Section Notes for PSTAT 213

Haosheng Zhou

Sept, 2023

Contents

This note contains extra exercises, examples and materials for PSTAT 213. The notes may be subject to typos, and you are welcome to email me at hzhou593@ucsb.edu for any possible advice.

Week 1

Example for Indicator

Lemma 1 (Example)**.** *The indicator I_A of event A is a random variable defined as*

$$I_A(\omega) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{else} \end{cases} \tag{1}$$

(a): Write PMF for $X = I_A$ and calculate $\mathbb{E}X, Var(X), G_X(s)$.

(b): For random variable $Y \geq 0$ and ϕ as any non-negative increasing function on $[0, +\infty)$, show that $\forall a > 0, \phi(a) \cdot \mathbb{P}(Y \geq a) \leq \mathbb{E}\phi(Y)$ so that $\forall \varepsilon > 0, \mathbb{P}(|Z| \geq \varepsilon) \leq \frac{\mathbb{E}Z^2}{\varepsilon^2}$ for any random variable Z.

(c): Assume Y is a random variable such that its MGF $M_Y(t) = \mathbb{E}e^{tY}$ is finite for all $t \in \mathbb{R}$, show that when $t \geq 0$, $\mathbb{P}(X \geq x) \leq e^{-tx} M_X(t)$ so that $\mathbb{P}(X \geq x) \leq \inf_{t \geq 0} e^{-tx} M_X(t)$.

Proof. (a): X has support $\{0, 1\}$ with $\mathbb{P}(X = 1) = \mathbb{P}(A), \mathbb{P}(X = 0) = \mathbb{P}(A^c) = 1 - \mathbb{P}(A)$ gives the PMF.

From the PMF, it's easy to calculate $\mathbb{E}X = \mathbb{P}(A), \mathbb{E}X^2 = \mathbb{P}(A)$ so $Var(X) = \mathbb{E}X^2 - \mathbb{E}^2X = \mathbb{P}(A) - [\mathbb{P}(A)]^2$.

$$G_X(s) = \mathbb{E}s^X = 1 \cdot \mathbb{P}(X = 0) + s \cdot \mathbb{P}(X = 1) = 1 - \mathbb{P}(A) + s \cdot \mathbb{P}(A) \tag{2}$$

(b): This is the classical trick on indicator

$$\forall a > 0, \phi(a) \cdot \mathbb{P}(Y \geq a) = \mathbb{E}[\phi(a)\mathbb{I}_{Y \geq a}] \leq \mathbb{E}[\phi(Y)\mathbb{I}_{Y \geq a}] \leq \mathbb{E}\phi(Y) \tag{3}$$

since ϕ is increasing and indicator is non-negative and takes value no larger than 1.

Consider $\phi(x) = x^2$ non-negative and increasing on $[0, +\infty)$ plugging in $Y = |Z| \geq 0, a = \varepsilon$ to conclude the proof.

(c):

Since for $t > 0$, e^{tx} is non-negative and increasing in x, resulting in

$$\mathbb{P}(X \geq x) = \mathbb{P}\left(e^{tX} \geq e^{tx}\right) \leq e^{-tx}\mathbb{E}e^{tX} = e^{-tx}M_X(t) \tag{4}$$

applying the conclusion in (b) for $Y = e^{tX}, a = e^{tx}, \phi(x) = x$. When $t = 0$, check $e^{-tx}M_X(0) = 1$ so $\mathbb{P}(X \geq x) \leq 1$ naturally holds. This proves that the inequality holds for $\forall t \geq 0$. Taking inf on both sides w.r.t. t concludes the proof.

□

Remark. *Part (c) is a very important technique that will appear once again in 213BC to derive the Chernoff bound of concentration of measures. The basic idea is to* ***introduce some unspecified parameter*** t***, build a bound for the probability and optimize the bound to get the tightest bound by specifying an appropriate value of*** t*.*

Example for Branching Process

Lemma 2 (Example)**.** *There is an isolated island with the original stock of* 100 *family surnames, and the survival of family names is modelled by branching process, different surnames' survivals are independent. Each surname has extinction probability* $\eta = \frac{9}{10}$*.*

(a): After many generations how many surnames do you expect to be on the island?

(b): Do you expect the total population on the island to be increasing or decreasing?

Proof. (a): Each surname has η probability of disappearing independent of other surnames so the number of surname survived after a long enough time denoted X has binomial distribution $X \sim B(100, 1-\eta)$. It's clear that $\mathbb{E}X = 100(1-\eta) = 10$.

(b): Since $\eta > 0, \eta \neq 1$, the branching process $\{Z_n\}$ for each family surname is in the supercritical phase with offspring mean $\mu > 1$. It's clear that $\mathbb{E}Z_n = \mu^n \to +\infty$ $(n \to \infty)$ so the expected total population is increasing.

□

Lemma 3 (Example)**.** *Branching process* $\{Z_n\}$ *originates from one individual, i.e.* $Z_0 = 1$ *has Poisson offspring distribution* $Z_1 \sim P(\lambda)$ $(\lambda > 1)$*. If it's known that a branching process conditional on extinction is still a branching process, i.e. let* A *stands for the event that* $\{Z_n\}$ *extinct,* $\{E_n\} = \{Z_n\}|_A$ *is still a branching process. Can you derive the offspring distribution for* $\{E_n\}$*?*

Proof. Since $E_0 = Z_0|_A = 1$, the offspring distribution for $\{E_n\}$ is just the distribution of E_1. Let's denote η as the extinction probability of $\{Z_n\}$, i.e. $\eta = \mathbb{P}(A)$ and $p_k = \mathbb{P}(Z_1 = k)$ as the offspring distribution PMF of $\{Z_n\}$.

$$\mathbb{P}(E_1 = k) = \mathbb{P}(Z_1 = k|A) = \frac{\mathbb{P}(A|Z_1 = k)\,\mathbb{P}(Z_1 = k)}{\mathbb{P}(A)} \tag{5}$$

using Bayes formula. Notice that conditional on $Z_1 = k$, extinction happens if and only if all k subtrees generated in generation 1 are extinct. Since all k subtrees are independent and follow the same offspring distribution, they have exactly the same probability of being extinct, resulting in

$$\mathbb{P}(A|Z_1 = k) = [\mathbb{P}(A|Z_1 = 1)]^k = [\mathbb{P}(A)]^k = \eta^k \tag{6}$$

where the second equation comes from the fact that if $Z_1 = 1$, restarting the branching process at generation 1 makes no difference to the extinction probability (this is actually the Markov property of branching process). At this point, we see that

$$\mathbb{P}(E_1 = k) = \eta^{k-1} p_k = \eta^{k-1} \frac{\lambda^k}{k!} e^{-\lambda} \tag{7}$$

Since $\lambda > 1$, the offspring mean is larger than 1, the extinction probability η is thus the fixed point of $G(s)$ with

$$G(s) = \mathbb{E}s^{Z_1} = \sum_{k=0}^{\infty} s^k \frac{\lambda^k}{k!} e^{-\lambda} = e^{s\lambda - \lambda} \tag{8}$$

telling us

$$e^{\eta\lambda - \lambda} = \eta \tag{9}$$

turning it into $e^{-\lambda} = \eta e^{-\eta\lambda}$ and replace the $e^{-\lambda}$ term in the expression of $\mathbb{P}(E_1 = k)$ to get

$$\mathbb{P}(E_1 = k) = \frac{(\eta\lambda)^k}{k!} e^{-\eta\lambda}, E_1 \sim P(\eta\lambda) \tag{10}$$

the offspring distribution of $\{E_n\}$ is still Poisson but it's $P(\eta\lambda)$.

□

Remark. *Actually **any branching process conditional on extinction is still a branching process**. Unfortunately, there is no easy approach to prove this conclusion since it's a statement for the whole process but not for pointwise evaluation of the process. Proving this conclusion requires the correspondence between branching process and random walk which we might have the chance to introduce in the future.*

However, we can do heuristic calculations as above to calculate the offspring distribution of the new branching process. From what we have shown above, the new branching process $\{E_n\}$ has offspring distribution with PMF

$$\mathbb{P}(E_1 = k) = p'_k = \eta^{k-1} p_k \tag{11}$$

*this is called **the duality principle of branching process**. In particular, Poisson branching process conditional on extinction still provides a Poisson branching process.*

Lemma 4 (Example)**.** *A branching process $\{Z_n\}$ is given such that $Z_0 = 8$ with offspring distribution PMF $p_0 = 0.2, p_1 = 0.5, p_2 = 0.3$.*

(a): Derive its extinction probability η.

