

# Notes of Probability

Wenxiao Yang\*

\*Department of Mathematics, University of Illinois at Urbana-Champaign

Last updated: 2022.09

## Contents

<b>1</b>	<b>Distribution</b>	<b>5</b>
1.1	Discrete . . . . .	5
1.1.1	Bernoulli Distribution – Bernoulli( $\pi$ ): an event happens with probability $\pi$ . .	5
1.1.2	Binomial distribution – $\text{bin}(n, \pi)$ : $n$ independent Bernoulli distributions . . . .	5
1.1.3	Multinomial Distribution . . . . .	5
1.1.4	Poisson Distribution – $\text{Pois}(\lambda)$ : an event happens $k$ times within unit time . .	5
1.2	Continuous . . . . .	6
1.2.1	Exponential distribution $\text{Exp}(\lambda)$ : interval between to independent identical event / the first time a event happened . . . . .	6
1.2.2	Gaussian/Normal Distribution . . . . .	7
1.2.3	Multivariate/Joint Gaussian/Normal Distribution (MVN) . . . . .	8
1.3	Poisson process: A sequence of arrivals in continuous time with rate $\lambda$ . . . . .	9
1.3.1	Definition . . . . .	9
1.3.2	$T_j$ : time of $j^{\text{th}}$ arrival . . . . .	9
1.3.3	Theorem (Conditional counts): $N(t_1) N(t_2) = n \sim \text{Bin}(n, \frac{t_1}{t_2})$ . . . . .	9
<b>2</b>	<b>Basis</b>	<b>10</b>
2.1	Covariance and Variance . . . . .	10
2.2	Conditional Expectation and Variance . . . . .	10
2.3	Gambler's Ruin . . . . .	10

2.4	Moment Generating Function (MGF)	11
2.5	Inequality	11
2.5.1	Cauchy-Schwarz inequality: $ \mathbb{E}XY  \leq \sqrt{\mathbb{E}X^2 \cdot \mathbb{E}Y^2}$	11
2.5.2	Jensen's Inequality: convex $g \Rightarrow \mathbb{E}(g(X)) \geq g(\mathbb{E}(X))$	12
2.5.3	Markov's Inequality: $P( X  \geq a) \leq \frac{\mathbb{E} X }{a}$	12
2.5.4	Chebychev's inequality: $P( X - \mu  \geq a) \leq \frac{\sigma^2}{a^2}$	12
2.5.5	Chernoff Inequality: $P(X \geq a) \leq \frac{\mathbb{E}e^{tX}}{e^{ta}}$	12
2.6	Law of Large Numbers (LLN)	12
2.6.1	Weak Law of Large Numbers (wLLN)	12
2.6.2	Strong Law of Large Numbers (sLLN)	13
2.6.3	Differences between <u>convergence in probability</u> (wLLN) and <u>wp1(a.s.)</u> (sLLN)	13
2.7	Central Limit Theorem (CLT)	14
<b>3</b>	<b>Markov Chain</b>	<b>15</b>
3.1	Definition	15
3.2	Matrix Computations	15
3.2.1	Chapman Kolmogorov Equations (C-K Equations) $P(X_{n+m} = j   X_0 = i) =$ $(P^{m+n})_{ij} = \sum_{k \in S} (P^m)_{ik} (P^n)_{kj}$	16
3.2.2	Marginal Distribution $P(X_n = j) = (\alpha P^n)_j$	16
3.3	States, Class	16
3.3.1	Irreducible, Reducible	16
3.3.2	Recurrent, Transient	16
3.4	Periodicity	17
3.4.1	Lemma: all states in an irreducible MC have the same period	18
3.4.2	Periodic, Aperiodic	18
3.5	Regular Matrix	18
3.5.1	Regular matrix: $\exists n \geq 1$ s.t. $P^n > 0$	18
3.5.2	Lemma: Finite MC is Irreducible, Aperiodic $\Leftrightarrow$ has Regular transition matrix	18
3.6	Long Run Behavior of Finite Markov Chains	18
3.6.1	Limiting Distribution	19
3.6.2	Stationary Distribution	19
3.6.3	Limiting Distribution = Expected Proportion of time in each state	20

