, j \in S$$

An equivalent definition is that for all initial distributions $X_0 \sim \alpha$ and all $j \in S$ we have

$$\lim_{n\to\infty} (\alpha P^n)_j = \lambda_j$$

♣

Example: $P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$. If $p + q = 1$, each rows of P is the same and $P^n = P$.

Assume $p+q \neq 1$,

$$\begin{aligned} P_{11}^n &= P_{11}^{n-1}(1-p) + P_{12}^{n-1}q \\ &= P_{11}^{n-1}(1-p) + (1-P_{11}^{n-1})q \\ P_{11}^n &= \frac{q}{p+q} + \frac{p}{p+q}(1-p-q)^n \to \frac{q}{p+q} \text{ as } n \to \infty \end{aligned}$$

$$\lim_{n\to\infty} P^n = \frac{1}{p+q}\begin{bmatrix} q & p \\ q & p \end{bmatrix}$$

Lemma 5.10

If λ is the limiting distribution for a MC with transition matrix P then λ satisfies the equation

$$\lambda P = \lambda$$

♡

Proof 5.5

$(\lambda P)_j = \sum_{i\in S} \lambda_i P_{ij} = \sum_{i\in S} \lim_{n\to\infty} P_{ki}^n P_{ij} = \lim_{n\to\infty} \sum_{i\in S} P_{ki}^n P_{ij} = \lim_{n\to\infty} P_{kj}^{n+1} = \lambda_j$

5.7.2 Stationary Distribution

Definition 5.4

A distribution π which satisfies the equation

$$\pi P = \pi$$

*is called a **stationary distribution** for the MC.*

♣

Note: A limiting distribution λ for the MC has to also be a stationary distribution. The converse is not always true.

5.7.3 Limiting Distribution = Expected Proportion of time in each state

The entries of the limiting distribution can also be interpreted as **the limit of the expected proportion of time the MC spends in each of the corresponding states.** For any state j, define the indicator random variable $I_k = 1(X_k = j)$. Now define

$$F_{n,j} = \frac{1}{n}\sum_{k=0}^{n-1} I_k$$

The random variable $F_{n,j}$ represents the proportion of time till time $n-1$ the MC spends in state j.

Lemma 5.11

If λ is the limiting distribution for a MC with transition matrix P then λ satisfies the equation $\lim_{n\to\infty} \mathbb{E}(F_{n,j}|X_0 = i) = \lambda_j$ *for all* $j, i \in S$

♡

Proof 5.6

We can write

$$\mathbb{E}(F_{n,j}|X_0 = i) = \mathbb{E}\frac{1}{n}\sum_{k=0}^{n-1}\mathbb{E}(I_k|X_0 = i) = \frac{1}{n}\sum_{k=0}^{n-1}P(X_k = j|X_0 = i) = \frac{1}{n}\sum_{k=0}^{n-1}P_{ij}^k$$

Therefore, taking limits we can conclude that

$$\lim_{n\to\infty}\mathbb{E}(F_{n,j}|X_0 = i) = \lim_{n\to\infty}\frac{1}{n}\sum_{k=0}^{n-1}P_{ij}^k = \lim_{n\to\infty}P_{ij}^n = \lambda_j$$

5.7.4 Fundamental Theorem for Irreducible, Aperiodic, Finite MC (Regular transition matrix) $\Rightarrow \exists$ unique limiting distribution π and $\pi_j > 0, \forall j$

Theorem 5.2

If P is the transition matrix for an irreducible, aperiodic (finite) Markov chain then there exists a unique stationary distribution or a unique solution to the equation $\pi = \pi P$ which satisfies the following two properties:

*(1) π is the **limiting distribution** of the MC. ($\lim_{n\to\infty}\alpha P^n = \pi, \forall\alpha$ initial distribution)*

*(2) π gives **positive** probability to each of the states. ($\pi_j > 0, \forall j \in S$)*

♡

5.7.5 Long run behavior for reducible and/or periodic chains

Question: What is the long run behavior for reducible and/or periodic chains?

Assume P is reducible with recurrent classes $R_1, \ldots, R_r$ and transient classes $T_1, \ldots, T_s$. Each recurrent class acts as a separate MC with transition matrix $P_1, \ldots, P_r$. Assume each P_k is aperiodic. Then by the fundamental theorem, there exists r different limiting distributions $\pi^1, \ldots, \pi^r$. The distribution π^k is supported on its own recurrent class; i.e. $\pi^k(j) = 0$ if $j \notin R_k$. There are three cases to consider:

1. If $i, j \in R_k$ (in the same recurrent class) then
$$\lim_{n\to\infty}P_{ij}^n = \pi^k(j)$$
2. **If i is any transient state then eventually it ends up in one of the recurrent states.** Therefore, if i, j are transient states then,
$$\lim_{n\to\infty}P_{ij}^n = 0$$
3. Let $\alpha_k(i)$ for $k = 1, ..., r$ be the probability that the chain starting in i eventually ends up in a recurrent class R_k. (We will see later how to calculate $\alpha_k(i)$.) Once the chain reaches the recurrent class R_k, it will settle down to the limiting distribution on R_k. Therefore, we have for a transient state i and $j \in R_k$,
$$\lim_{n\to\infty}P_{ij}^n = \alpha_k(i)\pi^k(j)$$

So, in this case there is a limit of P^n, but the limit will have different rows.

When an MC is irreducible but **periodic** (period $d > 1$), we can show there is no **limiting distribution**. P^n will keep switching according to whether $n|d$ has remainder $0, 1, ..., d-1$. Therefore, there cannot be a limit of P^n.

Although $\lim_{n\to\infty} P_{ij}^n$ doesn't exist in irreducible and periodic MC, $\lim_{n\to\infty} \frac{1}{n}\sum_{m=0}^{n-1} P_{ij}^m$ exists. It is the limit of the expected long run proportions of time spent in each state.

5.7.6 Fundamental Theorem for Irreducible, Finite MC: expected first return time $\mathbb{E}(T_j|X_0 = j) = \frac{1}{\pi_j}$

$$T_j = \min\{n > 0 : X_n = j\}$$

is the first time the chain returns to state i after time 0. This time is often also called the first passage time to the state i.

In a finite irreducible MC, $P(T_j < \infty) = 1, \forall i$.

Theorem 5.3

*Assume that $X_0, X_1, ...$ is a finite irreducible Markov chain. For each state j, let $\mu_j = \mathbb{E}(T_j|X_0 = j)$ be the expected return time to j. Then, μ_j is finite, and there exists a unique **positive** stationary distribution π such that*

$$\pi_j = \frac{1}{\mu_j}, \forall j$$

Furthermore, for all initial states i, limiting distribution on j equals to the expected proportion of time spends in j:

$$\pi_j = \lim_{n\to\infty} \frac{1}{n}\sum_{m=0}^{n-1} P_{ij}^m = \frac{1}{\mu_j}, \forall j$$

♡

Proof 5.7

The sum of k i.i.d. random variables $T_1 + T_2 + \cdots + T_k$ each of which follows the same distribution as T conditional on $X_0 = i$. For $k \to \infty$, by the Law of Large Numbers, $\lim_{k\to\infty} \frac{T_1+T_2+\cdots+T_k}{k} = \mathbb{E}(T|X_0 = i)$.

Consider this total time is $T_1 + T_2 + \cdots + T_k$ and the time we spent at i is k, the expected proportion of time the chain spends in state i is approximately $\lim_{k\to\infty} \frac{k}{T_1+T_2+\cdots+T_k} \approx \frac{1}{\mathbb{E}(T|X_0=i)}$. As we showed before, the expected proportion of time is $\pi_i \Rightarrow \pi_i = \frac{1}{\mathbb{E}(T|X_0=i)} = \frac{1}{\mu_i}, \forall i$

Example 5.1Two State MC Consider the transition matrix

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

Here, by the theorem

$$\mu_0 = \mathbb{E}[T_0|X_0 = 0] = \frac{1}{\pi(0)} = \frac{p+q}{q}$$

5.8 Return Times and Absorption Probabilities

5.8.1 Expected Number of Visits to a Transient State: $E(Y_i|X_0 = j) = M_{ji} = (I - Q)^{-1}_{ji}$

Let P be the transition matrix of a MC. Suppose P has some transient states and let Q be the submatrix of P which contains the rows and columns for the transient states. Hence, after reordering the states we can write

$$P = \begin{bmatrix} \tilde{P} & 0 \\ S & Q \end{bmatrix}$$

Let i be a transient state and let us define **a random variable which counts the total number of visits to the state** i

$$Y_i = \sum_{n=0}^{\infty} \mathbf{1}_{X_n=i}$$

Since i is transient, $Y_i < \infty$ w.p.1.

Lemma 5.12

Let Q denote the part of transition matrix indexed by the transient states. Define $M = (I - Q)^{-1}$. We have the following equality for any two transient states $i, j \in S$,

$$E(Y_i|X_0 = j) = M_{ji}$$

Thus, the matrix $(I - Q)^{-1}$ gives the expected number of visits to a transient state i when the MC starts at a transient state j.

♡

Proof 5.8

We can write

$$\mathbb{E}(Y_i|X_0 = j) = \sum_{n=0}^{\infty} P(X_n = i|X_0 = j) = \sum_{n=0}^{\infty} P^n_{ji} = \sum_{n=0}^{\infty} Q^n_{ji} = M_{ji}$$

The last equality holds because $I + Q + Q^2 + \cdots = \frac{I(I-Q^{\infty})}{1-Q} = (I - Q)^{-1}$

We can also extend the equation:

$$\mathbb{E}(Y_i|X_0 = j) = \mathbf{1}_{i=j} + \sum_{k \text{ transient}} \mathbb{E}(Y_i|X_1 = k)Q_{jk}$$

5.8.2 Expected Time till Absorption to a Recurrent Class:

$\mathbb{E}(T_{abs}|X_0 = j) = \sum_{i\in T_1\cup T_2\cup\cdots\cup T_s} M_{ji}$

Let's define

$$T_{abs} = \{\min_{n\geq 0} : X_n \in \text{a recurrent class}\}$$

which is **the waiting time till the chain enters a recurrent class**. T_{abs} also equals to the total time spent on transient states.

$$T_{abs} = \sum_{i\in T_1\cup T_2\cup\cdots\cup T_s} Y_i$$

Corollary 5.3

For any transient state $j \in S$,

$$\mathbb{E}(T_{abs}|X_0 = j) = \sum_{i\in T_1\cup T_2\cup\cdots\cup T_s} M_{ji}$$

♡

Example 5.2 Simple Random Walk (SRW) with absorbing boundaries on $\{0, 1, 2, 3, 4\}$

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 & 0 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

We can reorder it by $\{0, 4, 1, 2, 3\}$

$$P = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 1/2 & 0 & 0 & 1/2 & 0 \\ 0 & 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 & 1/2 & 0 \end{pmatrix} = \begin{bmatrix} I_{2\times 2} & 0 \\ S & Q \end{bmatrix}$$

where $Q = \begin{bmatrix} 0 & 1/2 & 0 \\ 1/2 & 0 & 1/2 \\ 0 & 1/2 & 0 \end{bmatrix}$, then

$$M = (I - Q)^{-1} = \begin{bmatrix} 3/2 & 1 & 1/2 \\ 1 & 2 & 1 \\ 1/2 & 1 & 3/2 \end{bmatrix}$$

Therefore $\mathbb{E}(Y_3|X_0=1)=\frac{1}{2}$, $\mathbb{E}(T_{abs}|X_0=1)=M_{11}+M_{12}+M_{13}=\frac{3}{2}+1+\frac{1}{2}=3$.

5.8.3 Expected first return time (different initial state) = Time till Absorption

We have computed $\mathbb{E}[T_i|X_0=i]=\frac{1}{\pi_i}$, we want to compute

$$\mathbb{E}[T_i|X_0=j], i\neq j$$

Method 1: Condition on first step: Let $a_j=\mathbb{E}[T_i|X_0=j]$

$$\begin{aligned}\mathbb{E}[T_i|X_0=j]&=P_{ji}\cdot 1+\sum_{k\neq i}P_{jk}\cdot(1+\mathbb{E}[T_i|X_0=k])\\&=1+\sum_{k\neq i}P_{jk}\cdot\mathbb{E}[T_i|X_0=k]\\\Rightarrow a_j&=1+\sum_{k\neq i}P_{jk}\cdot a_k\end{aligned}$$

Then the problem can be solved by solving the linear system for all $j\in S$.