(b): Derive the probability that the process is extinct in generation 3 but survives in generation 1 and generation 2.

Proof. (a): Such branching process is actually the sum of 8 branching process $\left\{Z_n^{(1)}\right\}, ..., \left\{Z_n^{(8)}\right\}$ with the same offspring distribution but with $Z_0^{(1)} = ... = Z_0^{(8)} = 1$. Moreover, those 8 branching processes are independent (by the definition of branching process).

Denote $E_n^{(i)}$ as the event that $\left\{Z_n^{(i)}\right\}$ is extinct in generation n and $S_n^{(i)}$ as the event that $\left\{Z_n^{(i)}\right\}$ survives in generation n, $E^{(i)}$ as the event that $\left\{Z_n^{(i)}\right\}$ is extinct. It's clear that $\{Z_n\}$ is extinct if and only if $\left\{Z_n^{(1)}\right\}, ..., \left\{Z_n^{(8)}\right\}$ are all extinct.

$$\eta = \mathbb{P}\left(E^{(1)}, E^{(2)}, ..., E^{(8)}\right) = \left[\mathbb{P}\left(E^{(1)}\right)\right]^8 \tag{12}$$

since offspring mean $\mu = 0.5 + 2 \times 0.3 = 1.1 > 1$, $\left\{Z_n^{(i)}\right\}$ is in supercritical phase, $\mathbb{P}\left(E^{(1)}\right)$ is the fixed point of $G(s)$. Let's first derive generating function

$$G(s) = 0.2 + 0.5s + 0.3s^2 \tag{13}$$

and solve $G(s) = s$ to get the solution $\mathbb{P}\left(E^{(1)}\right) = \frac{2}{3}$. We get the answer

$$\eta = \left(\frac{2}{3}\right)^8 \tag{14}$$

(b): $\{Z_n\}$ is extinct in generation 3 iff all $\left\{Z_n^{(i)}\right\}$ are extinct in generation 3. $\{Z_n\}$ survives in generation 2 iff there exists some $\left\{Z_n^{(i)}\right\}$ survive in generation 2. Notice that $\{Z_n\}$ survives in generation 2 implies $\{Z_n\}$ survives in generation 1 so the probability we want to find is the probability that $\{Z_n\}$ is extinct in generation 3 and survives in generation 2.

$$\mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)} \cap \bigcup_{i=1}^8 S_2^{(i)}\right) = \mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)}\right) - \mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)} \cap \left[\bigcup_{i=1}^8 S_2^{(i)}\right]^c\right) \tag{15}$$

$$= \mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)}\right) - \mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)} \cap \bigcap_{i=1}^8 \left[S_2^{(i)}\right]^c\right) \tag{16}$$

$$= \mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)}\right) - \mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)} \cap \bigcap_{i=1}^8 E_2^{(i)}\right) \tag{17}$$

is what we want to calculate by noticing $\forall n, i, \left[S_n^{(i)}\right]^c = E_n^{(i)}$. Use the fact that extinction in generation 2 implies extinction in generation 3, this tells us

$$\mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)} \cap \bigcap_{i=1}^8 E_2^{(i)}\right) = \mathbb{P}\left(\bigcap_{i=1}^8 (E_3^{(i)} \cap E_2^{(i)})\right) = \mathbb{P}\left(\bigcap_{i=1}^8 E_2^{(i)}\right) \tag{18}$$

the structure of independence helps us again

$$\mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)} \cap \bigcup_{i=1}^8 S_2^{(i)}\right) = \mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)}\right) - \mathbb{P}\left(\bigcap_{i=1}^8 E_2^{(i)}\right) = \left[\mathbb{P}\left(E_3^{(1)}\right)\right]^8 - \left[\mathbb{P}\left(E_2^{(1)}\right)\right]^8 \tag{19}$$

the final step is to calculate those two probabilities. Recall the property of generating function that $G_X(0) = \mathbb{P}(X = 0)$. Now $\mathbb{P}\left(E_2^{(1)}\right) = \mathbb{P}\left(Z_2^{(1)} = 0\right) = G_{Z_2^{(1)}}(0)$ and we have proved in class that $Z_2^{(1)}$ has generating function

$G(G(s))$. This tells us

$$\begin{cases} \mathbb{P}\left(E_2^{(1)}\right) = G(G(0)) = G(0.2) = 0.312 \\ \mathbb{P}\left(E_3^{(1)}\right) = G(G(G(0))) = G(0.312) = 0.3852 \end{cases} \tag{20}$$

so the probability we want to find is

$$0.3852^8 - 0.312^8 \tag{21}$$

□

Extra Materials: Total Progeny

For branching process $\{Z_n\}$ with $Z_0 = 1$, offspring distribution $\{p_k\}$ and generating function of offspring distribution $G(s)$, the **total progeny** is defined as

$$T = \sum_{n=0}^{\infty} Z_n \tag{22}$$

the overall number of individuals in the branching process. It's easy to see that if extinction probability $\eta = 1$, then $T < \infty$ *a.s.*, otherwise T has positive probability taking value ∞. Due to this fact, the generating function of the total progeny is defined as

$$G_T(s) = \mathbb{E}\left(s^T \cdot \mathbb{I}_{T<\infty}\right) \tag{23}$$

with the indicator added to make sure that $G_T(s)$ is well-defined. Deriving the generating function of T would provide us with a taste of how things work in branching process.

Theorem 1. ***(Generating Function of Total Progeny)***

$$\forall s \in [0,1), G_T(s) = s \cdot G(G_T(s)) \tag{24}$$

Proof. Tear apart the expectation w.r.t. the value of Z_1 to get

$$G_T(s) = \sum_{k=0}^{\infty} \mathbb{P}\left(Z_1 = k\right) \cdot \mathbb{E}(s^T \cdot \mathbb{I}_{T<\infty} | Z_1 = k) \tag{25}$$

now under the condition that $Z_1 = k$, $T = 1 + T_1 + ... + T_k$ where T_j denotes the total progeny of the descendants of the j-th person in generation 1

$$G_T(s) = \sum_{k=0}^{\infty} p_k \cdot s \cdot \mathbb{E}(s^{T_1}...s^{T_k} \cdot \mathbb{I}_{T_1<\infty}...\mathbb{I}_{T_k<\infty} | Z_1 = k) \tag{26}$$

notice that $T_1, ..., T_k, Z_1$ are independent and $T_1, ..., T_k$ are identically distributed, so

$$G_T(s) = \sum_{k=0}^{\infty} p_k \cdot s \cdot \mathbb{E}(s^{T_1} ... s^{T_k} \cdot \mathbb{I}_{T_1<\infty} ... \mathbb{I}_{T_k<\infty}) \tag{27}$$

$$= s \cdot \sum_{k=0}^{\infty} p_k \cdot \mathbb{E}(s^{T_1} \cdot \mathbb{I}_{T_1<\infty}) ... \mathbb{E}(s^{T_k} \cdot \mathbb{I}_{T_k<\infty}) \tag{28}$$

$$= s \cdot \sum_{k=0}^{\infty} p_k \cdot [G_{T_1}(s)]^k \tag{29}$$

$$= s \cdot G(G_{T_1}(s)) \tag{30}$$

at last notice that $T \overset{d}{=} T_1$ since the branching process starting from generation 0 with 1 individual is the same in distribution as the branching process starting from generation 1 with 1 individual, so the distribution of the total progeny in these two cases are the same. We conclude that

$$G_T(s) = s \cdot G(G_T(s)) \tag{31}$$

□

Remark. *By noticing the continuity of G_T and taking $s \to 1^-$, one may find that*

$$G_T(1) = G(G_T(1)) \tag{32}$$

when $\eta = 1$, it's obvious that $G_T(1) = \mathbb{P}(T < \infty) = 1$. When $\eta < 1$, however, $G_T(1) < 1$ and is the fixed point of the generating function $G(s)$. Since in supercritical phase, the fixed point of $G(s)$ in $[0, 1)$ exists and is uniquely the extinction probability η, we conclude that $G_T(1) = \mathbb{P}(T < \infty) = \eta$. This provides ***another perspective understanding the extinction probability****.*

Week 2

Interpretation of Markov Property

From what we have learnt about discrete-state discrete-time Markov chain, it's a stochastic process $\{X_n\}$ satisfying the Markov property

$$\mathbb{P}(X_n = i_n | X_0 = i_0, ..., X_{n-1} = i_{n-1}) = \mathbb{P}(X_n = i_n | X_{n-1} = i_{n-1}) \tag{33}$$

many different interpretations can be made on the Markov property. The most intuitive one is saying that conditional on the value of X_{n-1}, X_n is independent of $X_0, ..., X_{n-2}$. In short, **conditional on present (observation at time $n-1$), past (observation prior to time $n-1$) is independent of future (observation after time $n-1$).**

Another useful interpretation of Markov property is that Markov chain is a process that is **memoryless**. Since Markov property is saying that if one cares about the future behavior of Markov chain, only the most recent past matters, if we have already observed the event $\{X_n = 0\}$ happening, we can actually forget about $X_0, ..., X_{n-1}$ when investigating the behavior of the Markov chain after time n. This interpretation will be made clearer a little bit afterwards.