3.6.4	Fundamental Theorem for <u>Irreducible, Aperiodic, Finite MC</u> (Regular transition matrix) $\Rightarrow \exists$ unique limiting distribution $\pi$ and $\pi_j > 0, \forall j$ . . . . .	20
3.6.5	Long run behavior for reducible and/or periodic chains . . . . .	20
3.6.6	Fundamental Theorem for <u>Irreducible, Finite MC</u> : expected first return time $\mathbb{E}(T_j X_0 = j) = \frac{1}{\pi_j}$ . . . . .	21
3.7	Return Times and Absorption Probabilities . . . . .	22
3.7.1	Expected Number of Visits to a Transient State: $E(Y_i X_0 = j) = M_{ji} = (I - Q)_{ji}^{-1}$	22
3.7.2	Expected Time till Absorption to a Recurrent Class: $\mathbb{E}(T_{abs} X_0 = j) = \sum_{i \in T_1 \cup T_2 \cup \dots \cup T_s} M_{ji}$	23
3.7.3	Expected first return time (different initial state) = Time till Absorption . . . .	24
3.7.4	Probability of Eventually Entering a Given Recurrent Class: $A = (I - Q)^{-1}S = MS$ . . . . .	25
3.8	Examples of Finite MC . . . . .	26
3.8.1	Gambler's Ruin . . . . .	26
3.8.2	Simple Random Walk (SRW) on Undirected Graph . . . . .	27
<b>4</b>	<b>Countably infinite MC</b>	<b>28</b>
4.1	Recurrence and Transience . . . . .	28
4.1.1	Recurrent or Transient State . . . . .	28
4.1.2	Recurrent or Transient Class . . . . .	29
4.1.3	Lemma: Transient Class $\Leftrightarrow \sum_{n=0}^{\infty} P_{i,i}^n < \infty$ . . . . .	29
4.1.4	Recurrence/Transience of Simple Random Walk on Lattice . . . . .	30
4.1.5	Null and Positive Recurrence . . . . .	30
4.1.6	Stationary Distribution and Limiting Distribution . . . . .	30
4.2	Differences between Finite and (Countably) Infinite Markov Chains . . . . .	31
<b>5</b>	<b>Branching Process</b>	<b>32</b>
5.1	Extinction Probability in a Branching Process . . . . .	32
5.1.1	Expectation $\mathbb{E}X_n = \mu^n \mathbb{E}X_0$ . . . . .	32
5.1.2	Lemma: $\mu < 1 \Rightarrow P(\text{extinction}) = 1$ . . . . .	33
5.1.3	Variance: $VarX_n = \begin{cases} n\sigma^2, & \mu = 1 \\ \sigma^2 \mu^{n-1} \frac{\mu^n - 1}{\mu - 1}, & \mu \neq 1 \end{cases}$ . . . . .	33
5.1.4	Extinction probability $\rho = 1$ if $\mu \leq 1$ ; $\rho < 1$ if $\mu > 1$ . . . . .	34

5.1.5	$G_n(s) = G_{n-1}(\psi(s)) = \psi(\psi(\psi(\cdots \psi(s) \cdots))) = \psi(G_{n-1}(s))$	36
<b>6</b>	<b>Time Reversible Markov Chains</b>	<b>36</b>
6.1	Definition: Local Balance $\pi(i)P(i, j) = \pi(j)P(j, i), \forall i, j \in S$	36
6.2	Discussion about Local Balance	37
6.2.1	Flow: $Flow(A, B) = \sum_{i \in A} \sum_{j \in B} \pi(i)P_{ij}$	37
6.2.2	Lemma: $Flow(A, A^c) = Flow(A^c, A)$ for any subset $A \subset S$	37
6.2.3	Lemma: Local balance $\Rightarrow \pi$ is stationary	37
6.2.4	Lemma: All stationary birth and death chains are reversible	37
6.3	Example: Random Walk on an Undirected Graph	38
<b>7</b>	<b>Markov Chain Monte Carlo (MCMC)</b>	<b>38</b>
7.1	Strong Law of Large Numbers for Markov Chains	39
7.2	Example of Designing MC	39
7.3	Metropolis Hastings Algorithm	40