Method 2: This problem can be transformed into computing the **expected time till absorption to** i. (we can let i be an absorbing state)

Reorder the transition matrix P with i being the first state and make i an absorbing state

$$P=\begin{bmatrix}P_{ii} & R\\ S & Q\end{bmatrix}\Rightarrow\tilde{P}=\begin{bmatrix}1 & 0\\ S & Q\end{bmatrix}$$

Then

$$\mathbb{E}[T_i|X_0=j]=\mathbb{E}[T_{abs}|X_0=j]$$

Example 5.3 Simple Random Walk (SRW) with reflecting boundaries on $\{0,1,2,3,4\}$

$$P=\begin{pmatrix}0 & 1 & 0 & 0 & 0\\ 1/2 & 0 & 1/2 & 0 & 0\\ 0 & 1/2 & 0 & 1/2 & 0\\ 0 & 0 & 1/2 & 0 & 1/2\\ 0 & 0 & 0 & 1 & 0\end{pmatrix}$$

To compute $\mathbb{E}[T_0|X_0=j]$, we make 0 an absorbing state:

$$P=\begin{pmatrix}1 & 0 & 0 & 0 & 0\\ 1/2 & 0 & 1/2 & 0 & 0\\ 0 & 1/2 & 0 & 1/2 & 0\\ 0 & 0 & 1/2 & 0 & 1/2\\ 0 & 0 & 0 & 1 & 0\end{pmatrix}=\begin{bmatrix}1 & 0\\ S & Q\end{bmatrix}$$

where $Q = \begin{pmatrix} 0 & 1/2 & 0 & 0 \\ 1/2 & 0 & 1/2 & 0 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 0 & 1 & 0 \end{pmatrix}$, then we can calculate

$$M = (I - Q)^{-1} = \begin{pmatrix} 2 & 2 & 2 & 1 \\ 2 & 4 & 4 & 2 \\ 2 & 4 & 6 & 3 \\ 2 & 4 & 6 & 4 \end{pmatrix}$$

Now we can compute

$$\mathbb{E}[T_0|X_0 = 4] = M_{41} + M_{42} + M_{43} + M_{44} = 16$$

5.8.4 Probability of Eventually Entering a Given Recurrent Class: $A = (I - Q)^{-1}S = MS$

In some MC, there are more than one recurrent class. e.g. $\{0\}.\{N\}$ in absorbing boundary exmaple. We want to know **what is the probability that the MC eventually ends up in a given recurrent class starting from a transient state j.**

We can create a modified MC where each of the recurrent classes are seen as single states. Let these states be $r_1, ..., r_k$ with $P(r_i, r_i) = 1, \forall i \in \{1, ..., k\}$.

We denote all transient states as $t_1, ..., t_s$. And the transistion matrix is expressed by

$$P = \begin{bmatrix} I & 0 \\ S & Q \end{bmatrix}$$

Let α_{t_i,r_j} be the probability that the MC strating at t_i ends up at r_j. We set $\alpha_{r_i,r_i} = 1$ and $\alpha_{r_i,r_j} = 0, i \neq j$. Then, for any t_i, we can write by conditioning on the first step

$$\begin{aligned} \alpha_{t_i,r_j} &= P(X_n = r_j \text{ eventually } |X_0 = t_i) \\ &= \sum_{x \in S} P(X_1 = x|X_0 = t_i) P(X_n = r_j \text{ eventually } |X_1 = x) \\ &= \sum_{x \in S} P(t_i, x)\alpha_{x,r_j} \end{aligned}$$

(this S is the set of states.) Let $A_{s \times k}$ be the matrix with α_{t_i,r_j} being entries. The above equation can be written as

$$A = [S\ Q] \begin{bmatrix} I \\ A \end{bmatrix} = S + QA$$

$$\Rightarrow A = (I - Q)^{-1}S = MS$$

(this S is the submartix in P)

5.9 Examples of Finite MC

5.9.1 Gambler's Ruin

Example 5.4Gambler's Ruin Consider the asymmetric Gambler's Ruin with winning probability $p \in (0,1)$. The state space is $\{0,1,...,N\}$.

Let α_j be the probability that the MC get absorbed in state N stratinhg from state j. Clearly, $\alpha(0)=0, \alpha(N)=1$. For any $0<j<N$, we can condition on the first step to get

$$\begin{aligned}
\alpha(j) &= (1-p)\alpha(j-1) + p\alpha(j+1) \\
\Rightarrow \alpha(j+1) - \alpha(j) &= \frac{1-p}{p}(\alpha(j) - \alpha(j-1)) \\
\Rightarrow 1 = \alpha(N) - \alpha(0) &= \sum_{j=0}^{N-1} (\alpha(j+1) - \alpha(j)) \\
&= \sum_{k=0}^{N-1} \left(\frac{1-p}{p}\right)^k (\alpha(1) - \alpha(0)) \\
&= \begin{cases} N\alpha(1), & p = 0.5 \\ \frac{1-\left(\frac{1-p}{p}\right)^N}{1-\left(\frac{1-p}{p}\right)}\alpha(1), & p \neq 0.5 \end{cases}
\end{aligned}$$

$\alpha(1) = \alpha(1) - \alpha(0) = \begin{cases} \frac{1}{N}, & p = 0.5 \\ \frac{1-\left(\frac{1-p}{p}\right)}{1-\left(\frac{1-p}{p}\right)^N}, & p \neq 0.5 \end{cases}$. Then,

$$\alpha(j) = \sum_{k=0}^{j-1} \left(\frac{1-p}{p}\right)^k (\alpha(1) - \alpha(0)) = \begin{cases} \frac{j}{N}, & p = 0.5 \\ \frac{1-\left(\frac{1-p}{p}\right)^j}{1-\left(\frac{1-p}{p}\right)^N}, & p \neq 0.5 \end{cases}$$

5.9.2 Simple Random Walk (SRW) on Undirected Graph

Consider an undirected graph (V,E). The state space is V. Let the degree $deg(i)$ of a vertex i be the number of edges starting from i. Formally, we can write $deg(i) = \{j \in V : (i,j) \in E\}$. The transition matrix $P_{|V|\times|V|}$ is as follows.

$$P_{ij} = \frac{1}{deg(i)} \mathbf{1}_{(i,j)\in E}$$

The MC is irreducible iff the graph is connected. When assuming connected we can compute the unique

stationary distribution

$$\pi(v) = \frac{deg(v)}{2|E|} = \frac{deg(v)}{\sum_{v \in V} deg(v)}$$

The period of the chain is either 1 or 2. The period is 2 if and only if the graph is bipartite, meaning that the set of vertices can be divided into two subsets and each edge in the graph goes from one subset to another.

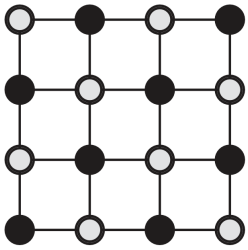

Figure 5.1: bipartite

If the period is 1 then π is the limiting distribution for this chain. If the period is 2 then π can still be interpreted as the limiting expected fraction of time spent in each of the states.

Chapter 6 Countably infinite MC

Countably infinite MC: Markov Chain in countable infinite state space (e.g. $\mathbb{Z}$). The transition matrix P is infinite large, but the sum of each row converges to 1.

Chapman Kolmogorov Equations (C-K Equations) also holds:

$P(X_{n+m} = j|X_0 = i) = (P^{m+n})_{ij} = \sum_{k\in S}(P^m)_{ik}(P^n)_{kj}$

Example:

(1) RW with partially reflecting boundary $S = \{0, 1, 2, ...\}$, $P_{x,x-1} = 1 - p, P_{x,x+1} = p, P_{0,1} = p, P_{0,0} = 1 - p$.

(2) Queuing Model: $X_n =$ # people at time n. $S = \{0, 1, 2, ...\}$. $P(x, x - 1) = q(1 - p)$; $P(x, x + 1) = (1 - q)p$; $P(x, x) = pq + (1 - p)(1 - q)$; $P(0, 0) = 1 - p$; $P(0, 1) = p$.

Difference: For the infinite, irreducible, and aperiodic MC, there may not exist stationary distribution.

Example 6.1 For Simple Random Walk: assume there exists a stationary distribution π, we have

$$\pi_j = \frac{1}{2}(\pi_{j-1} + \pi_{j+1}) \Rightarrow \pi_j - \pi_{j-1} = \pi_{j+1} - \pi_j$$

Let the difference between $\pi_j - \pi_{j-1} = \varepsilon$, there doesn't exist solution to

$$\sum_{i=0}^{\infty} \pi_i = 1; \pi_i = i\varepsilon, i = 0, 1, ...$$

6.1 Recurrence and Transience

6.1.1 Recurrent or Transient State

Suppose the first return time $T_j = \min\{n > 0 : X_n = j\}$.

Let the probability of the chain return to j given $X_0 = j$ is

$$f_j = P(T_j < \infty | X_0 = j)$$

Definition 6.1

*A state j is **recurrent** if $f_j = 1$ and **transient** if $f_j < 1$.* ♣

6.1.2 Recurrent or Transient Class

(Also class properties: states of a class should be all recurrent or all transient)

Lemma 6.1

If i, j are in the same class, i is recurrent $\Leftrightarrow$ j is recurrent. ♡

Proof 6.1

Suppose i is recurrent, $P(T_i < \infty | X_0 = i) = 1$. Since $i \sim j$, $\exists k > 0, P_{ij}^k > 0$.

Suppose $P(T_j < \infty | X_0 = j) < 1$ i.e., $P(T_j = \infty | X_0 = j) > 0$. Then,

$$\begin{aligned} P(T_i = \infty | X_0 = i) &\geq P(T_i = \infty | X_0 = j) P_{ij}^k \\ &= P(T_i = \infty | T_j = \infty, X_0 = j) P(T_j = \infty | X_0 = j) P_{ij}^k > 0 \end{aligned}$$

6.1.3 Lemma: Transient Class $\Leftrightarrow \sum_{n=0}^{\infty} P_{i,i}^n < \infty$

Lemma 6.2

*An irreducible MC is **transient** if and only if the expected number of visits to a state is finite; i.e.*

$$\sum_{n=0}^{\infty} P_{i,i}^n < \infty$$

♡

Proof 6.2

Let the total number of visits i in infinite time is $Y_i = \sum_{n=0}^{\infty} \mathbf{1}_{X_n = i}$. The expected number is $\mathbb{E}[Y_i | X_0 = i] = \sum_{n=0}^{\infty} P_{i,i}^n$.

$\Leftarrow$: If i is recurrent, the expected total number to visits i in infinite time should be infinite. Then, the MC can be proved to be transient if $\mathbb{E}[Y_i | X_0 = i] = \sum_{n=0}^{\infty} P_{i,i}^n < \infty$.

$\Rightarrow$: Suppose i is transient, let $f_i = P(T_i = \infty | X_0 = i) = q > 0$ (Probability of not return). Then, the expected number of returns to i is (follows geometric distribution)

$$\sum_{n=0}^{\infty} (1-q)^n q n = q(1-q) \frac{\partial \left(-\sum_{n=0}^{\infty} (1-q)^n\right)}{\partial q} = q(1-q) \frac{\partial \left(-\frac{1}{q}\right)}{\partial q} = \frac{1-q}{q}$$

which also equals to $\mathbb{E}[Y_i | X_0 = i] - 1 \Rightarrow \mathbb{E}[Y_i | X_0 = i] = \frac{1}{q} < \infty$

6.1.4 Recurrence/Transience of Simple Random Walk on Lattice

Is the d dimensional SRW recurrent or transient?