Due to the presence of Markov property, one can define the transition probability of Markov chain $p_{ij}^n(1) = \mathbb{P}(X_{n+1} = j | X_n = i)$ as the probability of transiting from state i to state j at time n. For simplification, we will only consider the time-homogeneous Markov chain, i.e. Markov chain such that $\mathbb{P}(X_{n+1} = j | X_n = i)$ does not depend on n so the same transition law applies at each time point. After knowing the transition probability, one last thing to know in order to fix the distribution of the whole Markov chain is just the information on where it starts, i.e. the initial distribution of X_0 denoted μ. As a result, **the distribution of a Markov chain is fixed iff the initial distribution and the transition probability are known**.

Remark. *At this point considering time-homogeneous Markov chain, one can always stop the Markov chain and restart it. For example, if we have already observed the event $\{X_n = 0\}$ happening, we can actually forget about $X_0, ..., X_{n-1}$ when we investigate the behavior of the Markov chain after time n. This is equivalent to* ***stopping the current Markov chain at time n and restarting it with the same transition rule but act as if it has initial value 0****. We will come back to this interpretation a lot of times in the future.*

Examples of Markov Chain

Let's look at some examples of Markov chain. The easiest one is the two-state Markov chain with state space $S = \{0, 1\}$ (state of the phone, 0 means free and 1 means busy). It's assumed that at each time point there's probability p that a call is coming in and if the phone was busy then there is q probability that the call will end at this time point. This results in the transition matrix

$$P = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}, \ p, q \in [0, 1] \tag{34}$$

however, if the system can put one caller on hold, the state space is extended to $S = \{0, 1, 2\}$ where state 1 means that the phone is busy but no caller is on hold and state 2 means that the phone is busy and there's also one caller on hold. The transition matrix now becomes

$$P = \begin{bmatrix} 1-p & p & 0 \\ q(1-p) & 1-p(1-q)-q(1-p) & p(1-q) \\ 0 & q & 1-q \end{bmatrix}, \ p, q \in [0, 1] \tag{35}$$

the second row comes from the fact that state 1 transits to state 2 with probability $p(1-q)$ (the old call has not ended and a new call comes in) while state 1 transits to state 0 with probability $q(1-p)$ (the old call has ended and no new call comes in). This provides a simple queueing model that we will be able to analyze later on.

On the other hand, Markov chain can also be formed in structures other than the real line $\mathbb{R}$. Consider the random walk on any undirected graph with $S = \{v_1, ..., v_n\}$ as the set of all vertices. Let $N(v_i) = \{v_j \in V : v_j \sim v_i\}$ be the neighborhood of vertex v_i so that $d_{v_i} = |N(v_i)|$ is called the degree of vertex v_i. When the state is at v_i, it has $\frac{1}{d_{v_i}}$ probability transiting to any one of the states in $N(v_i)$. It's called a **random walk on graph** and it also turns out to be a Markov chain. Another famous example would be the **random walk on infinite binary tree**, we will come back to this interesting example later. Different from the random walk on finite graph, this example is a random walk on infinite graph.

Another useful example to mention is that if $\{X_n\}$ is a Markov chain with state space S, the tuple $Z_n = (X_n, X_{n+1})$ that tracks the **two-step history** of $\{X_n\}$ is also a Markov chain with state space $S \times S$. An easy proof can be given below that

$$\begin{aligned}
&\mathbb{P}(Z_n = (i_n, i_{n+1}) | Z_0 = (i_0, i_1), ..., Z_{n-1} = (i_{n-1}, i_n)) && (36)\\
&= \mathbb{P}(X_n = i_n, X_{n+1} = i_{n+1} | X_0 = i_0, X_1 = i_1, ..., X_n = i_n) && (37)\\
&= \mathbb{P}(X_{n+1} = i_{n+1} | X_0 = i_0, X_1 = i_1, ..., X_n = i_n) && (38)\\
&= \mathbb{P}(X_{n+1} = i_{n+1} | X_{n-1} = i_{n-1}, X_n = i_n) && (39)\\
&= \mathbb{P}(X_n = i_n, X_{n+1} = i_{n+1} | X_{n-1} = i_{n-1}, X_n = i_n) && (40)\\
&= \mathbb{P}(Z_n = (i_n, i_{n+1}) | Z_{n-1} = (i_{n-1}, i_n)) && (41)
\end{aligned}$$

using the Markov property of $\{X_n\}$. An example would be to set $\{X_n\}$ as the output sequence one is getting by tossing a coin independently. For this Markov chain, $S = \{0, 1\}$ so $\{Z_n\}$ is a Markov chain on $\{(0,0), (0,1), (1,0), (1,1)\}$, with transition probability matrix

$$P = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix} \tag{42}$$

and initial distribution as that Z_0 has $\frac{1}{4}$ probability taking all four possible states. This Markov chain is useful if we

want to find, e.g. the expected time of tosses we have to make until we see two consecutive heads.

Remark. *Markov process theory is the core part of probability theory since in real life we rarely see a sequence of independent random variables. Markov process considers a sequence of dependent random variables by putting mild restrictions on the dependency.*

However, one might be curious about the way we deal with non-Markov processes. For example, consider process $\{X_n\}$ *with state space* $\{0,1\}$*, the transition rule is that*

- *If* $X_{n-1}=0, X_{n-2}=0$*, then* $X_n \sim B(1,\frac{1}{2})$
- *If* $X_{n-1}=0, X_{n-2}=1$*, then* $X_n \sim B(1,\frac{3}{4})$
- *If* $X_{n-1}=1, X_{n-2}=0$*, then* $X_n \sim B(1,\frac{1}{4})$
- *If* $X_{n-1}=1, X_{n-2}=1$*, then* $X_n \sim B(1,\frac{2}{3})$

it's quite obvious that $\{X_n\}$ *is not a Markov chain since the transition rule at time* n *differs according to different values of* X_{n-2}*.*

However, $Z_n=(X_n,X_{n+1})$ *turns out to be a Markov chain with transition matrix (please check)*

$$P=\begin{bmatrix} \frac{1}{2} & \frac{1}{2} & 0 & 0 \\ 0 & 0 & \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} & 0 & 0 \\ 0 & 0 & \frac{1}{3} & \frac{2}{3} \end{bmatrix} \tag{43}$$

this is because for $\{X_n\}$ *the transition rule is fixed based on its two-step history and the construction of* Z_n *as a tuple keeps track of the two-step history of* $\{X_n\}$ *at each time point so it's Markov. More generally,* ***when*** $\{X_n\}$ ***is not Markov but its transition rule depends on*** k***-step history, setting*** $Z_n=(X_n,X_{n+1},...,X_{n+k})$ ***always creates a Markov chain*** $\{Z_n\}$ ***at the cost of enlarging the state space****.*

Chapman-Kolmogorov Equation

It should be familiar that Markov property implies Chapman-Kolmogorov equation, however, here we raise a counterexample to show that the converse is not true.