# 1 Distribution

## 1.1 Discrete

### 1.1.1 Bernoulli Distribution – Bernoulli( $\pi$ ): an event happens with probability $\pi$

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

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

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

### 1.1.2 Binomial distribution – $\text{bin}(n, \pi)$ : $n$ independent Bernoulli distributions

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

$$P(y) = \binom{n}{y} \pi^y (1 - \pi)^{n-y}$$

with  $\binom{n}{y} = \frac{n!}{y!(n-y)!}$  and  $y = 0, 1, 2, \dots, n$

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

### 1.1.3 Multinomial Distribution

### 1.1.4 Poisson Distribution – $\text{Pois}(\lambda)$ : an event happens $k$ times within unit time

$\lambda$ : frequency of the event, i.e., the average number of event happens within unit time.

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

$$E(X) = \text{Var}(X) = \lambda$$

### Derivation process:

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

the probability the number of the event happens:

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

### 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 parameter is the sum of  $X$ 's and  $Y$ 's

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

Then,

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

## 1.2 Continuous

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

$\lambda$ : frequency of the event.

$X$  follows exponential distribution with parameter  $\lambda$  or  $\beta$ :

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

They are equivalent, 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 frequency of the occurrence of the event;  $\beta > 0$  is the probability of the event happens in each second. The range of exponential distribution is  $[0, \infty)$ .

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

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

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

CDF:

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

PDF:

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

### 1.2.2 Gaussian/Normal Distribution

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

**Theorem 1.** Suppose  $X \sim N(\mu_1, \sigma_1^2)$  and  $Y \sim N(\mu_2, \sigma_2^2)$  are independent, then  $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$ .

*Proof.* The MGF of  $X$  is

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

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  $Z \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$  □

### 1.2.3 Multivariate/Joint Gaussian/Normal Distribution (MVN)

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

p.d.f.

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

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)  $\boldsymbol{\mu}$  is a  $k$ -dimensional **mean vector**:

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

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

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

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

**Theorem 2.** *MVN distribution is completely specified by knowing  $\boldsymbol{\mu}$ ,  $\Sigma$ .*

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

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

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

□



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

**Theorem 3.** For *MVN*, "independence" is **equivalent** to "0 correlation"

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

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

□

**Theorem 4.** Independent  $X = [X_1, X_2, \dots, X_n] \sim MVN$  and  $Y = [Y_1, Y_2, \dots, Y_m] \sim MVN$ , then  $W = [X_1, X_2, \dots, X_n, Y_1, Y_2, \dots, Y_m] \sim MVN$ .

**Theorem 5.** Independent  $X = [X_1, X_2, \dots, X_n] \sim N(\mu_1, \Sigma_1)$  and  $Y = [Y_1, Y_2, \dots, Y_n] \sim N(\mu_2, \Sigma_2)$ , then  $X + Y \sim N(\mu_1 + \mu_2, \Sigma_1 + \Sigma_2)$ .

### 1.3 Poisson process: A sequence of arrivals in continuous time with rate $\lambda$

#### 1.3.1 Definition

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

$$\begin{aligned} N(t) &\sim Pois(\lambda t) \\ \Pr(N(t) = k) &= \frac{(\lambda t)^k e^{-\lambda t}}{k!} \end{aligned}$$

The number of arrivals in disjoint time intervals are independent.

#### 1.3.2 $T_j$ : time of $j^{th}$ arrival

$T_1 > t$  is same as  $N(t) = 0$ :  $P(T_1 > t) = P(N(t) = 0) = e^{-\lambda t}$   
 $\Rightarrow T_1 \sim Expo(\lambda) \Rightarrow T_j - T_{j-1} \sim Expo(\lambda); T_j \sim Gamma(j, \lambda)$

#### 1.3.3 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  $\frac{t_1}{t_2}$ )

## 2 Basis

### 2.1 Covariance and Variance

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

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

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

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

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

### 2.2 Conditional Expectation and Variance

(1) Conditional variance:

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

(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):**

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

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

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

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

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

### 2.3 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.

$$\begin{aligned} P(A|x=k) &= \frac{1}{2}P(A|x=k-1) + \frac{1}{2}P(A|x=k+1) \\ &\Rightarrow p_k - p_{k-1} = p_{k+1} - p_k \end{aligned}$$