We can first consider $d = 1$ case. We want to compute the probability of returning to state 0 (the same as

others). For $2n$ steps trajectories, there are $\binom{2n}{n}$ trajectories that can return to 0 and each has probability $\frac{1}{2^{2n}}$.

$$P_{0,0}^{2n} = \binom{2n}{n} \frac{1}{2^{2n}} = \frac{(2n)!}{n!n!2^{2n}}$$

Using Stirling's formula: $n! \sim \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$ that is $\lim_{n\to\infty} \frac{n!}{\sqrt{2\pi n}\left(\frac{n}{e}\right)^n} = 1$

$$P_{0,0}^{2n} \sim \frac{2\sqrt{\pi n}\left(\frac{2n}{e}\right)^{2n}}{2\pi n\left(\frac{n}{e}\right)^{2n} 2^{2n}} = \frac{1}{\sqrt{\pi n}}$$

So the $\sum_{n=N}^{\infty} P_{0,0}^{2n} = \sum_{n=N}^{\infty} \frac{1}{\sqrt{\pi}} n^{-\frac{1}{2}} = \infty$.

Note: $n^{-\alpha}$ diverges when $\alpha \in (0, 1]$ and converges when $\alpha > 1$.

For d dimensions,

$$P_{0,0}^{2n} \sim n^{-\frac{d}{2}}$$

Lemma 6.3

SRW is recurrent when $d = 1, 2$; SRW is transient when $d \geq 3$. ♡

6.1.5 Null and Positive Recurrence

$\mu_j = \mathbb{E}[T_j | X_0 = j]$

Definition 6.2

*A state j is **positive recurrent** if it is recurrent and $\mu_j < \infty$. A state j is **null recurrent** if it is recurrent and $\mu_j = \infty$.* ♣

Example of null recurrent: $P(T_i = n) = \frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}, n \geq 1$.

$$f_i = \sum_{n=1}^{\infty} P(T_i = n) = \sum_{n=1}^{\infty} \left(\frac{1}{n} - \frac{1}{n+1}\right) \Rightarrow \text{ recurrent}$$

$$\mu_i = \sum_{n=1}^{\infty} nP(T_i = n) = \sum_{n=1}^{\infty} \frac{1}{n+1} = \infty \Rightarrow \text{ null recurrent}$$

6.1.6 Stationary Distribution and Limiting Distribution

Limiting distribution

$$\lim_{n\to\infty} P_{y,x}^n = \pi(x), \ \forall x, y \in S$$

Obviously, when a chain is transient, $\lim_{n\to\infty} P_{y,x}^n = 0$, there will be no limiting distribution. We can also know $\lim_{n\to\infty} P_{y,x}^n = 0$ when the chain is null recurrent.

Lemma 6.4

For an irreducible MC, $\lim_{n\to\infty} P^n_{y,x} = 0$ *for each* $x, y \in S$ ***if and only if*** *the chain is transient or null recurrent.* ♡

Theorem 6.1 (Fundamental Theorem for General Discrete Markov Chains)

An ***irreducible, positive recurrent*** *MC has a* ***unique stationary distribution*** π *(which is positive everywhere) solving the equation*

$$\sum_{y\in S} \pi(y)P(y,x) = \pi(x),\ \forall x \in S$$

$\pi(j)$ *equals to the* ***expected visiting time*** *at* j

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P^k_{ij} = \pi(j)$$

If in addition, the MC is ***aperiodic****, then*

$$\lim_{n\to\infty} P^n_{ij} = \pi(j)$$

The stationary distribution π *is also* ***inversely*** *related to the* ***expected first return times****.*

$$\pi(j) = \frac{1}{\mathbb{E}(T_j|X_0 = j)} = \frac{1}{\mu_j}$$

Furthermore, if the irreducible chain is not positive recurrent then there does not exist a stationary distribution. ♡

Note: we can prove a MC is not positive recurrent by showing the MC doesn't have a stationary distribution.

6.2 Differences between Finite and (Countably) Infinite Markov Chains

1. An irreducible MC with finite S has to be recurrent. An irreducible MC with infinite S could be recurrent or transient.
2. An irreducible MC with finite S has to be positive recurrent. An irreducible recurrent MC with infinite S could be positive recurrent ($\mathbb{E}[T_j|X_0 = j] < \infty$) or null recurrent ($\mathbb{E}[T_j|X_0 = j] = \infty$).
3. An irreducible MC with finite S always has a unique stationary distribution. An irreducible recurrent MC with infinite S has a (unique) stationary distribution if and only if the MC is positive recurrent.

Chapter 7 Branching Process

(Sir Francis Galton, 1873) It is a stochastic model for population growth. Let X_n denote the number of individuals at time n. At each time interval, each individual will produce a random number of offsprings and then die.

Two Assumptions:

(1) Each individual produces offspring with the same probability distribution: there are given non-negative numbers $p_0, p_1, ...$ summing up to 1 so that the probability of an individual producing k children is p_k.

(2) The individuals reproduce independently.

We want to know

"What is the probability that the population eventually becomes extinct?"

The number of individuals at time n, X_n is a MC with state space $S = \{0, 1, 2, ...\} = \mathbb{Z}_+$. Note that 0 is an absorbing state. Suppose $X_n = k$. Then k individuals produce offspring for the next generation. Let $Y_1, ..., Y_k$ be i.i.d random variables with $P(Y_1 = j) = p_j$. Then we can write the transition probabilities as

$$P_{k,j} = P(Y_1 + \cdots + Y_k = j)$$

Since $P(X_1 = 0|X_0 = i) = p_0^i > 0$ for each $i > 0$, the any state $i > 0$ must be transient. From this, it can be shown that, with probability 1, the chain must either get absorbed in 0 eventually or approach ∞.

7.1 Extinction Probability in a Branching Process

7.1.1 Expectation $\mathbb{E}X_n = \mu^n \mathbb{E}X_0$

The mean number of offsprings produced by an individual:

$$\mu = \sum_{i=0}^{\infty} i p_i$$

The mean number of individuals in generation n,

$$\mathbb{E}X_n = \sum_{k=0}^{\infty} P(X_{n-1} = k)\mathbb{E}(X_n|X_{n-1} = k) = \sum_{k=0}^{\infty} P(X_{n-1} = k)k\mu = \mu\mathbb{E}X_{n-1}$$

Then, we can get

$$\mathbb{E}X_n = \mu^n \mathbb{E}X_0$$

7.1.2 Lemma: $\mu < 1 \Rightarrow P(extinction) = 1$

Lemma 7.1

If $\mu < 1$, then probability of extinction is 1. ♡

Proof 7.1

We know the event $\{X_{n-1} = 0\} \subseteq \{X_n = 0\}$

$$P(extinction) = P(\cup_{n=0}^{\infty}\{X_n = 0\}) = \lim_{n\to\infty} P(X_n = 0)$$

$$P(X_n \geq 1) = \sum_{k=1}^{\infty} P(X_n = k) \leq \sum_{k=1}^{\infty} kP(X_n = k) = \mathbb{E}X_n$$

Now, the probability of survival

$$\lim_{n\to\infty} P(X_n \geq 1) \leq \lim_{n\to\infty} \mathbb{E}X_n = \lim_{n\to\infty} \mu^n \mathbb{E}X_0 = 0$$

$$\Rightarrow \lim_{n\to\infty} P(X_n \geq 1) = 0$$

Then we can conclude

$$P(extinction) = \lim_{n\to\infty} P(X_n = 0) = 1 - \lim_{n\to\infty} P(X_n \geq 1) = 1$$

If $\mu = 1$, the expected population size remains constant while if $\mu > 1$, the expected population size grows.

7.1.3 Variance: $VarX_n = n\sigma^2$ if $\mu = 1$; $VarX_n = \sigma^2\mu^{n-1}\frac{\mu^n-1}{\mu-1}$ if $\mu \neq 1$

Let's calculate the variance of X_n. We denote the variance of the number of offsprings produced by an individual by σ^2. By the law of total variance,

$$\begin{aligned} VarX_n &= Var(\mathbb{E}X_n|X_{n-1}) + \mathbb{E}Var(X_n|X_{n-1}) \\ &= Var(\mu X_{n-1}) + \mathbb{E}(\sigma^2 X_{n-1}) \\ &= \mu^2 Var(X_{n-1}) + \sigma^2\mu^{n-1}\mathbb{E}X_0 \end{aligned}$$

(Assuming $X_0 = 1$ with probability 1)

$$VarX_n = \begin{cases} n\sigma^2, & \mu = 1 \\ \sigma^2\mu^{n-1}\frac{\mu^n-1}{\mu-1}, & \mu \neq 1 \end{cases}$$

7.1.4 Extinction probability $\rho = 1$ if $\mu \leq 1$; $\rho < 1$ if $\mu > 1$

To avoid trivial cases, we assume 1.$p_0 > 0$; 2. $p_0 + p_1 < 1$.

Let $a_n(k) = P(X_n = 0|X_0 = k)$ and let $a(k) = \lim_{n\to\infty} a_n(k)$ denote the probability that the population dies out eventually assuming that $X_0 = k$.

Since all k individuals act independently, we must have

$$a(k) = a(1)^k$$

We simply denote $a(1)$ by ρ.

$$\rho = a(1) = P(extinction|X_0 = 1) = \lim_{n\to\infty} P(X_n = 0|X_0 = 1)$$

By conditioning on the first step, we can write

$$\rho = \sum_{k=0}^{\infty} P(X_1 = k|X_0 = 1)P(extinction|X_1 = k) = \sum_{k=0}^{\infty} p_k \rho^k = \psi(\rho)$$

where $\psi : [0, 1] \to \mathbb{R}$ is given by $\psi(z) = \sum_{k=0}^{\infty} p_k z^k$. Then the ρ satisfies $z = \psi(z)$

Definition 7.1

*If a random variable X takes values in $\mathbb{Z}$, the **probability generating function (pgf)** of X is the function $\psi : [0, 1] \to \mathbb{R}$ given by*

$$\psi(s) = \psi_X(s) = \mathbb{E}(s^X) = \sum_{k=0}^{\infty} s^k P(X = k)$$

♣

We now note some important properties of the function ψ.

1. $\psi'(x) = \sum_{k=1}^{\infty} x^{k-1} k p_k > 0$ for $x \in (0, 1) \Rightarrow \psi$ is an **increasing** function.
2. $\psi''(x) = \sum_{k=2}^{\infty} x^{k-2} k(k-1) p_k > 0$ for $x \in (0, 1) \Rightarrow \psi$ is a **convex** function.
3. $\psi(0) = p_0 > 0$
4. $\psi(1) = 1$
5. $\psi'(1) = \sum_{k=1}^{\infty} k p_k = \mu$
6. **Probability Generating Functions characterize the distribution**: if two discrete random variables have their pgf the same then they have the same distribution.
7. $\psi_{X+Y}(s) = \mathbb{E}(s^{X+Y}) = \mathbb{E}(s^X)\mathbb{E}(s^Y) = \psi_X(s)\psi_Y(s)$

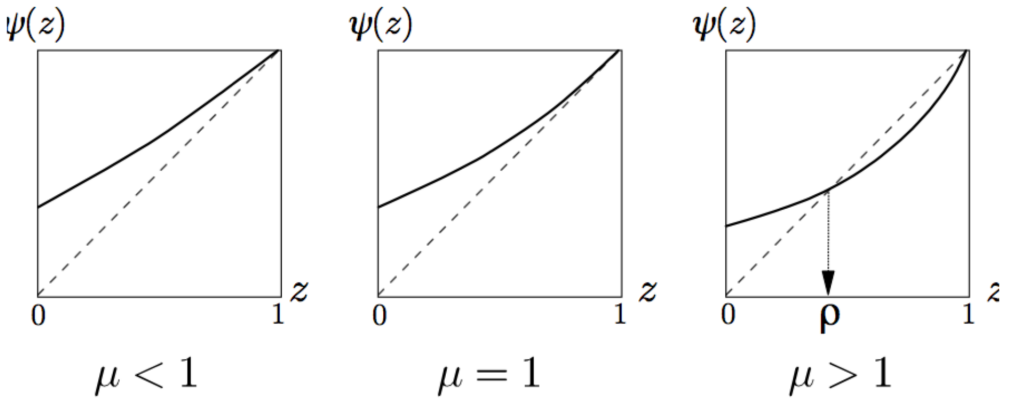

Figure 7.1: Fixed Point of pgf

From the pictures we can find that $\rho = 1$ is the unique fixed point of $\psi(z)$ when $\mu \leq 1$ and there exists another fixed point $\rho = r \in (0, 1)$ when $\mu > 1$.