Consider $Y_1, Y_3, ...$ as *i.i.d.* random variables taking value ± 1 with probability $\frac{1}{2}$ and set $Y_{2k}=Y_{2k-1}Y_{2k+1}$ so $Y_2, Y_4, ...$ is also a sequence of *i.i.d.* random variables taking value ± 1 with probability $\frac{1}{2}$. Moreover, the sequence of random variables $Y_1, Y_2, Y_3, ...$ are pairwise independent. Those facts can be checked below

$$\mathbb{P}(Y_2=1)=\mathbb{P}(Y_1Y_3=1)=\mathbb{P}(Y_1=1,Y_3=1)+\mathbb{P}(Y_1=-1,Y_3=-1) \tag{44}$$

$$=\mathbb{P}(Y_1=1)\,\mathbb{P}(Y_3=1)+\mathbb{P}(Y_1=-1)\,\mathbb{P}(Y_3=-1)=\frac{1}{4}+\frac{1}{4}=\frac{1}{2} \tag{45}$$

due to independence of Y_1 and Y_3

$$\mathbb{P}(Y_2 = 1, Y_4 = 1, ..., Y_{2k} = 1) = \mathbb{P}(Y_1Y_3 = 1, Y_3Y_5 = 1, ..., Y_{2k-1}Y_{2k+1} = 1) \tag{46}$$

$$= \mathbb{P}(Y_1 = 1, Y_3 = 1, ..., Y_{2k+1} = 1) + \mathbb{P}(Y_1 = -1, Y_3 = -1, ..., Y_{2k+1} = -1) \tag{47}$$

$$= \frac{1}{2^{k+1}} + \frac{1}{2^{k+1}} = \frac{1}{2^k} = \mathbb{P}(Y_2 = 1)\,\mathbb{P}(Y_4 = 1)\,...\mathbb{P}(Y_{2k} = 1) \tag{48}$$

due to the sequence $Y_1, Y_3, ...$ being *i.i.d.*, for the pairwise independence of $Y_1, Y_2, Y_3, ...$, we only have to check the independence of Y_1 and Y_{2k} without loss of generality

$$\mathbb{P}(Y_1 = 1, Y_{2k} = 1) = \mathbb{P}(Y_1 = 1, Y_{2k-1}Y_{2k+1} = 1) \tag{49}$$

$$= \mathbb{P}(Y_1 = 1, Y_{2k-1} = 1, Y_{2k+1} = 1) + \mathbb{P}(Y_1 = 1, Y_{2k-1} = -1, Y_{2k+1} = -1) \tag{50}$$

$$= \frac{1}{8} + \frac{1}{8} = \frac{1}{4} = \mathbb{P}(Y_1 = 1)\,\mathbb{P}(Y_{2k} = 1) \tag{51}$$

Using the facts mentioned above, the transition probability is well-defined

$$\forall i, j \in \{-1, 1\}, \forall m, p_{ij}(m) = \mathbb{P}(Y_{n+m} = j | Y_n = i) = \mathbb{P}(Y_{n+m} = j) = \frac{1}{2} \tag{52}$$

and the m-step transition matrix is

$$P^{(m)} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} \tag{53}$$

let's check Chapman-Kolmogorov equation

$$P^{(m)}P^{(1)} = \begin{bmatrix} \frac{1}{2} & \frac{1}{2} \\ \frac{1}{2} & \frac{1}{2} \end{bmatrix} = P^{(m+1)} \tag{54}$$

proves that $P^{(m)} = [P^{(1)}]^m$. However, this process $\{Y_n\}$ is not a Markov chain since

$$\mathbb{P}(Y_3 = 1 | Y_2 = 1, Y_1 = 1) = \mathbb{P}(Y_3 = 1 | Y_1Y_3 = 1, Y_1 = 1) = 1 \tag{55}$$

$$\mathbb{P}(Y_3 = 1 | Y_2 = 1) = \mathbb{P}(Y_3 = 1) = \frac{1}{2} \tag{56}$$

violates the Markov property.

Remark. *When one is asked to prove that a process is Markov, merely arguing the existence of the transition matrix or checking Chapman-Kolmogorov equation does not suffice.*

Week 3

Examples of Markov Chain

Lemma 5 (Example)**.** *X, Y are two independent homogeneous Markov chains on the same state space S, show that $Z_n = (X_n, Y_n)$ is a Markov chain on the state space $S \times S$ and derive transition probability.*

Proof. By definition of Markov chain, let's check

$$\mathbb{P}\left(Z_{n+1} = (x_{n+1}, y_{n+1}) | Z_0 = (x_0, y_0), ..., Z_n = (x_n, y_n)\right) \tag{57}$$
$$= \mathbb{P}\left(X_{n+1} = x_{n+1}, Y_{n+1} = y_{n+1} | X_0 = x_0, Y_0 = y_0, ..., X_n = x_n, Y_n = y_n\right) \tag{58}$$
$$= \frac{\mathbb{P}\left(X_0 = x_0, Y_0 = y_0, ..., X_{n+1} = x_{n+1}, Y_{n+1} = y_{n+1}\right)}{\mathbb{P}\left(X_0 = x_0, Y_0 = y_0, ..., X_n = x_n, Y_n = y_n\right)} \tag{59}$$
$$= \frac{\mathbb{P}\left(X_0 = x_0, ..., X_{n+1} = x_{n+1}\right)\mathbb{P}\left(Y_0 = y_0, ..., Y_{n+1} = y_{n+1}\right)}{\mathbb{P}\left(X_0 = x_0, ..., X_n = x_n\right)\mathbb{P}\left(Y_0 = y_0, ..., Y_n = y_n\right)} \tag{60}$$

due to the independence of X, Y, then use Markov property of X, Y and the independence once more

$$\frac{\mathbb{P}\left(X_0 = x_0, ..., X_{n+1} = x_{n+1}\right)\mathbb{P}\left(Y_0 = y_0, ..., Y_{n+1} = y_{n+1}\right)}{\mathbb{P}\left(X_0 = x_0, ..., X_n = x_n\right)\mathbb{P}\left(Y_0 = y_0, ..., Y_n = y_n\right)} \tag{61}$$
$$= \mathbb{P}\left(X_{n+1} = x_{n+1} | X_0 = x_0, ..., X_n = x_n\right)\mathbb{P}\left(Y_{n+1} = y_{n+1} | Y_0 = y_0, ..., Y_n = y_n\right) \tag{62}$$
$$= \mathbb{P}\left(X_{n+1} = x_{n+1} | X_n = x_n\right)\mathbb{P}\left(Y_{n+1} = y_{n+1} | Y_n = y_n\right) \tag{63}$$
$$= \frac{\mathbb{P}\left(X_n = x_n, X_{n+1} = x_{n+1}\right)\mathbb{P}\left(Y_n = y_n, Y_{n+1} = y_{n+1}\right)}{\mathbb{P}\left(X_n = x_n\right)\mathbb{P}\left(Y_n = y_n\right)} \tag{64}$$
$$= \frac{\mathbb{P}\left(X_n = x_n, X_{n+1} = x_{n+1}, Y_n = y_n, Y_{n+1} = y_{n+1}\right)}{\mathbb{P}\left(X_n = x_n, Y_n = y_n\right)} \tag{65}$$
$$= \mathbb{P}\left(X_{n+1} = x_{n+1}, Y_{n+1} = y_{n+1} | X_n = x_n, Y_n = y_n\right) \tag{66}$$
$$= \mathbb{P}\left(Z_{n+1} = (x_{n+1}, y_{n+1}) | Z_n = (x_n, y_n)\right) \tag{67}$$

concludes the proof.

When it comes to the transition probability of Z in terms of X, Y,

$$p^Z_{(x_0,y_0),(x_1,y_1)} = \mathbb{P}\left(X_1 = x_1, Y_1 = y_1 | X_0 = x_0, Y_0 = y_0\right) \tag{68}$$
$$= \frac{\mathbb{P}\left(X_1 = x_1, Y_1 = y_1, X_0 = x_0, Y_0 = y_0\right)}{\mathbb{P}\left(X_0 = x_0, Y_0 = y_0\right)} \tag{69}$$
$$= \frac{\mathbb{P}\left(X_1 = x_1, X_0 = x_0\right)\mathbb{P}\left(Y_1 = y_1, Y_0 = y_0\right)}{\mathbb{P}\left(X_0 = x_0\right)\mathbb{P}\left(Y_0 = y_0\right)} \tag{70}$$
$$= \mathbb{P}\left(X_1 = x_1 | X_0 = x_0\right)\mathbb{P}\left(Y_1 = y_1 | Y_0 = y_0\right) \tag{71}$$
$$= p^X_{x_0,x_1} \cdot p^Y_{y_0,y_1} \tag{72}$$

gives the representation. □

Remark. *The tuple of two independent Markov chain is still a Markov chain. This idea does not seem interesting at the first glance but it turns out to be an important technique called* ***independent coupling****. We will see how this coupling technique helps us when proving the convergence theorem for ergodic Markov chain.*

Lemma 6 (Example)**.** *X is a Markov chain with state space S and $h : S \to T$ is bijective. Show that $Y_n = h(X_n)$ is a Markov chain on T.*

Proof. Again from the definition,

$$\mathbb{P}\left(Y_{n+1} = y_{n+1} | Y_0 = y_0, ..., Y_n = y_n\right) = \mathbb{P}\left(h(X_{n+1}) = y_{n+1} | h(X_0) = y_0, ..., h(X_n) = y_n\right) \tag{73}$$

$$= \mathbb{P}\left(X_{n+1} = x_{n+1} | X_0 = x_0, ..., X_n = x_n\right) \tag{74}$$

$$= \mathbb{P}\left(X_{n+1} = x_{n+1} | X_n = x_n\right) \tag{75}$$

where $x_0 = h^{-1}(y_0), ..., x_{n+1} = h^{-1}(y_{n+1})$ is well-defined and the Markov property of X is applied. Now we just need to go back from X to Y

$$\mathbb{P}\left(X_{n+1} = x_{n+1} | X_n = x_n\right) = \mathbb{P}\left(Y_{n+1} = y_{n+1} | Y_n = y_n\right) \tag{76}$$

to see that Y also has Markov property. □

Remark. *It's left as an exercise to the readers how to construct an example of function $f : S \to T$ such that X is a Markov chain but $Z_n = f(X_n)$ is not a Markov chain.*

The following exercise is left to the reader.