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

## 2.4 Moment Generating Function (MGF)

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

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

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

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

**Why MGF is useful?**

(1) If  $X, Y$  are independent, then

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

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

## 2.5 Inequality

**2.5.1 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 1** (Second 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,

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

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

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

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

**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.

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

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

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

### 2.5.5 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.*  $P(X \geq a) = P(e^{tX} \geq e^{ta}) \leq \frac{\mathbb{E}e^{tX}}{e^{ta}}$  □

## 2.6 Law of Large Numbers (LLN)

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

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

### 2.6.1 Weak Law of Large Numbers (wLLN)

**Theorem 6** (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.*

$$\bar{X}_n \xrightarrow{P} \mu \quad \text{when } n \rightarrow \infty.$$

That is, for any positive number  $\varepsilon$ ,

$$\lim_{n \rightarrow \infty} \Pr(|\bar{X}_n - \mu| < \varepsilon) = 1.$$

*Proof.* Prove by Chebychev's inequality.

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

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

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

□

## 2.6.2 Strong Law of Large Numbers (sLLN)

**Theorem 7** (Strong Law of Large Numbers (sLLN)).

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

$$\bar{X}_n \xrightarrow{\text{a.s.}} \mu \quad \text{when } n \rightarrow \infty.$$

*That is,*

$$\Pr\left(\lim_{n \rightarrow \infty} \bar{X}_n = \mu\right) = 1.$$

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

a) Weak Law of Large Numbers (wLLN)

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

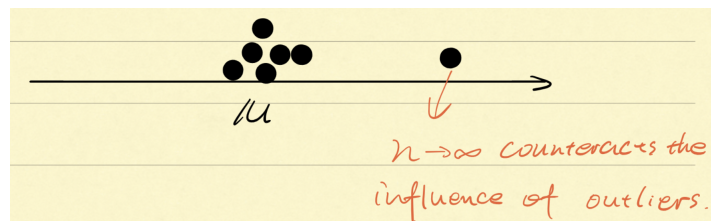


Figure 1: convergence in probability

b) Strong Law of Large Numbers (sLLN)

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

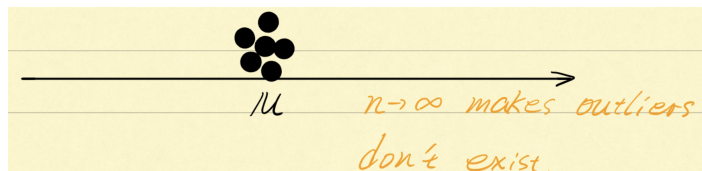


Figure 2: wp1(a.s.)

## 2.7 Central Limit Theorem (CLT)

**Theorem 8** (Central Limit Theorem (CLT)).

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

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

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

*Proof.* 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  $\sqrt{n}\bar{X}$ :

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

$$\begin{aligned} \lim_{n \rightarrow \infty} \log M_1(t) &= \lim_{n \rightarrow \infty} n \log M_0(\frac{t}{\sqrt{n}}) \\ &\quad (\text{let } y = \frac{1}{\sqrt{n}}) \\ &= \lim_{y \rightarrow 0} \frac{\log M_0(yt)}{y^2} \\ &\quad (\text{L'Hôpital's rule}) \\ &= \lim_{y \rightarrow 0} \frac{tM'_0(yt)}{2yM_0(yt)} \\ &\quad (\text{L'Hôpital's rule}) \\ &= \lim_{y \rightarrow 0} \frac{t^2 M''_0(yt)}{2M_0(yt) + 2ytM'_0(yt)} \\ &= \frac{t^2}{2} \end{aligned}$$

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.  $\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \xrightarrow{D} N(0, 1)$  as  $n \rightarrow \infty$  □

### 3 Markov Chain

#### 3.1 Definition

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

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

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

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**.