Suppose $\mu > 1$. Denote $q_n = a_n(1) = P(X_n = 0|X_0 = 1)$, where $\lim_{n\to\infty} q_n = \rho$. Defining r to be the smaller solution of $\psi(z) = z$.

We want to prove $q_n \leq r, \forall n \geq 0$. Prove by induction:

1. Let $q_0 = 0$.
2. Assume that $q_n \leq r$,

$$q_{n+1} = P(X_{n+1} = 0|X_0 = 1) = \sum_{k=0}^{\infty} P(X_{n+1} = 0|X_1 = k)p_k$$
$$= \sum_{k=0}^{\infty} P(X_n = 0|X_0 = k)p_k = \sum_{k=0}^{\infty} q_n^k p_k = \psi(q_n)$$

Since ψ is increasing, we have $q_{n+1} = \psi(q_n) \leq \psi(r) = r$

Theorem 7.1

If $\mu < 1$ or $\mu = 1$, the extinction probability $\rho = 1$. If $\mu > 1$, then the extinction probability $\rho < 1$ and equals to the unique root of $z = \psi(z), z \in (0, 1)$. ♡

Example 7.1 $p_0 = \frac{1}{8}, p_1 = \frac{3}{8}, p_2 = \frac{3}{8}, p_3 = \frac{1}{8}$.

Since $\mu = \frac{3}{2} > 1$, we can solve $\frac{1}{8} + \frac{3}{8}r + \frac{3}{8}r^2 + \frac{1}{8}r^3 = r$. Because $r = 1$ is always a solution $\Rightarrow$ $(r - 1)(r^2 + 4r - 1) = 0$ $r^* = \sqrt{5} - 2$.

7.1.5 $G_n(s) = G_{n-1}(\psi(s)) = \psi(\psi(\psi(\cdots\psi(s)\cdots))) = \psi(G_{n-1}(s))$

For $n \geq 0$, let

$$G_n(s) = \sum_{k=0}^{\infty} s^k P(Z_n = k)$$

be the generating function of the n^{th} generation size Z_n.

$$G_1(s) = \psi(s)$$

We have

$$G_n(s) = \psi_{Z_n}(s) = \mathbb{E}(s^{Z_n}) = \mathbb{E}\left(\mathbb{E}(s^{\sum_{k=1}^{z} X_k})|Z_{n-1} = z\right)$$
$$= \mathbb{E}\left(\prod_{k=1}^{z} \mathbb{E}(s^{X_k})|Z_{n-1} = z\right) = \mathbb{E}\left((\psi(s))^z|Z_{n-1} = z\right)$$
$$= \mathbb{E}\left[(\psi(s))^{Z_{n-1}}\right] = G_{n-1}(\psi(s))$$

Since $G_2(s) = G_1(\psi(s)) = \psi(\psi(s)) = \psi(G_1(s))$, we can infer

$$G_n(s) = G_{n-1}(\psi(s)) = \psi(\psi(\psi(\cdots\psi(s)\cdots))) = \psi(G_{n-1}(s))$$

Chapter 8 Time Reversible Markov Chains

8.1 Definition: Local Balance $\pi(i)P(i,j) = \pi(j)P(j,i), \forall i,j \in S$

Definition 8.1

*We say that a MC is **time reversible** if, for each $n \geq 1$, the distribution of $(X_0, \ldots, X_n)$ is the same as the distribution of $(X_n, \ldots, X_0)$. Equivalently, for any $x_0, \ldots, x_n \in \mathcal{S}$ we have*

$$P(X_0 = x_0, X_1 = x_1, \ldots, X_n = x_n) = P(X_n = x_0, X_{n-1} = x_{n-1}, \ldots, X_0 = x_n).$$

In words, the probability of a given trajectory is the same as the probability of the reverse trajectory. ♣

Lemma 8.1 (Local Balance)

*The Markov chain $X_0, X_1, \ldots$ is time-reversible **if and only if** the distribution π of X_0 satisfies the condition*

$$\pi(i)P(i,j) = \pi(j)P(j,i), \forall i,j \in S$$

♡

8.2 Discussion about Local Balance

8.2.1 Flow: $Flow(A,B) = \sum_{i\in A}\sum_{j\in B}\pi(i)P_{ij}$

Definition 8.2

*For a distribution π on the state space S and any two subsets of the state space A, B define the **Flow***

$$Flow(A,B) = \sum_{i\in A}\sum_{j\in B}\pi(i)P_{ij}$$

♣

8.2.2 Lemma: $Flow(A,A^c) = Flow(A^c,A)$ for any subset $A \subset S$

Lemma 8.2

$Flow(A,A^c) = Flow(A^c,A)$ for any subset $A \subset S$.

♡

Proof 8.1

$$Flow(A,A^c) = \sum_{i\in A}\sum_{j\in A^c}\pi(i)P_{ij} = \sum_{i\in A}\pi(i)(1-\sum_{j\in A}P_{ij}) = \sum_{i\in A}\pi(i) - \sum_{i\in A}\sum_{j\in A}\pi(i)P_{ij}$$

$$Flow(A^c,A) = \sum_{i\in A^c}\sum_{j\in A}\pi(i)P_{ij} = \sum_{j\in A}(\pi(j) - \sum_{i\in A}\pi(i)P_{ij}) = \sum_{j\in A}\pi(j) - \sum_{j\in A}\sum_{i\in A}\pi(i)P_{ij}$$

8.2.3 Lemma: Local balance $\Rightarrow \pi$ is stationary

Lemma 8.3

If the local balance equations "$\pi(i)P(i,j) = \pi(j)P(j,i), \forall i,j \in S$" hold then π is stationary. ♡

Proof 8.2

$$(\pi P)_i = \sum_{j \in S} \pi_j P_{ji} = \sum_{j \in S} \pi_i P_{ij} = \pi_i$$

8.2.4 Lemma: All stationary birth and death chains are reversible

Lemma 8.4

All stationary birth and death chains are reversible. (i.e. For a MC with $P_{i,j} = 0, \forall |i-j| > 1$, $\pi(i)P(i,j) = \pi(j)P(j,i), \forall i,j \in \mathbb{Z}_+$) ♡

Proof 8.3

It is enough to show the equation holds when $j = i+1$. For $A = \{0,1,2,...,i\}$,

$$Flow(A, A^c) = \sum_{0 \le k \le i} \sum_{j > i} \pi(k) P_{kj} = \pi(i) P_{i,i+1}$$

$$= Flow(A^c, A) = \sum_{j > i} \sum_{0 \le k \le i} \pi(j) P_{jk} = \pi(i+1) P_{i+1,i}$$

8.3 Example: Random Walk on an Undirected Graph

Lemma 8.5

Any stationary random walk on a weighted undirected graph is time reversible. On the other hand, any time reversible MC can be thought of as a random walk on a weighted undirected graph. ♡

Proof 8.4

Consider a RW on a weighted undirected graph $G = (V, W)$. Recall that every potential edge or a pair of states i, j has some weight $W_{ij} \ge 0$. Since the graph is undirected this means that the edge weights $W_{ij} = W_{ji}$ are symmetric. The transition probabilities are $P_{v,u} = \frac{W_{vu}}{\sum_{v \in S} W_{vu}}$, where $S = \{v : W_{uv} \neq 0\}$. By the symmetric property, $P_{v,u} = \frac{W_{vu}}{\sum_{v \in S} W_{vu}} = \frac{W_{uv}}{\sum_{v \in S} W_{uv}}$. Let's denote $W = \sum_{(i,j) \in V \times V} W_{ij}$.

We know from an earlier lecture that the stationary distribution is given by $\pi(v) = \frac{\sum_{v \in S} W_{uv}}{W}$. Let's now

check that this π satisfies the local balance.

$$\pi(v)P_{v,u} = \frac{\sum_{v\in\mathcal{S}} W_{uv}}{W} \frac{W_{uv}}{\sum_{v\in\mathcal{S}} W_{uv}} = \frac{W_{uv}}{W}.$$

The right hand side above is symmetric in u, v so local balance must hold. On the other hand, lets consider a time reversible MC. Build a graph where the set of vertices is same as the state space of this MC. Define the edge weights to be $W_{ij} = \pi_i P_{ij}$. Since local balance holds we have $W_{ij} = W_{ji}$. Now we can imagine a random walk on this weighted undirected graph. What is the transition probability Q of this random walk? It has to be

$$Q_{uv} = \frac{W_{uv}}{\sum_{v\in\mathcal{S}} W_{uv}} = \frac{\pi(v)P_{v,u}}{\sum_{u\in\mathcal{S}} \pi(v)P_{v,u}} = P_{v,u}.$$

Therefore, this random walk describes the same MC *as the original one.*

Chapter 9 Markov Chain Monte Carlo (MCMC)

Given a probability distribution π, the goal of MCMC is to simulate a random variable X whose distribution is π.

The MCMC algorithm constructs an ergodic (irreducible and aperiodic) Markov chain whose limiting distribution is the desired π.

9.1 Strong Law of Large Numbers for Markov Chains

Theorem 9.1

Assume that $X_0, X_1, ...$ is an ergodic Markov chain with stationary distribution π. Let r be a bounded and real-valued function. Let Y be a random variable with distribution π. Then, with probability 1,

$$\lim_{n\to\infty} \frac{r(X_1) + \cdots + r(X_n)}{n} = \mathbb{E}_Y[r(Y)]$$

where $\mathbb{E}_Y[r(Y)] = \sum_j r(j)\pi_j$

♡

What different to LLN is $X_1, .., X_n$ are not i.i.d.

9.2 Example of Designing MC

1. All binary sequence consisted of 0 or 1 of length d, $S = \{0,1\}^d$, $|S| = 2^d$. π is a uniform distribution on all S, obviously, each sequence has probability $\frac{1}{2^d}$. We can use a MC of tossing a coin to simulate it.
2. All binary sequence consisted of 0 or 1 of length d and **no consecutive 1**, $B = \{0,1\}^d$. π is a uniform distribution on all B, but it is hard to get the $|B|$ in this situation.
   - **Ejection Sampling:** sample a random sequence from S, eject the sequence is not in B. (issue: usually the $|B|$ is small compared to $|S|$, the method works badly, e.g. when $d = 100$, $|B| \sim 10^{21}, |S| \sim 10^{30}$)
   - **MCMC:** Given sequence $(x_1, ..., x_d)$, pick a coordinate at random (with probability $\frac{1}{d}$ each). If the coordinate is 1 then flip it to 0. If the coordinate is 0, then flip it to 1 if this results in a sequence in B, otherwise do not flip it.

     We can identify the facts of the MC:

     (1) The MC is irreducible;

     (2) Period = 1;

(3) For $i \neq j$, $P_{ij} = P_{ji} = 1$ or $= 0 \Rightarrow$ local balance is satisfied for π

The above three facts imply that the uniform distribution π is the stationary and limiting distribution of this chain

Interested in calculating the expected number of 1 in a sequence $\mathbb{E}f(\pi)$ (the function of counting 1 in a sequence is represented by $f(\cdot)$), suppose $X \sim \text{Unif}(B)$. We can calculate

$$\frac{f(X_1) + \cdots + f(X_n)}{n} \to \mathbb{E}f(\pi)$$

9.3 Metropolis Hastings Algorithm

Given a proposal chain T, we want local balance holds:

$$\pi_i T_{ij} = \pi_j T_{ji}$$

However, sometimes the equation may not hold. We modify the transition matrix by

$$\pi_i T_{ij} A_{ij} = \pi_j T_{ji}, \text{ where } A_{ij} = \frac{\pi_j T_{ji}}{\pi_i T_{ij}}$$

Assume at time n, the chain is at state i or equivalently, $X_n = i$. The next step of the chain X_{n+1} is determined by the following two-step procedure.