Lemma 7 (Exercise)**.** *Let X be a Markov chain and $Y_n = X_{kn}$ for some fixed positive integer k, prove that Y is also a Markov chain and find the transition matrix of Y in terms of the transition matrix of X.*

Gambler's Ruin

Imagine a person starts gambling with j dollar, each time of gambling he either wins 1 dollar with probability p or loses 1 dollar with probability q where $p + q = 1, p \neq q$. When the person has zero dollar, he loses all his money (ruin state) and when the person reaches N dollar, he stops gambling with his wealth reaches the maximum possible. We want to find what's the probability that the person is in the ruin state if infinitely many times of gambling is allowed.

We shall first build a mathematical model for this process. It's clear that if we denote $\{X_n\}$ as the amount of dollar this person has at time n (before he gambles at time n), it's a Markov chain on $S = \{0, 1, ..., N\}$ and $X_0 = j$ with

$$\forall i \in \{1, 2, ..., N-1\}, \forall j \in S, p_{i,j} = \begin{cases} p & j = i+1 \\ q & j = i-1 \end{cases} \tag{77}$$

the corner case is that

$$p_{0,0} = 1, p_{N,N} = 1 \tag{78}$$

so it's actually a simple asymmetric random walk with absorbing boundary.

Let $\alpha(j)$ denote the probability that the gambler eventually ruins with initially j dollars, we naturally discuss by case based on the value of X_1

$$\forall j \in \{1, 2, ..., N-1\}, \alpha(j) \tag{79}$$
$$= \mathbb{P}(X_1 = j+1 | X_0 = j)\, \mathbb{P}(\text{ruins} | X_1 = j+1, X_0 = j) + \mathbb{P}(X_1 = j-1 | X_0 = j)\, \mathbb{P}(\text{ruins} | X_1 = j-1, X_0 = j) \tag{80}$$
$$= p \cdot \mathbb{P}(\text{ruins} | X_1 = j+1, X_0 = j) + q \cdot \mathbb{P}(\text{ruins} | X_1 = j-1, X_0 = j) \tag{81}$$

by Markov property, we can stop the chain at time 1 and restart it. On observing $\{X_1 = j+1\}$, the person acts as if he is starting the gambling with initially $j+1$ dollars.

$$p \cdot \mathbb{P}(\text{ruins} | X_1 = j+1, X_0 = j) + q \cdot \mathbb{P}(\text{ruins} | X_1 = j-1, X_0 = j) \tag{82}$$
$$= p \cdot \mathbb{P}(\text{ruins} | X_0 = j+1) + q \cdot \mathbb{P}(\text{ruins} | X_0 = j-1) \tag{83}$$
$$= p \cdot \alpha(j+1) + q \cdot \alpha(j-1) \tag{84}$$

now the only work is to solve this recurrence relationship

$$\alpha(j) = p \cdot \alpha(j+1) + q \cdot \alpha(j-1), \alpha(0) = 1, \alpha(N) = 0 \tag{85}$$

Since this recurrence relationship is linear, homogeneous (no extra constants) and has constant coefficients (no dependence on j in the coefficients), the approach of using the root of characteristic equation works. To be clear with that, the characteristic equation is

$$x = px^2 + q \tag{86}$$

solve this to get two distinct roots $x_1 = 1, x_2 = \frac{q}{p}$ (since $p \neq q$), the formula of $\alpha(j)$ must have the form

$$\alpha(j) = c_1 x_1^j + c_2 x_2^j = c_1 + c_2 \left(\frac{q}{p}\right)^j \tag{87}$$

according to the conditions $\alpha(0) = 1, \alpha(N) = 0$, it's possible to solve out

$$c_1 = -\frac{\left(\frac{q}{p}\right)^N}{1 - \left(\frac{q}{p}\right)^N}, c_2 = \frac{1}{1 - \left(\frac{q}{p}\right)^N} \tag{88}$$

provides the formula

$$\alpha(j) = \frac{\left(\frac{q}{p}\right)^j - \left(\frac{q}{p}\right)^N}{1 - \left(\frac{q}{p}\right)^N} \tag{89}$$

as the **ruin probability**.

Remark. *A lot of interesting interpretations can be made from this formula. Consider taking the limit* $N \to \infty$ *(greedy gambler who never quits gambling), when* $p < \frac{1}{2} < q$, $\lim_{N\to\infty} \alpha(j) = 1$, *and when* $q < \frac{1}{2} < p$, $\lim_{N\to\infty} \alpha(j) = \left(\frac{q}{p}\right)^j$. *When the gamble is for the person, the ruin probability is exponentially decaying w.r.t. the amount of initial asset* j. *When the gamble is against the person, the gambler almost surely ruins. That is to say, even if the gamble is designed to be slightly for the person, e.g.* $p = \frac{51}{100}, q = \frac{49}{100}$, *with the amount of initial asset* $j = 50$, *the person still has a non-negligible* 13.53% *probability of getting ruined.*

One might find that we have skipped the case where $p = q = \frac{1}{2}$. This part will be left to the reader, but one has to be careful that when $p = q$, the characteristic equation has two identical roots so $\alpha(j)$ must have the form

$$\alpha(j) = c_1 x_1^j + c_2 j x_2^j \tag{90}$$

repeating the same procedure, one would find out

$$\alpha(j) = 1 - \frac{j}{N} \to 1 \ (N \to \infty) \tag{91}$$

surprisingly, **even if the gamble is fair, a greedy gambler almost surely ruins**.

Week 4

Interpretation of Recurrence and Transience

By definition, state s of a Markov chain is recurrent iff

$$\mathbb{P}_s\left(T_s < \infty\right) = 1 \tag{92}$$

where the subscript s under the probability means that the Markov chain starts with initial state $X_0 = s$ and the stopping time is defined as

$$T_s \overset{def}{=} \inf\left\{n \geq 1 : X_n = s\right\} \tag{93}$$

the first hitting time to state s except time 0.

We have also proved in class that a state is recurrent iff it is almost surely visited for infinitely many times. This is due to the fact that Markov chain can be restarted at any stopping time (strong Markov property), restarting Markov chain at time T_s gives a new Markov chain as if it has initial state $X_{T_s} = s$. Since recurrent state will be visited in finitely many time and the time horizon is infinite, such restarting of Markov chain must happen infinitely often.

Intuitively, recurrence can be understood in terms of the **trend** of stochastic process. If a stochastic process has a certain pattern of trend, it must be transient. An example would be the one in the homework showing that if we have a random walk S_n with *i.i.d.* integrable increments $X_1, X_2, ...$ such that $\mathbb{E}X_1 \neq 0$, then state 0 must be transient. The interpretation is that the strong law of large number provides the conclusion that $\frac{S_n}{n} \overset{a.s.}{\to} \mathbb{E}X_1$ $(n \to \infty)$, saying S_n either goes to $+\infty$ or $-\infty$ depending on the sign of $\mathbb{E}X_1$. In other words, such a random walk asymptotically moves toward $+\infty$ or moves toward $-\infty$, getting farther away from 0 so there's no reason to expect that state 0 will be hit after a long enough period of time, which naturally shows the fact that $p_{s,s}(n) \to 0$ $(n \to \infty)$ for any transient state s.