#### 3.2 Matrix Computations

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

**Lemma 1** (Distribution of Entire Sequence).

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

**Lemma 2** (Markov Property).

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

**Lemma 3** (Transition Probability after  $n$  states).

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

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

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

**3.2.1 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  $j$ :

$$P(X_{n+m} = j | X_0 = i) = (P^{m+n})_{ij} = \sum_{k \in S} (P^m)_{ik} (P^n)_{kj} \quad (1)$$

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

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

**Lemma 4** (Marginal Distribution). *Given initial distribution  $X_0 \sim \alpha$  and transition matrix  $P$ .  $\alpha$  is distribution vector ( $1 \times |S|$ ) with  $\sum_{i \in S} \alpha_i = 1$ .*

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

**Corollary 1** (Distribution of Subsequence).

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

### 3.3 States, Class

#### 3.3.1 Irreducible, Reducible

- Accessible:  $j$  is accessible from  $i$  if  $\exists n$  s.t.  $P_{ij}^n > 0$ .
- 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.

#### 3.3.2 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  $\sum_{n=0}^{\infty} f_i^n(1 - f_i) = 1 - f_i < \infty$ ).
- 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.** *If  $i$  is recurrent,  $i \leftrightarrow j \Rightarrow j$  is recurrent.*

**Theorem 9.** *The states of a communication class are either all recurrent or all transient.*

**Corollary 2.** *For a finite irreducible Markov chain, all states are recurrent.*

### Canonical Decomposition

**Definition 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 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.* (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. □

### 3.4 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

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

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

**Lemma 7.** *For an irreducible MC, all states have the same period.*

*Proof.* 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  $J_j$ .  $\square$

### 3.4.2 Periodic, Aperiodic

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

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

## 3.5 Regular Matrix

### 3.5.1 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 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  $n \geq 1$ .*

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

**Lemma 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**.

## 3.6 Long Run Behavior of Finite Markov Chains

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

- (1) Convergence. ( $P^{n+1} = P^n$ )
- (2) Forgetting the initial states. (each row is identical)

### 3.6.1 Limiting Distribution

**Definition 3.** A MC is said to have a **limiting distribution**  $\lambda$  if we have

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

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

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

**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  $p+q \neq 1$ ,

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

**Lemma 9.** If  $\lambda$  is the limiting distribution for a MC with transition matrix  $P$  then  $\lambda$  satisfies the equation

$$\lambda P = \lambda$$

*Proof.*  $(\lambda P)_j = \sum_{i \in S} \lambda_i P_{ij} = \sum_{i \in S} \lim_{n \rightarrow \infty} P_{ki}^n P_{ij} = \lim_{n \rightarrow \infty} \sum_{i \in S} P_{ki}^n P_{ij} = \lim_{n \rightarrow \infty} P_{kj}^{n+1} = \lambda_j \quad \square$

### 3.6.2 Stationary Distribution

**Definition 4.** A distribution  $\pi$  which satisfies the equation

$$\pi P = \pi$$

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

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

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

*Proof.* We can write

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

Therefore, taking limits we can conclude that

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

□

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

**Theorem 10.** *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)  $\pi$  is the **limiting distribution** of the MC. ( $\lim_{n \rightarrow \infty} \alpha P^n = \pi, \forall \alpha$  initial distribution)
- (2)  $\pi$  gives **positive** probability to each of the states. ( $\pi_j > 0, \forall j \in S$ )

### 3.6.5 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  $R_1, \dots, R_r$  and transient classes  $T_1, \dots, T_s$ . Each recurrent class acts as a separate MC with transition matrix  $P_1, \dots, P_r$ . Assume each  $P_k$  is aperiodic.

Then by the fundamental theorem, there exists  $r$  different limiting distributions  $\pi^1, \dots, \pi^r$ . The distribution  $\pi^k$  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

$$\lim_{n \rightarrow \infty} P_{ij}^n = \pi^k(j)$$

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,

$$\lim_{n \rightarrow \infty} P_{ij}^n = 0$$

3. Let  $\alpha_k(i)$  for  $k = 1, \dots, r$  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  $j \in R_k$ ,

$$\lim_{n \rightarrow \infty} P_{ij}^n = \alpha_k(i) \pi^k(j)$$

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  $0, 1, \dots, d-1$ . Therefore, there cannot be a limit of  $P^n$ .