1. Choose a new state according to the transition matrix T. That is, choose j with probability T_{ij}. State j is called the proposal state.
2. Define

$$A_{ij} = \min\left\{1, \frac{\pi_j T_{ji}}{\pi_i T_{ij}}\right\} \quad \text{(Actually, it is fine to let } A_{ij} = \frac{\pi_j T_{ji}}{\pi_i T_{ij}}\text{)}$$

Generate a uniformly random number between 0 and 1 as $U \sim U(0,1)$. If $U \leq A_{ij}$ then j is accepted as the next state of the chain. If $U > A_{ij}$ then j is not accepted as the next state of the chain and $X_{n+1} = i$.

Lemma 9.1

Let P denote the modified transition matrix of the Metropolis-Hastings algorithm. Then π satisfies local balance with respect to P. ♡

Therefore, π is stationary and if the MC with the new transition dynamics is ergodic then π is limiting. If we start out with an irreducible chain then the final chain is also irreducible.

Note: If the proposal chain is ergodic so is the resulting Metropolis Hastings chain.

9.3.1 Example of generate standard normal distribution with uniform

Suppose we want to generate a standard normal random variable using only a uniform random number generator. The target density function is $\pi(t) = \frac{\exp(-t^2/2)}{\sqrt{2\pi}}$. For the proposal distribution, we choose the uniform

distribution of length 2 centered at the current state. From state s, the proposal chain moves to t, where t is uniformly distributed on $(s-1, s+1)$. The conditional density $T(s,t) = 1/2$ if $|s-t| \leq 2$ and 0 otherwise. The acceptance function then becomes

$$A(s,t) = \min\left\{1, \frac{\pi(t)T_{ts}}{\pi(s)T_{st}}\right\} = \min\left\{1, \exp\left(\left[-t^2 + s^2\right]/2\right)\right\}$$

9.3.2 Without MCMC: Box Muller Transform

There are methods to sample exactly from the standard normal distribution without using MCMC. For any continuous random variable X with CDF F, the random variable $F^{-1}(U)$ has the same distribution as X when $U \sim Unif(0,1)$. For the standard normal the function F^{-1} is not available in closed form. There is another method called the *Box Muller Transform*.

The basic idea is as follows. X, Y are two independent standard normal random variables if and only if (R, Θ) are independent, Θ follows $Unif(0, 2\pi)$ and R^2 follows a Chi Squared distribution with degrees of freedom 2 which is the same as the Exponential Distribution with mean 2. Here (R, Θ) are the polar coordinates corresponding to the cartesian coordinates (X, Y). Therefore, to sample two independent standard normals it is enough to sample R and Θ. Sampling $\Theta \sim \text{Unif}(0, 2\pi)$ is easy and sampling $R = \sqrt{R^2} \sim \sqrt{\text{Exponential }(2)}$ is easy by the inverse CDF method.

9.4 Gibbs Sampling

Gibbs sampling is a MCMC algorithm for obtaining approximate draws from a joint distribution, based on sampling from **conditional distributions** one at a time: at each stage, one variable is updated (keeping all the other variables fixed) by drawing from the conditional distribution of that variable given all the other variables. This approach is especially useful in problems where the conditional distributions are simple enough to simulate from but the overall joint distribution is complicated.

There are several forms of Gibbs samplers, depending on the order in which updates are done. We will introduce two major kinds of Gibbs sampler: systematic scan, in which the updates sweep through the components in a deterministic order, and random scan, in which a randomly chosen component is updated at each stage.

9.4.1 Systematic scan Gibbs sampler

Let X and Y be discrete r.v.s with joint PMF $p(x,y) = P(X = x, Y = y)$. We wish to construct a two-dimensional Markov chain (X_n, Y_n) whose stationary distribution is p. The *systematic scan Gibbs sampler* proceeds by updating the X component and the Y component in alternation. If the current state is $(X_n, Y_n) =$

(x_n, y_n), then we update the X component while holding the Y component fixed, and then update the Y component while holding the X component fixed:

1. Draw a value x_{n+1} from the conditional distribution of X given $Y = y_n$, and set $X_{n+1} = x_{n+1}$
2. Draw a value y_{n+1} from the conditional distribution of Y given $X = x_{n+1}$, and set $Y_{n+1} = y_{n+1}$
3. Repeating steps 1 and 2 over and over, the stationary distribution of the chain (X_0, Y_0), (X_1, Y_1), (X_2, Y_2), ... is p.

Why is the last statement true? Suppose we are updating the X coordinate. Suppose $(X, Y) \sim p$. We transition to (X', Y) where X' is drawn from the conditional distribution of p given Y. So we can write

$$P\left(X' = x, Y = y\right) = P\left(X' = x \mid Y = y\right) P(Y = y) = p(x \mid y)p(y) = p(x, y)$$

The second equality is true because $(X, Y) \sim p$. The above display shows that p is stationary for this chain.

9.4.2 Random Scan Gibbs sampler

Similarly, we wish to construct a two-dimensional Markov chain (X_n, Y_n) whose stationary distribution is p. However, each move of the *random scan Gibbs sampler* picks a uniformly random component and updates it, according to the conditional distribution given the other component:

1. Choose which component to update, with equal probabilities.
2. If the X-component was chosen, draw a value x_{n+1} from the conditional distribution of X given $Y = y_n$, and set $X_{n+1} = x_{n+1}, Y_{n+1} = y_n$. Similarly, if the Y-component was chosen, draw a value y_{n+1} from the conditional distribution of Y given $X = x_n$, and set $X_{n+1} = x_n, Y_{n+1} = y_{n+1}$.
3. Repeating steps 1 and 2 over and over, the stationary distribution of the chain (X_0, Y_0), (X_1, Y_1), (X_2, Y_2), ... is p.

Gibbs sampling generalizes naturally to higher dimensions. If we want to sample from a d dimensional joint distribution, the Markov chain we construct will be a sequence of d dimensional random vectors. At each stage, we choose one component of the vector to update, and we draw from the conditional distribution of that component **given the most recent values of the other components**. We can either cycle through the components of the vector in a systematic order, or choose a random component to update each time.

The Gibbs sampler is less flexible than the Metropolis-Hastings algorithm in the sense that we don't get to choose a proposal distribution; this also makes it simpler in the sense that we don't have to choose a proposal distribution. The flavors of Gibbs and Metropolis-Hastings are rather different, in that Gibbs emphasizes conditional distributions while Metropolis-Hastings emphasizes acceptance probabilities. But the algorithms are closely connected, as we show below.

Theorem 9.2 (Random scan Gibbs as Metropolis-Hastings)

The random scan Gibbs sampler is a special case of the Metropolis-Hastings algorithm, in which the proposal is always accepted. In particular, it follows that the stationary distribution of the random scan Gibbs sampler is as desired.

♡

The random scan Gibbs sampler proposal chain satisfies local balance.

9.4.3 Example: Bivariate Normal Distribution

Consider a bivariate standard normal distribution with a correlation of ρ. If (X, Y) has a bivariate normal distribution then the conditional distribution of $X \mid Y = y$ is normal with mean ρy and variance $1 - \rho^2$. Similarly, the conditional distribution of $Y \mid X = x$ is normal with mean ρx and variance $1 - \rho^2$. Therefore, we can implement Gibbs sampler by simply generating normal random variables each time. We write the steps when using the deterministic scan version, although the random scan version is equally applicable.

(a) Initialize $(x_0, y_0) = (0, 0)$. Also initialize $n = 1$.

(b) Generate $x_n \sim N\left(\rho y_{n-1}, 1 - \rho^2\right)$.

(c) Generate $y_n \sim N\left(\rho x_n, 1 - \rho^2\right)$.

(d) Update $n = n + 1$.

(e) Return to Step (b).

Remark. Recall that there is a simple exact method to sample standard Bivariate Normal with correlation ρ. First sample two i.i.d $Z_1, Z_2 \sim N(0, 1)$. Now let $X = Z_1$ and $Y = \rho Z_1 + \sqrt{1 - \rho^2} Z_2$.

9.4.4 Example: Potts model (Ising model)

The Potts model is a generalization of the Ising model, and is used regularly in statistical physics and computer science. Fix a graph $G = (V, E)$. Let $\sigma : V \to \{1, \cdots, q\}$ for some $q > 1$. The Potts model is a probability distribution on such functions σ, and is specified as follows:

$$\pi(\sigma) = \frac{1}{Z} \exp\left(\beta \sum_{i \sim j} \mathbf{1}\left(\sigma_i = \sigma_j\right)\right)$$

where $\beta \in \mathbb{R}$ is a constant. Derive a Gibbs sampler for the Potts model.

Cycle over the coordinates. At each step, given $\sigma_{-i} := (\sigma_1, ..., \sigma_{i-1}, \sigma_{i+1}, ..., \sigma_n)$, sample σ_i from the distribution

$$\pi_i(k \mid \sigma_{-i}) = \frac{1}{Z} \exp\left(\beta \sum_{j \sim i} \mathbf{1}\left(\sigma_j = k\right)\right)$$

(a) Initialize $\sigma_{0,j} = 1, \forall j \in V$. Also initialize $n = 1$.

(b) For i in V: Generate $\sigma_{n,i} \sim \pi_i(k \mid \sigma_{n-1,-i})$.

(c) Update $n = n + 1$.

(d) Return to Step (b).

9.5 A Linear Algebraic Condition for Convergence

Let's assume a finite reversible ergodic MC with transition matrix P and stationary distribution π. Suppose the cardinality of the state space is k. Let Q be the diagonal matrix with diagonals square root of the entries of π. Let $A = QPQ^{-1}$. We can check that

$$A_{ij} = \sum_{r=1}^{k}\sum_{s=1}^{k} Q_{ir}P_{rs}Q^{-1}_{sj} = Q_{ii}P_{ij}Q^{-1}_{jj} = \sqrt{\frac{\pi_i}{\pi_j}}P_{ij}$$

Since the chain is reversible, we obtain

$$A_{ij} = \sqrt{\frac{\pi_i}{\pi_j}}P_{ij} = \frac{\pi_i P_{ij}}{\sqrt{\pi_i\pi_j}} = \frac{\pi_j P_{ji}}{\sqrt{\pi_i\pi_j}} = A_{ji}.$$

We will now use some Linear Algebra facts which we state below.

(1) **Spectral decomposition theorem:** A real symmetric matrix A can be orthogonally diagonalized, that is

$$A = SDS^T$$

where S is an orthonormal real matrix and D is a diagonal real matrix with A's eigenvalues as entries.

(2) Since $P = Q^{-1}AQ$, P has the same eigenvalues as A.

$(det(P - \lambda I) = 0 \Rightarrow 0 = det(QP - \lambda Q) = det(AQ - \lambda Q) = 0 \Rightarrow det(A - \lambda I) = 0)$

(3) Since P is ergodic (irreducible and aperiodic), P has regular transition matrix. Then the eigenvalues can be written

$$1 = \lambda_1 > \lambda_2 \geq \lambda_3 \geq \cdots \geq \lambda_k > -1$$

Now we can write

$$P = Q^{-1}AQ = Q^{-1}\left(SDS^T\right)Q = \left(Q^{-1}S\right)D\left(S^TQ\right)$$

where D has diagonal entries $1, \lambda_2, \ldots, \lambda_k$. Now, for any integer $n \geq 1$, we can compute $P^n = \left(Q^{-1}S\right)D^n\left(S^TQ\right)$. Taking the ij th entry,

$$P^n_{ij} = \sum_{t=1}^{k}\left(Q^{-1}S\right)_{it}\lambda_t^n\left(S^TQ\right)_{tj} = \sqrt{\frac{\pi_j}{\pi_i}}\sum_{t=1}^{k}\lambda_t^n S_{it}S_{jt} = \sqrt{\frac{\pi_j}{\pi_i}}S_{i1}S_{j1} + \sqrt{\frac{\pi_j}{\pi_i}}\sum_{t=2}^{k}\lambda_t^n S_{it}S_{jt}$$

Since $\lim_{n\to\infty} P^n_{ij} = \pi_j$ this means that $\sqrt{\frac{\pi_j}{\pi_i}}S_{i1}S_{j1} = \pi_j$ and moreover we can write

$$\left|P^n_{ij} - \pi_j\right| = \left|\sqrt{\frac{\pi_j}{\pi_i}}\sum_{t=2}^{k}\lambda_t^n S_{it}S_{jt}\right| \leq \underbrace{\sqrt{\frac{\pi_j}{\pi_i}}\sum_{t=2}^{k}|S_{it}S_{jt}|}_{T_1}\underbrace{\max_{2\leq t\leq n}|\lambda_t^n|}_{T_2}.$$

The T_1 term is a constant and does not change with n. The term T_2 decreases geometrically because it is strictly less than 1. The above display shows that the rate of convergence of a reversible ergodic MC is governed by how close the second largest (in absolute value) eigenvalue is to 1 in absolute value. This gap between 1 and the second-largest eigenvalue (in absolute value) is often called the spectral gap and if this gap is not too small then the convergence happens exponentially fast.