Remark. *Be careful that the converse in not always true that such interpretation might fail in certain cases. For example, consider simple symmetric random walk in $\mathbb{Z}^d$. Since each increment has equal probability of going in each direction, there is a nice symmetricity for this process, meaning that the process does not have a trend of going somewhere particularly. However, when $d \leq 2$ the process is recurrent and when $d \geq 3$ the process is transient, which is a surprising fact characterizing the essential difference between two dimensional and three dimensional space.*

Remark. *One might suspect that what I have mentioned above is too "unmathematical" since nothing seems to be rigorously stated. It's actually the opposite that the idea is always the most important thing to get while the proof can often be made rigorous without too much effort. The following theorems exactly come from the interpretation above and the readers are welcome to check more details if interested.*

Theorem 2 (Chung-Fuchs)**.** *For random walk on $\mathbb{R}$, if WLLN holds in the form $\frac{S_n}{n} \overset{p}{\to} 0$ $(n \to \infty)$, then $\{S_n\}$ is recurrent.*

Theorem 3. *If S_n is a random walk on $\mathbb{R}^2$ and $\frac{S_n}{\sqrt{n}} \xrightarrow{d} N(\mu, \sigma^2)$ with $\sigma > 0$ non-degenerate, then $\{S_n\}$ is recurrent.*

When discussing recurrence, **irreducibility** is always a useful criterion since all states in the same communication class has the same recurrence or transience property. Recall that if the Markov chain is not irreducible (there exists more than one communication class), one can always first do the canonical decomposition of state space and then discuss the recurrence of each communication class.

Examples of Recurrent and Transient Markov Chain

Lemma 8 (SSRW on Binary Tree)**.** *Consider $\{S_n\}$ as simple symmetric random walk on binary tree where the state space is $S = \{1, 2, ..., 2^N - 1\}$ and SSRW always starts from the root of the tree (node 1), i.e. $S_0 = 1$. The node indices are sorted in the order that node 1 has edges with 2 and 3, node 2 has edges with $4, 5$ and node 3 has edges with $6, 7$, etc. Whenever S_{n-1} is at a node with d degree, S_n transits to all nodes in the neighborhood of S_{n-1} with probability $\frac{1}{d}$. Discuss the recurrence of the Markov chain (be careful that here N can take value as any finite positive integer or $+\infty$).*

Proof. Whatever value N takes, the Markov chain is always irreducible so we only need to consider the recurrence property of a single state, e.g. state 1.

Let's first look at the case where $N < \infty$ so the state space is finite. In this case, there must exist at least one recurrent state (refer to the remark below for the proof and explanation) so the whole chain is recurrent.

When $N = \infty$, however, the Markov chain is transient. To see this fact, let's define another process $\{T_n\}$ where T_n is the height of S_n in the binary tree and it's also a Markov chain. In more detail, the root state 1 has height 0, the node $2, 3$ has height 1, etc. If $\{S_n\}$ is recurrent, $\{T_n\}$ must also be recurrent.

At this point, let's figure out the transition rule of $\{T_n\}$ that $T_0 = 0$ and condition on observing $\{T_n = k\}, k \neq 0$,

$$T_{n+1} = \begin{cases} k+1 & \text{w.p. } \frac{2}{3} \\ k-1 & \text{w.p. } \frac{1}{3} \end{cases} \tag{94}$$

and state 0 is a reflection wall, i.e. state 0 necessarily transits to state 1 for $\{T_n\}$. In other words, $\{T_n\}$ is just a simple asymmetric random walk with reflection boundary, it has a trend of going rightward (going rightward has probability $\frac{2}{3} > \frac{1}{3}$) so it's transient, a contradiction.

In all, $\{S_n\}$ is recurrent on finite binary tree and transient on infinite binary tree. Recurrence property can be very different on finite graph compared to infinite graph! □

Remark. *To see that an irreducible Markov chain with finite state space S must be recurrent, just prove by contradiction that it's otherwise transient so $\forall r, s \in S, p_{r,s}(n) \to 0$ $(n \to \infty)$. Consider*

$$\sum_{s \in S} p_{r,s}(n) = 1 \tag{95}$$

take limit on both sides as $n \to \infty$, the limit goes in since it's a finite sum

$$0 = \sum_{s \in S} \lim_{n \to \infty} p_{r,s}(n) = 1 \tag{96}$$

a contradiction! The explanation for this fact is that since there are only finitely many states but infinite time horizon, there must exists some states that are visited infinitely many times regardless of where the chain starts (the probability mass cannot escape when there are only finitely many states).

Remark. *A rigorous proof of the fact that simple asymmetric random walk on $S = \{0, 1, 2, ...\}$ with reflection wall at 0 must be transient is provided as follow. Consider the recurrence property of state 0 since we have an irreducible Markov chain, intuitively we shall have an exponentially small chance hitting the reflection wall*

$$\mathbb{P}(X_{2k} = 0) = \binom{2k}{k} \left(\frac{1}{3}\right)^k \left(\frac{2}{3}\right)^k = \binom{2k}{k} \left(\frac{2}{9}\right)^k \tag{97}$$

from the Taylor series $(1-x)^{-\frac{1}{2}} = \sum_{n=0}^{\infty} \binom{2n}{n} \left(\frac{x}{4}\right)^n$, we see that

$$\sum_{k=0}^{\infty} \mathbb{P}(X_{2k} = 0) = \left(1 - \frac{8}{9}\right)^{-\frac{1}{2}} = 3 < \infty \tag{98}$$

by Borel-Cantelli lemma (we will learn this in 213B), $\mathbb{P}(X_{2k} = 0 \text{ i.o.}) = 0$ so almost surely $\exists N, \forall n > N$, X_{2n} never hits the reflection wall so the reflection wall is hit for at most finitely many times. From the characterization of recurrence that it requires hitting state 0 to occur infinitely often, we know that state 0 must be transient.

Lemma 9. *Consider the **renewal chain** as a Markov chain with state space $S = \{0, 1, 2, ...\}$ such that $\forall n > 0, p_{n,n-1} = 1$ and $p_{0,n} = p_n$ for some given PMF of the renewal distribution $\{p_n\}_{n \geq 0}$. Show that this chain is always recurrent but positive recurrent iff the renewal distribution has finite mean.*

Proof. This Markov chain is irreducible so we only need to figure out the recurrence property of state 0.

$$\begin{aligned}
\mathbb{P}_0(T_0 = k) &= \mathbb{P}_0(X_1 \neq 0, X_2 \neq 0, ..., X_{k-1} \neq 0, X_k = 0) && (99)\\
&= \mathbb{P}_0(X_1 = k-1, X_2 = k-2, ..., X_{k-1} = 1, X_k = 0) && (100)\\
&= \mathbb{P}_0(X_1 = k-1) = p_{k-1} && (101)
\end{aligned}$$

check the definition of recurrence

$$\mathbb{P}_0(T_0 < \infty) = \sum_{k=1}^{\infty} \mathbb{P}_0(T_0 = k) = \sum_{k=1}^{\infty} p_{k-1} = 1 \tag{102}$$

and check the definition of positive recurrence

$$\mathbb{E}_0 T_0 = \sum_{k=1}^{\infty} k p_{k-1} \tag{103}$$

is finite iff $\sum_{k=0}^{\infty} k p_k < \infty$ proves the conclusion. □

Week 5

Independent Coupling

We have mentioned in the previous context that if $\{X_n\}$ is a Markov chain and $\{Y_n\}$ is an independent copy of $\{X_n\}$ then $Z_n = (X_n, Y_n)$ is also a Markov chain with transition probability $p^Z_{(x_0,y_0),(x_1,y_1)} = p^X_{x_0,x_1} \cdot p^Y_{y_0,y_1}$.

Now let's consider $\{X_n\}$ to be an aperiodic positive recurrent irreducible Markov chain with stationary distribution π. By the lemma shown in lecture notes, irreducible aperiodic Markov chain always has $\forall i, j \in S, \exists n_0 = n_0(i,j)$ such that $\forall n \geq n_0, p_{i,j}(n) > 0$, in other words, the n step transition probability between any two states is strictly positive eventually. As a result, $\{Z_n\}$ **must be irreducible**. To see this fact intuitively, for any two states of $\{Z_n\}$ denoted $(x_0, y_0), (x_1, y_1)$, $\exists n_0, n_1$ such that $\forall n \geq n_0, p_{x_0,x_1}(n) > 0$ and $\forall n \geq n_1, p_{y_0,y_1}(n) > 0$. We would expect $p^Z_{(x_0,y_0),(x_1,y_1)}(n_0 + n_1) = p^X_{x_0,x_1}(n_0 + n_1) \cdot p^Y_{y_0,y_1}(n_0 + n_1) > 0$ so $\{Z_n\}$ is irreducible.