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

### 3.6.6 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 11.** Assume that  $X_0, X_1, \dots$  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,  $\mu_j$  is finite, and there exists a unique **positive** stationary distribution  $\pi$  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  $j$ :

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

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

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

**Example 2** (Two State MC). Consider the transition matrix

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

Here, by the theorem

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

### 3.7 Return Times and Absorption Probabilities

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

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 11.** 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.* We can write

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

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

We can also extend the equation:

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

**3.7.2 Expected Time till Absorption to a Recurrent Class:**  $\mathbb{E}(T_{abs} | X_0 = j) = \sum_{i \in T_1 \cup T_2 \cup \dots \cup T_s} M_{ji}$

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.

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

**Corollary 3.** For any transient state  $j \in S$ ,

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

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

### 3.7.3 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  $a_j = \mathbb{E}[T_i|X_0 = j]$

$$\mathbb{E}[T_i|X_0 = j] = P_{ji} \cdot 1 + \sum_{k \neq i} P_{jk} \cdot (1 + \mathbb{E}[T_i|X_0 = k])$$

$$= 1 + \sum_{k \neq i} P_{jk} \cdot \mathbb{E}[T_i|X_0 = k]$$

$$\Rightarrow a_j = 1 + \sum_{k \neq i} P_{jk} \cdot a_k$$

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 4.** *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

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

**3.7.4 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  $r_1, \dots, r_k$  with  $P(r_i, r_i) = 1, \forall i \in \{1, \dots, k\}$ .

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

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

Let  $\alpha_{t_i, r_j}$  be the probability that the  $MC$  starting 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

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

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

$$\begin{aligned} A &= [S \ Q] \begin{bmatrix} I \\ A \end{bmatrix} = S + QA \\ \Rightarrow A &= (I - Q)^{-1}S = MS \end{aligned}$$

(this  $S$  is the submatrix in  $P$ )

### 3.8 Examples of Finite MC

#### 3.8.1 Gambler's Ruin

**Example 5** (Gambler's Ruin). *Consider the asymmetric Gambler's Ruin with winning probability  $p \in (0, 1)$ . The state space is  $\{0, 1, \dots, N\}$ .*

Let  $\alpha_j$  be the probability that the MC get absorbed in state  $N$  starting from state  $j$ . Clearly,

$\alpha(0) = 0, \alpha(N) = 1$ . For any  $0 < j < N$ , we can condition on the first step to get

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

### 3.8.2 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.

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

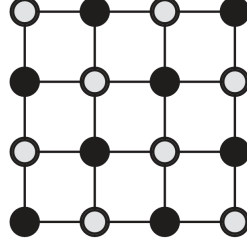


Figure 3: bipartite

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

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

**Example 6.** For Simple Random Walk: assume there exists a stationary distribution  $\pi$ , 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

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

### 4.1 Recurrence and Transience

#### 4.1.1 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 5.** A state  $j$  is **recurrent** if  $f_j = 1$  and **transient** if  $f_j < 1$ .

#### 4.1.2 Recurrent or Transient Class

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

**Lemma 12.** If  $i, j$  are in the same class,  $i$  is recurrent  $\Leftrightarrow j$  is recurrent.

*Proof.* 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,

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

□

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

**Lemma 13.** 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.* 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)

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

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

□

#### 4.1.4 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  $\lim_{n \rightarrow \infty} \frac{n!}{\sqrt{2\pi n} \left(\frac{n}{e}\right)^n} = 1$

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

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

#### 4.1.5 Null and Positive Recurrence

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

**Definition 6.** 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:**  $P(T_i = n) = \frac{1}{n(n+1)} = \frac{1}{n} - \frac{1}{n+1}, n \geq 1$ .

$$\begin{aligned} f_i &= \sum_{n=1}^{\infty} P(T_i = n) = \sum_{n=1}^{\infty} \left( \frac{1}{n} - \frac{1}{n+1} \right) \Rightarrow \text{recurrent} \\ \mu_i &= \sum_{n=1}^{\infty} n P(T_i = n) = \sum_{n=1}^{\infty} \frac{1}{n+1} = \infty \Rightarrow \text{null recurrent} \end{aligned}$$

#### 4.1.6 Stationary Distribution and Limiting Distribution

Limiting distribution

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

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

**Lemma 15.** *For an irreducible MC,  $\lim_{n \rightarrow \infty} P_{y,x}^n = 0$  for each  $x, y \in S$  **if and only if** the chain is transient or null recurrent.*

**Theorem 12** (Fundamental Theorem for General Discrete Markov Chains). *An **irreducible, positive recurrent** MC has a **unique stationary distribution**  $\pi$  (which is positive everywhere) solving the equation*

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

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

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

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

$$\lim_{n \rightarrow \infty} P_{ij}^n = \pi(j)$$

The stationary distribution  $\pi$  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.