Remark. In principle, for any MCMC method we can just compute its spectral gap to know how fast it will converge. In practice, this is often not possible as the state space is too large to compute eigenvalues of the transition matrix.

Chapter 10 Poisson Process

10.1 Basics of Poisson Process

10.1.1 Counting Process

Definition 10.1 (discrete time)

A counting process $(N_t)_{t\geq 0}$ is a collection of non-negative integer valued random variables such that if $0 \leq s \leq t$, then $N_s \leq N_t$.

♣

We use $N(t)$ to denote counting process.

10.1.2 Poisson Distribution

Lemma 10.1

If λ_n is a sequence of positive numbers with $\lim_{n\to\infty} \lambda_n = \lambda$ then

$$\lim_{n\to\infty} \left(1 - \frac{\lambda_n}{n}\right)^n = e^{-\lambda}$$

♡

Theorem 10.1

If $n \to \infty$ and $p \to 0$ in such a way that $Np \to \lambda$, then the Binomial distribution with parameters (N, p) converges to the Poisson λ distribution.

♡

Consider a unit time (the unit is divided into n equal subparts, $n \to \infty$), there is an event may occur with every subpart, the number of the event happens should follow binomial distribution $B(n, p)$, where $n \to \infty, p \to 0$; $\lambda = n \cdot p$ is the expected number of events in this period of time.

The probability the number of the event happens:

$$\begin{aligned}
\Pr(X = k) &= \lim_{n\to\infty} \binom{n}{k} (\frac{\lambda}{n})^k (1 - \frac{\lambda}{n})^{n-k} \\
&= \lim_{n\to\infty} \frac{n!}{(n-k)!k!} (\frac{\lambda}{n})^k (1 - \frac{\lambda}{n})^n (1 - \frac{\lambda}{n})^{-k} \\
&= \lim_{n\to\infty} \frac{n!}{(n-k)!k!} (\frac{\lambda}{n})^k e^{-\lambda} \\
&= \frac{\lambda^k e^{-\lambda}}{k!} \lim_{n\to\infty} \frac{n!}{(n-k)!n^k} \\
&= \frac{\lambda^k e^{-\lambda}}{k!} \lim_{n\to\infty} \frac{n}{n} \frac{n-1}{n} \cdots \frac{n-k+1}{n} \\
&= \frac{\lambda^k e^{-\lambda}}{k!}
\end{aligned}$$

For a Poisson distribution $Pois(\lambda)$. λ is frequency of the event, i.e., the average number of event happens within unit time.

$$\Pr(X{=}k) = \frac{\lambda^k e^{-\lambda}}{k!},\ k = 0, 1, 2, 3...$$

$$E(X) = Var(X) = \lambda$$

10.1.3 Definition of Poisson Process

Definition 10.2 (Definition 1 of PP)

*A **Poisson Process with parameter** λ is a counting process $(N_t)_{t\geq 0}$ with the following properties:*

1. $N_0 = 0$.
2. *For all $t > 0$, N_t has a Poisson distribution with parameter λt. $N_t \sim Pois(\lambda t)$.*
$$P(N_t = k) = \frac{(\lambda t)^k e^{-\lambda t}}{k!}$$
3. *For all $s, t > 0$, the increment $N_{t+s} - N_s$ has the same distribution as N_t. This property is called **stationary increments**.*
4. *For $0 \leq q < r \leq s < t$, the increments $N_t - N_s$ and $N_r - N_q$ are independent random variables. This property is called **independent increments**.*

♣

The stationary increments property says that the distribution of the number of arrivals in an interval only depends on the length of the interval. The independent increment property says that the number of arrivals on disjoint intervals are independent random variables. Since N_t has a Poisson distribution, $\mathbb{E}N_t = \lambda t$. So, we expect about λt arrivals in t time units. We say that the rate of arrivals is λ.

Proposition 10.1 (Translated PP)

Let N_t be a PP with parameter λ. For $t_0 > 0$, let

$$\tilde{N}_t = N_{t+t_0} - N_{t_0}$$

for $t \geq 0$. Then $(\tilde{N}_t)_{t\geq 0}$ is a Poisson process with parameter λ.

♠

10.2 Inter-Arrival Times

10.2.1 First arrival time: Exponential distribution $Exp(\lambda)$

For a PP with parameter λ, let X denote the time of the first arrival. Then, the event $\{X > t\}$ happens if and only if there are no arrivals in $[0, t]$. Thus, for any $t \geq 0$,

$$P(X > t) = P(N_t = 0) = e^{-\lambda t}$$

Hence, X has an exponential distribution with parameter λ or mean $\frac{1}{\lambda}$.

$$f_X(t) = \lambda e^{-\lambda t}, \text{ for } t > 0$$
$$\mathbb{E}(X) = \frac{1}{\lambda};\ Var(X) = \frac{1}{\lambda^2}$$

Recall that the pdf of the Exponential distribution with parameter 1 is given by $f(x) = \exp(-x)$ for $x > 0$, and an Exponential with parameter λ can always be represented as $\frac{1}{\lambda}\,\text{Exp}(1)$.

Definition 10.3 (Definition 2 of PP)

Let $X_1, X_2, \ldots$ be a sequence of i.i.d exponential random variables with parameter λ or mean $\frac{1}{\lambda}$. For $t > 0$, let

$$N_t = \max\{n : X_1 + \cdots + X_n \leq t\}$$

with $N_0 = 0$. Then $(N_t)_{t\geq 0}$ defines a Poisson process with parameter λ.

♣

Claim 10.1 (Simulation 1.)

The above definition leads to a direct method for simulating a Poisson Process in $[0, t]$.

♠

10.2.2 k^{th} arrival time: Gamma distribution $Gamma(n, \lambda)$

Remark. The above definition says that a PP is a counting process for which interarrival times are i.i.d exponential random variables. Let $S_n = X_1 + \cdots + X_n$ for $n = 1, 2, \ldots$ We call $S_1, S_2, \ldots$ as arrival times of the process, where S_k is the k^{th} arrival. Furthermore, $X_k = S_k - S_{k-1}$ is the k^{th} interarrival time between the $(k-1)^{th}$ arrival and the k^{th} arrival, with $S_0 = 0$.

For a PP, each arrival time S_n is a sum of n i.i.d exponential inter arrival times. A sum of i.i.d exponential (λ) distribution has a $Gamma(n, \lambda)$ distribution. The pdf of a Gamma (n, λ) is

$$f(t; n, \lambda) = \frac{\lambda^n t^{n-1} e^{-\lambda t}}{(n-1)!}, \text{ for } t > 0$$

The cdf is,

$$F(t; n, \lambda) = 1 - \sum_{i=0}^{n-1} \frac{(\lambda t)^i}{i!} e^{-\lambda t} = e^{-\lambda t} \sum_{i=n}^{\infty} \frac{(\lambda t)^i}{i!}.$$

Intuitively,

$$\mathbb{E}(S_n) = \frac{n}{\lambda};\ Var(S_n) = \frac{n}{\lambda^2}$$

10.2.3 Memorylessness of the Exponential Random Variable

Definition 10.4 (Memorylessness)

A positive random variable X possesses the memoryless property if for every $x \geq 0$ and $t > 0$,

$$P(X > t + x) = P(X > t)P(X > x)$$

or equivalently,

$$P(X > t + x \mid X > x) = P(X > t).$$

♣

Lemma 10.2

If X is a continuous random variable then it satisfies memorylessness if and only if it is an Exponential random variable with some parameter $\lambda > 0$, $P(X > t) = e^{-\lambda t}$.

♡

Proof 10.1

For an exponential r.v. X of rate $\lambda > 0, P(X > x) = \exp(-\lambda x)$ for $x \geq 0$. This satisfies the memorylessness equation so X is memoryless. Conversely, an arbitrary continuous random variable X is memoryless only if it is exponential. To see this, let $h(x) = \log[P(X > x)]$ and observe that $h(x)$ is strictly decreasing. In addition, the memorylessness equation says that $h(t + x) = h(x) + h(t)$ for all $x \geq 0, t > 0$. These two statements imply that $h(x)$ must be linear in x with negative slope and hence $\Pr(X > x)$ must be exponential in x.

Remark. The only discrete random variable which has the memoryless property is the geometric distribution. This is not a surprise as the exponential distribution can be thought of as a continuous version of the Geometric distribution.

Bus Example: Assume that Amy and Zach each want to take a bus. Buses arrive at a bus stop according to a PP with rate $\frac{1}{30}$ per minute. Unlucky Amy gets to the bus stop just as a bus leaves the stop. Her waiting time for the next bus is Exponential with mean 30 minutes. Suppose no bus arrives in the next 10 minutes and at this moment Zach arrives. The waiting time for Zach is also Exponential with mean 30 minutes and remarkably the additional waiting time of Amy also has the same distribution.

10.3 Conditioning on the number of arrivals in a Poisson Process: Uniform

What happens when we take a Poisson process and condition on the total number of events in an interval? In other words, given that $N(1) = k$ how are the k points within $[0, 1]$ distributed?

First, let us consider a Bernoulli process with a large N and a small p. Conditioning on $X_1 + \cdots + X_N = k$ what is the joint distribution of $(X_1, \ldots, X_N)$? It should be uniform over all binary vectors with k ones and

$n-k$ zeroes. (In fact this holds irrespective of what N and p is.) The following theorem can be thought of as a limiting version of this fact.

Theorem 10.2

*Given that $N(1)=k$, the k points are **uniformly distributed** on $[0,1]$.*

That is, for any partition $J_1,\ldots,J_m$ of $[0,1]$ into non overlapping intervals ($J_1+J_2+\cdots+J_m=1$),

$$P\left(N\left(J_i\right)=k_i,\forall i\in[1:m]\mid N(1)=k\right)=\frac{k!}{k_1!\ldots k_m!}\prod_{i=1}^{m}\left|J_i\right|^{k_i}$$

for any non-negative integers $k_1,\ldots,k_m$ summing up to k. Here we are abusing notation and denoting the number of arrivals within the interval J_i by $N\left(J_i\right)$ and we denote the length of J_i by $|J_i|$. ♡

(More general, for $N(t)=k$:

$$P\left(N\left(J_i\right)=k_i,\forall i\in[1:m]\mid N(t)=k\right)=\frac{k!}{k_1!\ldots k_m!}\prod_{i=1}^{m}\left(\frac{|J_i|}{t}\right)^{k_i}$$

where $J_1+J_2+\cdots+J_m=t$)

Remark. The above theorem is saying that the distribution of the random vector $\left(N\left(J_1\right),\ldots,N\left(J_m\right)\right)$ is distributed as Multinomial with number of trials n and probabilities $\left(|J_1|,\ldots,|J_m|\right)$

Proof 10.2

The random variables $N\left(J_i\right)$ are independent Poisson r.v.s with means $\lambda|J_i|$ by the definition of a Poisson process. Hence, for any nonnegative integers $k_1,\ldots,k_m$ with $k_1+k_2+\cdots+k_m=k$.

$$P\left(N\left(J_i\right)=k_i,\forall i\in[1:m]\right)=\prod_{i=1}^{m}\left[\left(\lambda\left|J_i\right|\right)^{k_i}\frac{e^{-\lambda|J_i|}}{k_i!}\right]=\lambda^k e^{-\lambda}\prod_{i=1}^{m}\frac{|J_i|^{k_i}}{k_i!}$$

Dividing this by

$$P(N(1)=k)=e^{-\lambda}\frac{\lambda^k}{k!}$$

yields the desired conditional probability. Finally, to obtain the connection with the uniform distribution, observe that if one were to drop k points independently in $[0,1]$ according to the uniform distribution then the probability that interval J_i would contain exactly k_i points for each $i=1,2,\ldots,m$ would also be given by the same probability.