Remark. *Think about a counterexample where the lack of aperiodicity results in Z_n to be not irreducible. Hint: think about two-state alternating Markov chain $\{X_n\}$ with $S = \{0, 1\}, X_0 = 0$ and transition matrix $P = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$. Make an independent copy of $\{X_n\}$ denoted $\{Y_n\}$ with $Y_0 = 0$ then $\{Z_n\}$ can only take values $(0,0), (1,1)$.*

Lemma 10. *Try to find the stationary distribution of $\{Z_n\}$ under the condition above.*

Proof. Since $\{Z_n\}$ is irreducible, its stationary distribution, if exists, is unique. Let's try to guess its stationary distribution and verify it. It's natural to guess that $\pi_{(x,y)} = \pi_x \pi_y$ is the stationary distribution of $\{Z_n\}$.

$$\sum_{x_0,y_0} \pi_{(x_0,y_0)} \cdot p^Z_{(x_0,y_0),(x_1,y_1)} = \sum_{x_0} \pi_{x_0} \cdot p^X_{x_0,x_1} \cdot \sum_{y_0} \pi_{y_0} \cdot p^Y_{y_0,y_1} = \pi_{x_1} \cdot \pi_{y_1} = \pi_{(x_1,y_1)} \tag{104}$$

also check the normalization property

$$\sum_{x,y} \pi_{(x,y)} = \sum_x \pi_x \sum_y \pi_y = 1 \cdot 1 = 1 \tag{105}$$

so such $\pi_{(x,y)}$ is the unique stationary distribution of $\{Z_n\}$. □

At this point, one could go back to the lecture notes and check the proof of the ergodic theorem for Markov chain. The trick is to make independent coupling with $X_0 \sim \mu$ following any initial distribution but $y_0 \sim \pi$ following stationary distribution. Whenever $\{Z_n\}$ first hits the diagonal set $D = \{(i,i) : i \in S\}$, stop the Markov chain and restart it so $\{Z_n\}$ forgets about the fact that X_0, Y_0 has different distribution but acts as if the chain starts at $Z_0 = (i,i)$ for some state $i \in S$. At this point after $\{Z_n\}$ has hit the diagonal, $\{X_n\}, \{Y_n\}$ can be viewed as Markov chains having the same initial distribution and transition rule so they have the same distribution at each time step, resulting in X_n converging to the stationary distribution (since $Y_0 \sim \pi$, we know $\forall n, Y_n \sim \pi$). One would now be amazed at how smart the proof of ergodic theorem is using independent coupling.

Birth Death Chain (BDC)

In this section we talk about the analysis of discrete-time BDC to be prepared for the study of continuous-time Markov chain and BDC. In the context, we restrict ourselves to BDC on $S = \mathbb{N}$, the set of natural numbers. The transition rule is intuitive

$$p_{i,i+1} = p_i, p_{i,i} = r_i, p_{i,i-1} = q_i \ (p_i + q_i + r_i = 1) \tag{106}$$

with the Markov chain assumed to be irreducible (p_i, q_i are strictly positive except the corner case that $q_0 = 0$). It's not obvious how to analyze the recurrence and positive recurrence property of this BDC since the state space is infinite and one cannot figure out the "trend" of the process easily.

In this case, we depart from the definition, set F_i as the first hitting time to state i. Directly figuring out the distribution of F_i is hard so we calculate instead $\mathbb{P}(F_i < F_j | X_0 = m)$ for $0 \leq i < m < j$, the probability that starting from state m the BDC visits state i before state j.

Similar to what we have done in the example of gambler's ruin, we hope that Markov property provides us with a recurrence relationship for

$$u(k) = \mathbb{P}(F_i < F_j | X_0 = k) \ \ (i \leq k \leq j) \tag{107}$$

it's clear that $u(i) = 1, u(j) = 0$. Conduct the first-step decomposition

$$\forall i < k < j, u(k) = q_k \mathbb{P}(F_i < F_j | X_0 = k, X_1 = k-1) + r_k \mathbb{P}(F_i < F_j | X_0 = k, X_1 = k) \tag{108}$$

$$+ p_k \mathbb{P}(F_i < F_j | X_0 = k, X_1 = k+1) \tag{109}$$

$$= q_k u(k-1) + r_k u(k) + p_k u(k+1) \tag{110}$$

from Markov property. This is a linear homogeneous recurrence relationship with non-constant coefficients (coefficients has dependence on k) so the characteristic equation method fails. However, if we notice that $p_k + q_k + r_k = 1$, it's still possible to solve

$$p_k u(k) + q_k u(k) = q_k u(k-1) + p_k u(k+1) \tag{111}$$

$$[u(k+1) - u(k)] = \frac{q_k}{p_k}[u(k) - u(k-1)] \tag{112}$$

concludes

$$u(k+1) - u(k) = \prod_{a=i+1}^{k} \frac{q_a}{p_a}[u(i+1) - u(i)] = \frac{P_k}{P_i}[u(i+1) - u(i)] \tag{113}$$

with the notation $P_k = \prod_{i=1}^{k} \frac{q_i}{p_i}$ summing up both sides for $k \in \{i, i+1, ..., j-1\}$ to get

$$-1 = u(j) - u(i) = \sum_{k=i}^{j-1} \frac{P_k}{P_i}[u(i+1) - u(i)] \tag{114}$$

solves out

$$u(i+1) - u(i) = -\frac{P_i}{\sum_{k=i}^{j-1} P_k} \tag{115}$$

so

$$u(s+1) - u(s) = -\frac{P_s}{\sum_{l=i}^{j-1} P_l} \tag{116}$$

To get the expression of $u(k)$, it's again summing both sides w.r.t. $s \in \{k, k+1, ..., j-1\}$

$$u(j) - u(k) = -\sum_{s=k}^{j-1} \frac{P_s}{\sum_{l=i}^{j-1} P_l} \tag{117}$$

this tells us

$$u(k) = \frac{\sum_{i=k}^{j-1} P_i}{\sum_{l=i}^{j-1} P_l} \tag{118}$$

Theorem 4 (Recurrence of BDC)**.** *Irreducible BDC on $S = \mathbb{N}$ is recurrent iff $\sum_{l=1}^{\infty} P_l = \infty$, i.e. $\sum_{l=1}^{\infty} \prod_{i=1}^{l} \frac{q_i}{p_i} = \infty$.*

Proof. Irreducible BDC is recurrent iff state 0 is recurrent iff $\mathbb{P}_0(F_0 < \infty) = 1$. To use our calculations above, we need the starting state of the Markov chain to be strictly larger than 0, let's think about if it's possible to start the Markov chain at state 1. It turns out that first step decomposition provides

$$\mathbb{P}_0(F_0 < \infty) = p_0 \mathbb{P}_0(F_0 < \infty | X_1 = 1) + r_0 \mathbb{P}_0(F_0 < \infty | X_1 = 0) \tag{119}$$

$$= p_0 \mathbb{P}_1(F_0 < \infty) + r_0 \tag{120}$$

with $p_0 + r_0 = 1$ so $\mathbb{P}_0(F_0 < \infty) = 1$ iff $\mathbb{P}_1(F_0 < \infty) = 1$.

To set up a first stopping time to the state larger than 1, let's take $j > 1$ and notice that $F_j \overset{\mathbb{P}_1 - a.s.}{\to} +\infty$ $(j \to +\infty)$, so there's enough reason to believe that

$$\mathbb{P}_1(F_0 < \infty) \overset{?}{=} \lim_{j \to +\infty} \mathbb{P}_1(F_0 < F_j) \tag{121}$$

the question mark here means that this step is not rigorously argued and some further work is required. It's left to the readers.

Now plug in the calculation result to see

$$\mathbb{P}_1\left(F_0<\infty\right)=\lim_{j\to+\infty}\frac{\sum_{i=1}^{j-1}P_i}{\sum_{l=0}^{j-1}P_l}=1-\frac{P_0}{\sum_{l=0}^{\infty}P_l} \tag{122}$$

this limit is 1 iff $\sum_{l=1}^{\infty}P_l=\infty$ concludes the proof ($P_0=1$ is defined separately for consistency). □

Remark. *The construction of $u(k)$ actually contains the martingale perspective of BDC. We are not able to talk about that approach due to the lack of tools (martingale convergence theorem) but the readers are welcome to come back to this argument after studying martingale theory.*

The following examples are special cases of the BDC mentioned above.