## 4.2 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.

## 5 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  $p_0, p_1, \dots$  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  $S = \{0, 1, 2, \dots\} = \mathbb{Z}_+$ . Note that 0 is an absorbing state. Suppose  $X_n = k$ . Then  $k$  individuals produce offspring for the next generation. Let  $Y_1, \dots, Y_k$  be i.i.d random variables with  $P(Y_1 = j) = p_j$ . Then we can write the transition probabilities as

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

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  $\infty$ .

### 5.1 Extinction Probability in a Branching Process

#### 5.1.1 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

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



**5.1.2 Lemma:**  $\mu < 1 \Rightarrow P(\text{extinction}) = 1$

**Lemma 16.** *If  $\mu < 1$ , then probability of extinction is 1.*

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

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

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

Now, the probability of survival

$$\begin{aligned} \lim_{n \rightarrow \infty} P(X_n \geq 1) &\leq \lim_{n \rightarrow \infty} \mathbb{E}X_n = \lim_{n \rightarrow \infty} \mu^n \mathbb{E}X_0 = 0 \\ \Rightarrow \lim_{n \rightarrow \infty} P(X_n \geq 1) &= 0 \end{aligned}$$

Then we can conclude

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

□

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

**5.1.3 Variance:** 
$$\text{Var}X_n = \begin{cases} n\sigma^2, & \mu = 1 \\ \sigma^2 \mu^{n-1} \frac{\mu^n - 1}{\mu - 1}, & \mu \neq 1 \end{cases}$$

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

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

(Assuming  $X_0 = 1$  with probability 1)

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

#### 5.1.4 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  $a(k) = \lim_{n \rightarrow \infty} a_n(k)$  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  $\rho$ .

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

By conditioning on the first step, we can write

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

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

**Definition 7.** If a random variable  $X$  takes values in  $\mathbb{Z}$ , the **probability generating function (pgf)** of  $X$  is the function  $\psi : [0, 1] \rightarrow \mathbb{R}$  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  $\psi$ .

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.  $\psi_{X+Y}(s) = \mathbb{E}(s^{X+Y}) = \mathbb{E}(s^X) \mathbb{E}(s^Y) = \psi_X(s) \psi_Y(s)$

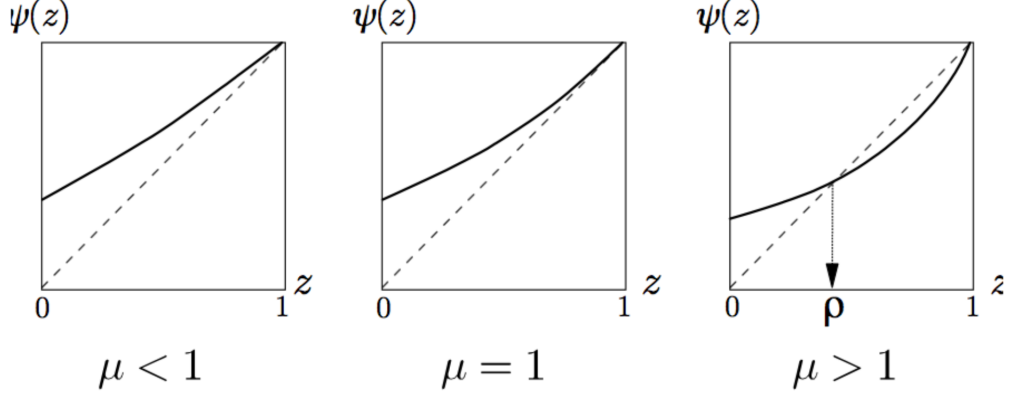


Figure 4: Fixed Point of pgf

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  $\lim_{n \rightarrow \infty} q_n = \rho$ . 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  $q_n \leq r$ ,

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

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

**Theorem 13.** *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.**  $p_0 = \frac{1}{8}, p_1 = \frac{3}{8}, p_2 = \frac{3}{8}, p_3 = \frac{1}{8}$ .