Claim 10.2 (Simulation 2.)

This suggests another way to simulate a Poisson point process of rate λ in $[0,a]$ for any integer $a\geq 1$. First, construct the counts $N[0,1]$, $N[1,2]$, $N[2,3]$, ... by i.i.d. sampling from the Poisson distribution with mean λ. Then, independently, throw down $N[i,i+1]$ points at random in the interval $[i,i+1]$ according to the uniform distribution. ♠

Corollary 10.1

*Let $S_1, S_2, ...$ be the occurrence/arrival times in a Poisson process $N(t)$ of rate λ. Then conditional on the event $N(1) = m$, the random variables $S_1, ..., S_m$ are distributed in the same manner as the **order statistics** of a sample of m i.i.d. $uniform[0, 1]$ random variables.* ♡

The probability density function of the order statistic $U_{(k)}$ is equal to

$$f_{U_{(k)}}(u) = \frac{n!}{(k-1)!(n-k)!} u^{k-1}(1-u)^{(n-k)}$$

That is the $U_{(k)} \sim Beta(k, n+1-k)$

Example 10.1(enter and leave) Students enter a campus building according to a Poisson process $(N_t)_{t\geq 0}$ with parameter λ. The times spent by each student in the building are i.i.d. random variables with continuous cumulative distribution function $F(t)$. Find the probability mass function of the number of students in the building at time t, assuming there are no students in the building at time 0.

Solution *Let B_t denote the number of students in the building at time t. Conditioning on N_t,*

$$\begin{aligned} P(B_t = k) &= \sum_{n=k}^{\infty} P(B_t = k | N_t = n) P(N_t = n) \\ &= \sum_{n=k}^{\infty} P(B_t = k | N_t = n) \frac{e^{-\lambda t}(\lambda t)^n}{n!} \end{aligned}$$

Assume that n students enter the building by time t, with arrival times $S_1, ..., S_n$. Let Z_k be the length of time spent in the building by the k^{th} student, for $1 \leq k \leq n$. Then, $Z_1, ..., Z_n$ are i.i.d. random variables with cdf F, and students leave the building at times $S_1 + Z_1, ..., S_n + Z_n$. There are k students in the building at time t if and only if k of the departure times $S_1 + Z_1, ..., S_n + Z_n$ exceed t. This gives

$$\begin{aligned} P(B_t = k | N_t = n) &= P(k \text{ of the } S_1 + Z_1, ..., S_n + Z_n \text{ exceed } t | N_t = n) \\ &= P(k \text{ of the } U_{(1)} + Z_1, ..., U_{(n)} + Z_n \text{ exceed } t) \\ &= P(k \text{ of the } U_1 + Z_1, ..., U_n + Z_n \text{ exceed } t) \\ &= \binom{n}{k} p^k (1-p)^{n-k} \end{aligned}$$

where $p = P(U_1 + Z_1 > t) = \frac{1}{t}\int_0^t P(Z_1 > t - x) = \frac{1}{t}\int_0^t [1 - F(x)]dx$. *Then,*

$$\begin{aligned} P(B_t = k) &= \sum_{n=k}^{\infty} P(B_t = k | N_t = n) \frac{e^{-\lambda t}(\lambda t)^n}{n!} \\ &= \sum_{n=k}^{\infty} \binom{n}{k} p^k (1-p)^{n-k} \frac{e^{-\lambda t}(\lambda t)^n}{n!} \\ &= e^{-\lambda t} \frac{p^k (\lambda t)^k}{k!} \sum_{n=k}^{\infty} \frac{(1-p)^{n-k}(\lambda t)^{n-k}}{(n-k)!} \\ &= \frac{(\lambda p t)^k e^{-\lambda p t}}{k!} \end{aligned}$$

That is B_t *has a Poisson distribution with parameter* λpt.

$$B_t \sim \text{Pois}(\lambda p t)$$

When Z_i follows exponential distribution with parameter λ', $F(x) = 1 - e^{-\lambda' x}$. Then, $p = \frac{1 - e^{-\lambda' t}}{\lambda' t}$,

$$B_t \sim \text{Pois}(\frac{\lambda}{\lambda'}(1 - e^{-\lambda' t}))$$

As $t \to \infty$, $\mathbb{E}(B_t) = \frac{\lambda}{\lambda'}$.

Claim 10.3 (Littles law)

The long-run average number of customers in a stable system is the long-term average arrival rate multiplied by the average time a customer spends in the system.

10.4 Superposition

10.4.1 Independent Poisson variables: $\sum_{i=1}^n Y_i \sim \text{Poi}\left(\sum_{i=1}^n \lambda_i\right)$

Lemma 10.3

If $Y_1, Y_2, \ldots, Y_n$ *are independent Poisson random variables with means* $\lambda_1, \lambda_2, \ldots, \lambda_n$ *then*

$$\sum_{i=1}^{n} Y_i \sim \text{Poi}\left(\sum_{i=1}^{n} \lambda_i\right)$$

Proof 10.3

There are various ways to prove this, none of them especially hard. For instance, you can use (1) probability generating functions. Alternatively, you can do (2) a direct calculation of the probability mass function when $n = 2$*, and then induct on* n*. But the clearest way to see that this theorem must be true is to use the Law of Small Numbers. (3) Consider, for definiteness, the case* $n = 2$*. Consider independent Bernoulli trials* X_i*, with small success probability* p*. Let* $N_1 = \left\lfloor \frac{\lambda_1}{p} \right\rfloor$ *and* $N_2 = \left\lfloor \frac{\lambda_2}{p} \right\rfloor$*. Clearly we have* $\sum_{i=1}^{N_1} X_i \sim \text{Bin}(N_1, p)$, $\sum_{i=N_1+1}^{N_2} X_i \sim \text{Bin}(N_2, p)$ *and* $\sum_{i=1}^{N} X_i \sim \text{Bin}(N, p)$ *where* $N = N_1 + N_2$.

The Law of Small Numbers implies that when p is small and N_1, N_2 and N are correspondingly large, the three sums above have distributions which are close to Poisson, with means λ_1, λ_2 and λ respectively.

10.4.2 Superposition Theorem: PP with λ_1 + PP with λ_2 = PP with $\lambda_1 + \lambda_2$

Theorem 10.3 (Superposition Theorem)

Let $\{N_1(t), t > 0\}$ and $\{N_2(t), t > 0\}$ be independent Poisson processes with rates λ_1 and λ_2 respectively. Then the combined process $N(t) = N_1(t) + N_2(t)$ is a Poisson process with rate $\lambda_1 + \lambda_2$. ♡

Proof 10.4

Lets verify the properties in the definition of Poisson process.

For all $t \geq 0, N_1(t) \sim \text{Pois}(\lambda_1 t)$ and $N_2(t) \sim \text{Pois}(\lambda_2 t)$, independently, so $N(t) \sim$ Pois $(\lambda_1 t + \lambda_2 t)$ by previous Lemma. The same argument applies for any interval of length t, not just intervals of the form $(0, t]$. Arrivals in disjoint intervals are independent in the combined process because they are independent in the two individual processes, and the individual processes are independent of each other.

Remark: If $X \sim Expo(\lambda_1)$ and $Y \sim Expo(\lambda_2)$, $\min\{X, Y\} \sim Expo(\lambda_1 + \lambda_2)$.

10.4.3 Probability of type 1 event before type 2 event: $\frac{\lambda_1}{\lambda_1+\lambda_2}$

Lemma 10.4

Let X, Y be i.i.d. $Expo(\lambda)$, $\frac{X}{X+Y} \sim Unif(0, 1)$. ♡

Proof 10.5

X, Y can be seen as S_1, S_2 of a Poisson process with λ. Then given $S_1 + S_2$, S_1 follows $Unif(0, S_1 + S_2)$. Hence, $\frac{S_1}{S_1+S_2} \sim Unif(0, 1)$.

Theorem 10.4 (Probability of type 1 event before type 2 event)

If independent Poisson processes of rates λ_1 and λ_2 are superposed, the probability of a type 1 event before a type 2 event in the combined Poisson process is $\frac{\lambda_1}{\lambda_1+\lambda_2}$. ♡

Proof 10.6

Let T be the time until the first type 1 event and let V be the time until the first type 2 event. Let $\tilde{T} = \lambda_1 T$ and $\tilde{V} = \lambda_2 V$ and i.i.d. $Expo(1)$. According to the lemma, we know $U = \frac{\tilde{T}}{\tilde{T}+\tilde{V}} \sim Unif(0, 1)$. We have

$$P(T \leq V) = P\left(\frac{\tilde{T}}{\lambda_1} \leq \frac{\tilde{V}}{\lambda_2}\right) = P\left(\frac{\tilde{T}}{\tilde{T}+\tilde{V}} \leq \frac{\lambda_1}{\lambda_2}\frac{\tilde{V}}{\tilde{T}+\tilde{V}}\right) = P(U \leq \frac{\lambda_1}{\lambda_1+\lambda_2}) = \frac{\lambda_1}{\lambda_1+\lambda_2}$$

Corollary 10.2

If independent Poisson processes of rates λ_1 and λ_2 are superposed, the type of arrival is i.i.d with $P(\text{type } 1) = \frac{\lambda_1}{\lambda_1+\lambda_2}$ *and* $P(\text{type } 2) = \frac{\lambda_2}{\lambda_1+\lambda_2}$.

This yields an alternative path to simulate a superposition of two Poisson processes: we can first generate an $Expo(\lambda_1 + \lambda_2)$ r.v. to decide when the next arrival occurs, and then independently flip a coin with probability $\frac{\lambda_1}{\lambda_1+\lambda_2}$ of heads to decide what kind of arrival it is.

Example 10.2 (Competing risks) The lifetime of Freds refrigerator is $Y_1 \sim \text{Expo}(\lambda_1)$ and the lifetime of Freds dishwasher is $Y_2 \sim \text{Expo}(\lambda_2)$, independent of Y_1. Show that $\min\{Y_1, Y_2\}$ the time of the first appliance failure, is independent of $I(Y_1 < Y_2)$, the indicator that the refrigerator failed first.

Solution *There is an entire Poisson process of refrigerator failures with rate λ_1 and a Poisson process of dishwasher failures with rate λ_2. We can interpret Y_1 as the waiting time for the first arrival in the refrigerator process and Y_2 as the waiting time for the first arrival in the dishwasher process.*

Furthermore, $\min\{Y_1, Y_2\}$ is the waiting time for the first arrival in the superposition of the two Poisson processes, and $I(Y_1 < Y_2)$ is the indicator of this arrival being a type 1 event. We know $\min\{Y_1, Y_2\} \sim \text{Expo}(\lambda_1 + \lambda_2)$ and $P(Y_1 < Y_2) = \frac{\lambda_1}{\lambda_1+\lambda_2}$. Now consider the conditional probability

$$P(Y_1 < Y_2 \mid \min\{Y_1, Y_2\} > t) = P(Y_1 < Y_2 \mid Y_1 > t, Y_2 > t)$$

Given $Y_1 > t$ and $Y_2 > t$ by memorylessness, the additional waiting times after t are also independent exponentials and hence the above conditional probability would again equal $\frac{\lambda_1}{\lambda_1+\lambda_2}$. This shows that ***the waiting times*** *and* ***event types*** *in a superposed Poisson process can be generated completely independently!.*

10.5 Thinning: PP can be divided into two independent PP

The third property of Poisson processes is thinning: if we take a Poisson process and, for each arrival, independently flip a coin to decide whether it is a type-1 event or type-2 event, we end up with two independent Poisson processes. This is the converse of superposition.

Lemma 10.5 (Thinning Property of Poisson)

Suppose that

(1) $N \sim Poi(\lambda)$.