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  $\Rightarrow (r - 1)(r^2 + 4r - 1) = 0$   $r^* = \sqrt{5} - 2$ .

$$\mathbf{5.1.5} \quad G_n(s) = G_{n-1}(\psi(s)) = \psi(\psi(\psi(\cdots \psi(s) \cdots))) = \psi(G_{n-1}(s))$$

For  $n \geq 0$ , let

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

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

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

We have

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

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

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

## 6 Time Reversible Markov Chains

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

**Definition 8.** We say that a MC is **time reversible** if, for each  $n \geq 1$ , the distribution of  $(X_0, \dots, X_n)$  is the same as the distribution of  $(X_n, \dots, X_0)$ . Equivalently, for any  $x_0, \dots, x_n \in \mathcal{S}$  we have

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

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

**Lemma 17** (Local Balance). The Markov chain  $X_0, X_1, \dots$  is time-reversible **if and only if** the distribution  $\pi$  of  $X_0$  satisfies the condition

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

## 6.2 Discussion about Local Balance

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

**Definition 9.** For a distribution  $\pi$  on the state space  $S$  and any two subsets of the state space  $A, B$  define the **Flow**

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

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

**Lemma 18.**  $Flow(A, A^c) = Flow(A^c, A)$  for any subset  $A \subset S$ .

*Proof.*

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

□

**6.2.3 Lemma:** Local balance  $\Rightarrow \pi$  is stationary

**Lemma 19.** If the local balance equations " $\pi(i)P(i, j) = \pi(j)P(j, i), \forall i, j \in S$ " hold then  $\pi$  is stationary.

*Proof.*

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

□

**6.2.4 Lemma:** All stationary birth and death chains are reversible

**Lemma 20.** 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.* It is enough to show the equation holds when  $j = i + 1$ . For  $A = \{0, 1, 2, \dots, i\}$ ,

$$\begin{aligned} Flow(A, A^c) &= \sum_{0 \leq k \leq i} \sum_{j > i} \pi(k) P_{kj} = \pi(i) P_{i, i+1} \\ &= Flow(A^c, A) = \sum_{j > i} \sum_{0 \leq k \leq i} \pi(j) P_{jk} = \pi(i+1) P_{i+1, i} \end{aligned}$$

□

### 6.3 Example: Random Walk on an Undirected Graph

**Lemma 21.** *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.* 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  $W_{ij} \geq 0$ . 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  $\pi$  satisfies the local balance.

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

The right hand side above is symmetric in  $u, v$  so local balance must hold. On the other hand, let's 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

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

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

## 7 Markov Chain Monte Carlo (MCMC)

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

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

## 7.1 Strong Law of Large Numbers for Markov Chains

**Theorem 14.** Assume that  $X_0, X_1, \dots$  is an ergodic Markov chain with stationary distribution  $\pi$ . Let  $r$  be a bounded and real-valued function. Let  $Y$  be a random variable with distribution  $\pi$ . Then, with probability 1,

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

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

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

## 7.2 Example of Designing MC

1. All binary sequence consisted of 0 or 1 of length  $d$ ,  $S = \{0, 1\}^d$ ,  $|S| = 2^d$ .  $\pi$  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**,  $B = \{0, 1\}^d$ .  $\pi$  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  $(x_1, \dots, x_d)$ , 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  $\pi$

The above three facts imply that the uniform distribution  $\pi$  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

$$\frac{f(X_1) + \dots + f(X_n)}{n} \rightarrow \mathbb{E}f(\pi)$$

### 7.3 Metropolis Hastings Algorithm

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

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

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

$$A_{ij} = \min \left\{ 1, \frac{\pi_j T_{ji}}{\pi_i T_{ij}} \right\}.$$

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  $X_{n+1} = i$ .

Box-Muller Method:

Fact:  $X, Y \sim N(0, 1)$ , then define the angle  $\theta = \tan^{-1} \frac{y}{x}$  and the radius's square  $R^2 = x^2 + y^2 \sim X_{(2)}^2$  ( $\exp(2)$ ).