(2) $X_i, i = 1, 2, ...$ are independent, identically distributed $Bernoulli(p)$ random variables independent of N

Then the sum $S_N = \sum_{i=1}^{N} X_i$ follows $Poi(\lambda p)$. Similarly, $N - S_N$ follows $Poi(\lambda(1-p))$. Moreover,

S_N and $N - S_N$ are independent. ♡

It says that if for each occurrence counted in N you toss a p coin, and then record only those occurrences for which the coin toss is a Head, then you still end up with a Poisson random variable.

Proof 10.7

$$\begin{aligned} P(S_N = k) &= \sum_{n=k}^{\infty} P(S_N = k | N = n) P(N = n) \\ &= \sum_{n=k}^{\infty} \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k} \frac{\lambda^n e^{-\lambda}}{n!} \\ &= \frac{(\lambda p)^k e^{-\lambda p}}{k!} \sum_{n=k}^{\infty} \frac{[\lambda(1-p)]^{n-k} e^{-\lambda(1-p)}}{(n-k)!} \\ &= \frac{(\lambda p)^k e^{-\lambda p}}{k!} \end{aligned}$$

From the pmf, we can conclude $S_N \sim Poi(\lambda p)$.

Similarly, we can get $P(N - S_N)$ by switching p to $1 - p \Rightarrow N - S_N \sim Poi(\lambda(1-p))$.

$$\begin{aligned} P(S_N = k_1 | N - S_N = k_2) &= \frac{P(S_N = k_1, N = k_1 + k_2)}{P(N - S_N = k_2)} \\ &= \frac{P(S_{k_1+k_2} = k_1) P(N = k_1 + k_2)}{P(N - S_N = k_2)} \\ &= \frac{\frac{(k_1+k_2)! p^{k_1} (1-p)^{k_2}}{k_1! k_2!} \frac{\lambda^{k_1+k_2} e^{-\lambda}}{(k_1+k_2)!}}{\frac{[\lambda(1-p)]^{k_2} e^{-\lambda(1-p)}}{k_2!}} \\ &= \frac{(\lambda p)^{k_1} e^{-\lambda p}}{k_1!} \end{aligned}$$

Hence, S_N and $N - S_N$ are independent.

Theorem 10.5

Let $N(t)$ be a Poisson process with rate λ, and we classify each arrival in the process as a type 1 event with probability p and a type 2 event with probability $1 - p$, independently. Then

(1) the type 1 events form a Poisson process with rate λp,

(2) the type 2 events form a Poisson process with rate $\lambda(1 - p)$,

(3) and these two processes are independent. ♡

Proof 10.8

In an interval of length T, $N(T) \sim Poi(T\lambda)$ which can be separated into number of event 1 $N_1(T) \sim Poi(T\lambda p)$ and number of event 2 $N_2(T) \sim Poi(T\lambda(1-p))$. $N_1(T)$ and $N_2(T)$ are independent. Since $N(t)$ is menmoryless, so are $N_1(t)$ and $N_2(t)$. We can infer $N_1(t)$ is PP with λ, $N_2(t)$ is PP with $1 - \lambda$, and they are independent.

Example 10.3 Birthday Problem Consider a sequential variant of the birthday problem where people enter a room one by one. Let K be the number of people in the room when for the first time two people share the same birthday? We want to calculate the mean and standard deviation of K.

Solution *Consider a continuous-time version of the question. People enter a room according to Poisson process N_t with rate $\lambda = 1$. Each person is independently marked with one of 365 birthdays, where all birthdays are equally likely. The procedure creates 365 thinned Poisson processes, one for each birthday.*

For $k = 1, ..., 365$, let Z_k be the time when the second person marked with birthday k enters the room. Then, the first time two people in the room have the same birthday is $T = \min_{1 \le k \le 365} Z_k$. Z_ks follow $Gamma(2, \frac{1}{365})$ and are independent of each other. Then, the cdf of T is

$$P(T \le t) = 1 - P(T > t) = 1 - P(Z_k > t)^{365}$$

10.6 Variants of Poisson process

10.6.1 Spatial Poisson Process (dimension ≥ 2)

Poisson processes in multiple dimensions are defined analogously to the 1D Poisson process: we just replace the notion of length with the notion of area or volume. For concreteness, we will now define 2D Poisson processes, after which it should also be clear how to define Poisson processes in higher dimensions.

Definition 10.5 (2D Poisson Process)

Events in the 2D plane are considered a 2D Poisson process with intensity λ if

1. *the number of events in a region A is distributed* Pois(λarea(A))
2. *the numbers of events in disjoint regions are independent of each other.*

♣

Conditionally Uniform: As one might guess, conditioning, superposition, and thinning properties apply to 2D Poisson processes. Let N(A) be the number of events in a region A, and let $B \subset A$. Given $N(A) = n$, the conditional distribution of $N(B)$ is Binomial:

$$N(B) \mid N(A) = n \sim \text{Bin}\left(n, \frac{\text{Area}(A)}{\text{Area }(B)}\right).$$

Conditional on the total number of events in the larger region A, the probability of an event falling into a subregion is proportional to the area of the subregion; thus the locations of the events are conditionally Uniform, and we can generate a 2D Poisson process in A by first generating the number of events $N(A) \sim$ Pois(λ area (A)) and then placing the events uniformly at random in A.

Superposition and Thinning: As in the 1D case, the superposition of independent 2D Poisson processes is a 2D Poisson process, and the intensities add. We can also thin a 2D Poisson process to get independent 2D

Poisson processes.

10.6.2 Non Homogeneous Poisson Process

Arrivals may be more or less likely at certain times. This is not captured by the Poisson Process model. To allow this, we can let the rate parameter λ vary over time.

Definition 10.6 (Non Homogeneous Poisson Process (NHPP))

A counting process Nt is a Non Homogenous Poisson Process (NHPP) with intensity function $\lambda(t)$ if

1. $N_0 = 0$.
2. *For any $t > 0$, N_t has Poisson distribution with mean $\mathbb{E}N_t = \int_0^t \lambda(x)dx$. In general,*
$$N_{t+s} - N_s \sim Poi\left(\int_s^{t+s} \lambda(x)dx\right)$$
3. *For any $0 \leq q < r \leq s < t$, counts in disjoint intervals $N_r - N_q$ and $N_t - N_s$ are independent.*

Remark: NHPP has independent but not stationary increments.

Chapter 11 Brownian Motion

11.1 Brownian Motion

11.1.1 Definition

BM is a stochastic process that models random continuous motion.

Definition 11.1 (Brownian Motion)

A Brownian Motion (BM) or a Weiner process with variance parameter σ^2 is a stochastic process X_t taking values in real numbers satisfying

1. $X_0 = 0$.
2. ***Independent Increments:*** *For any $s_1 < t_1 < s_2 < t_2 \cdots < s_n < t_n$, the random variables $X_{t_1} - X_{s_1}$,...,$X_{t_n} - X_{s_n}$ are independent.*
3. ***Stationary Normal Increments:*** *For any $s < t$, the random variable $X_t - X_s$ has a normal distribution with mean 0 and variance $(t-s)\sigma^2$.*

$$X_t - X_s \sim \mathcal{N}(0, (t-s)\sigma^2)$$

which also shows $X_t \sim \mathcal{N}(0, t\sigma^2), \forall t \in \mathbb{R}_+$.

4. *The paths are **continuous**, i.e, the function $X_t : \mathbb{R}_+ \to \mathbb{R}$ is a continuous function of t.*

♣

Example 11.1 For $0 < s < t$, find the distribution of $B_s + B_t$.

Solution *Since $B_s \sim \mathcal{N}(0, s\sigma^2), B_t - B_s \sim \mathcal{N}(0, (t-s)\sigma^2)$ are two independent normal distributions, $B_s + B_t = 2B_s + (B_t - B_s)$ is also normal.*

$$\mathbb{E}[B_s + B_t] = 0$$

$$\text{Var}[B_s + B_t] = 4\text{Var}(B_s) + \text{Var}(B_t - B_s) = 4s\sigma^2 + (t-s)\sigma^2 = (t+3s)\sigma^2$$

Hence, $B_s + B_t \sim \mathcal{N}(0, (t+3s)\sigma^2)$.

11.1.2 Sufficient Condition for BM

Proposition 11.1

If a stochastic process X has continuous paths and stationary, independent increments, then X is a Brownian motion.

♠

Note: Normality of the increments can be given be the assumptions freely.

11.1.3 Standard Brownian Motion and Transformations

Definition 11.2

Standard Brownian Motion (SBM) is a BM with $\sigma^2 = 1$.

Claim 11.1 (Transformation)

1. $X_0 = x$: *We can also speak of a BM starting from* x*; this is a process satisfying conditions 2 to 4 in the above definition along with the initial condition* $X_0 = x$*. If* X_t *is a SBM then the process* $Y_t = X_t + x$ *is a BM starting at* x.
2. $X_t - X_s \sim \mathcal{N}((t-s)\mu, (t-s)\sigma^2)$: *We can also speak of Brownian Motion with drift* μ*. If* X_t *is a SBM and* $Y_t = X_t + \mu t$ *then* Y_t *is a BM with drift* μ*. We can refer to a* (μ, σ^2) *BM as a Brownian motion where the mean and variance increases at rate* μ *and* σ^2 *per second, respectively.*

11.1.4 Brownian Motion as a limit of Random Walk

Let $X_1, X_2, ...$ be an i.i.d. sequence with each X_i taking values ± 1 with probability $\frac{1}{2}$ each. Set $S_0 = 0$ and for any integer $t > 0$, let $S_t = X_1 + \cdots + X_t$. Then, $S_0, S_1, S_2, ...$ is a simple symmetric random walk with $\mathbb{E}(S_t) = 0$ and $\text{Var}(S_t) = t$ for $t = 0, 1, ...$ As a sum of i.i.d. random variables, for large t, S_t is approximately normally distributed by the central limit theorem.

Let's fix the time interval $[0, 1]$. Let's take n random walk steps of size $\pm\delta$ at time gaps of $\frac{1}{n}$. This defines a piecewise linear continuous function or rather a distribution over continuous functions C_0. We should have the variance of the process at time 1 be 1. Therefore, we should take $\delta = \frac{1}{\sqrt{n}}$.

Now we can imagine letting $n \to \infty$. For each n, we get a distribution over the space C_0. This sequence of distributions converges to a limiting distribution which is precisely the Brownian Motion.

11.2 Gaussian Process

BM is a special case of Gaussian process.

Definition 11.3 (Gaussian Process)

Two equivalent characterizations:

1. W *is a Gaussian process if* $(W(t_1), ..., W(t_n))$ *has a joint normal distribution for all numbers* n *and times* $t_1, ..., t_n$.
2. W *is a Gaussian process if* $a_1 W(t_1) + \cdots + a_n W(t_n)$ *is normally distributed for all* $t_1, ..., t_n$ *and real numbers* $a_1, ..., a_n$.

Claim 11.2

For standard Brownian motion W, $\text{Cov}(W_s, W_t) = \min\{s, t\}$.

Proof 11.1

$\text{Cov}(W_s, W_t) = \text{Var}(W_s) = s, \forall s \leq t$

Proposition 11.2

If a Gaussian process having (1). Continuous paths, (2). Mean 0*, and (3). Covariance function* $\text{Cov}(W_s, W_t) = \min\{s, t\}$ *is a* ***standard Brownian motion****.*

This characterization of Brownian motion can be a convenient and powerful tool and can be used to show that other transformed processes are also BM.

11.3 Transformations and Properties

Lemma 11.1

Let $(B_t)_{t\geq 0}$ be a standard Brownian motion. Then, each of the following transformations is a standard Brownian motion.

1. ***Rescaling:*** *For any $a > 0$,*
$$\frac{1}{\sqrt{a}} B_{at}$$
2. ***Time Inversion:*** *The process $(X_t)_{t\geq 0}$ defined by $X_0 = 0$ and $X_t = tB_{\frac{1}{t}}$ for $t > 0$.*

Proposition 11.3

Suppose that W is a standard Brownian motion, and let $c > 0$. Define $X(t) = W(c + t) - W(c)$. Then $\{X(t) : t \geq 0\}$ is a standard Brownian motion that is independent of $\{W(t) : 0 \leq t \leq c\}$.