Lemma 11 (Example)**.** *Prove that the simple random walk on $S=\mathbb{N}$ with reflection boundary at 0 is recurrent iff $p\leq\frac{1}{2}$ (p is the probability the increment is taking value 1).*

Lemma 12 (Example)**.** *Prove that the irreducible BDC on $S=\mathbb{N}$ with reflection boundary at 0 has*

$$\frac{q_n}{p_n}\to l\ (n\to\infty) \tag{123}$$

prove that if $l<1$ the chain is transient and if $l>1$ the chain is recurrent.

After considering the recurrence property, let's consider the positive recurrence of such irreducible BDC. Actually, the criterion of positive recurrence is easier to derive by noticing its connection with the invariant measure. It's clear from the lecture note that irreducible Markov chain has unique invariant measure μ (up to a constant multiple) and it's positive recurrent iff $\sum_{s\in S}\mu_s<\infty$, i.e. it can be normalized to a stationary distribution.

For BDC, if μ is an invariant measure, it's necessary that

$$\sum_{j\in S}\mu_j p_{j,k}=\mu_{k-1}p_{k-1}+\mu_k r_k+\mu_{k+1}q_{k+1}=\mu_k \tag{124}$$

use the fact $1=p_k+r_k+q_k$ so

$$\mu_{k+1}q_{k+1}-\mu_k p_k=\mu_k q_k-\mu_{k-1}p_{k-1}=...=\mu_1 q_1-\mu_0 p_0 \tag{125}$$

assume $\mu_1=\frac{p_0}{q_1},\mu_0=1$ (the selection is not unique) then $\mu_k q_k-\mu_{k-1}p_{k-1}=0$ solves out

$$\mu_k=\prod_{j=1}^{k}\frac{p_{j-1}}{q_j} \tag{126}$$

check for $k=0$ (corner case) gives $\mu_0=1=\mu_1 q_1+\mu_0 r_0$ so it's exactly an invariant measure. The following theorem follows immediately from our knowledge on Markov chain.

Theorem 5 (Positive Recurrence of BDC)**.** *Irreducible BDC on $S=\mathbb{N}$ is positive recurrent iff $\sum_{k=0}^{\infty}\mu_k<\infty$, i.e. $\sum_{k=0}^{\infty}\prod_{j=1}^{k}\frac{p_{j-1}}{q_j}<\infty$.*

Lemma 13 (Example)**.** *Discuss positive recurrence property of simple random walk on $S = \mathbb{N}$ with reflection boundary at 0.*

Lemma 14 (Example)**.** *Find some nontrivial transient/null recurrent/positive recurrent examples of time-inhomogeneous random walk on $S = \mathbb{N}$ with reflection boundary at 0, i.e. p_j, q_j must depend on j.*

Hint: When $p_j = \frac{1}{2} + \frac{1}{4\sqrt{j}}, q_j = \frac{1}{2} - \frac{1}{4\sqrt{j}}, r_j = 0$, it's transient. When $p_j = \frac{j+1}{2j+1}, q_j = \frac{j}{2j+1}, r_j = 0$, it's null recurrent. When $p_j = \frac{(j+1)^2}{2[(j+1)^2+(j+2)^2]}, q_j = \frac{(j+1)^2}{2[(j+1)^2+j^2]}, r_j = 1 - p_j - q_j$, it's positive recurrent.

Week 6

Metropolis-Hastings (MH) Algorithm

The MH algorithm is a kind of acceptance-rejection algorithm that enables one to draw random samples from the distribution with likelihood (or PDF) $f(x)$ only knowing $h \propto f$. In other words, one does not have to have a full knowledge on the likelihood (the normalization constant) but only the structure of the likelihood matters. The details of the algorithm is presented below for the purpose of completeness. From the lecture, we see the proof why this algorithm samples from the correct distribution. To be concise, it's because $\{X_n\}$ generated is a time reversible Markov chain with the PDF of stationary distribution as f.

Algorithm 1 Metropolis-Hastings

Input: Integrable function $f(x)$, reference density $q(y|x)$

Output: Random samples generated $X_0, X_1, ..., X_n$ as a Markov chain converging in distribution to the probability distribution with density h where $h \propto f$.

1: Choose arbitrary initial state X_0 and assume that we have already generated random samples $X_0, ..., X_i$ and we work on generating X_{i+1}.

2: Generate random sample (proposal) $Y \sim q(y|X_i)$

3: Evaluate $r \equiv r(X_i, Y)$ where $r(x, y) = \min\left\{\frac{h(y)q(x|y)}{h(x)q(y|x)}, 1\right\}$

4: Set $X_{i+1} = \begin{cases} Y & w.p.\ r \\ X_i & w.p.\ 1-r \end{cases}$

Remark. *The motivation of MH is to construct transition probability such that the detailed balance condition holds*

$$f(x)p(x, y) = f(y)p(y, x) \tag{127}$$

since this implies that f is the density of the stationary distribution. To make a comment here, think about the detailed balance condition in the sense of Physics that $f(x)p(x, y)$ is the mass of substance transmitting from state x to y and $f(y)p(y, x)$ is the mass of substance transmitting from state y to x and they are equal for any pair of states x, y. That's the reason we call it "detailed balance".

WLOG, assume $f(x)q(y|x) > f(y)q(x|y)$ so $r(x, y) = \frac{f(y)q(x|y)}{f(x)q(y|x)} < 1, r(y, x) = 1$. In this case,

$$p(x, y) = q(y|x)r(x, y), p(y, x) = q(x|y) \tag{128}$$

so the detailed balance condition always holds. The smart point is to consider likelihood ratio $\frac{f(y)}{f(x)} = \frac{h(y)}{h(x)}$ which is exactly known since the unknown normalization constant cancels out.

Markov Chain Monte Carlo (MCMC)

MH is actually a special case of the so-called MCMC method. We shall be familiar with normal Monte Carlo method that it's an application of the law of large numbers. Naturally, normal Monte Carlo procedure involves

generating *i.i.d.* random samples and using the sample mean to approximate the unknown expectation.

However, MCMC goes against this idea. The main reason is that *i.i.d.* random samples typically has restricted capacities and sometimes dependency between random variables helps. The easiest model for a sequence of dependent random variables is just the Markov chain. As a result, MCMC generates random samples as a Markov chain and hope to play with those concepts in the Markov chain theory.

One broad class of MCMC algorithm makes use of the stationary distribution π and tries to design the transition probability of $\{X_n\}$ such that it's stationary distribution matches with what we hope to see. In MH, we design the transition probability such that the ergodic theorem of Markov chain works and guarantees the convergence towards stationary measure π. If the distribution we want to sample from is exactly π, then our work is done by simply simulating the Markov chain for a long enough time! This is a very clever idea and turns out to be very practically useful.

Lemma 15 (Example)**.** *If the only source of randomness we can have is from a black box that generates i.i.d. random samples from the distribution $B(1, \frac{1}{4})$, try to construct a random number generator that generates random samples from the distribution $B(1, \frac{1}{2})$. Try to make it as efficient as possible.*

Remark (Hint)**.** *Think about MCMC method to design $B(1, \frac{1}{2})$ as the stationary distribution of some Markov chain. One possible example is $\{X_n\}$ with state space $S = \{0, 1\}$, starting from $X_0 = 0$ with transition matrix*

$$P = \begin{bmatrix} \frac{3}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{3}{4} \end{bmatrix} \tag{129}$$

check the stationary distribution and think about why we can simulate the trajectory of this Markov chain. Then think about how efficient this algorithm is.

Implementing Metropolis-Hastings Algorithm

After understanding the fact that MH actually makes use of the ergodic theorem of Markov chain, it's natural to expect that the convergence takes time. As a result, after we start simulating the Markov chain $\{X_n\}$, the samples are not immediately useable and we have to wait some time until the Markov chain converges to the stationary distribution. This period is called the **burn-in period**. One might be wondering: how long is the burn-in period typically? How can we check whether the burn-in period has ended?

There are much details hidden behind but we are able to provide some superficial details here. Typically, the convergence in the ergodic theorem depends on the transition diagram of the Markov chain, but it's often exponentially decaying, i.e.

$$||\mu P^t - \pi||_2 \leq Ce^{-rt} \tag{130}$$

where C is some constant that does not depend on t and $r > 0$ is the rate. Just to clarify, μP^t is the distribution of X_t with initial distribution μ and the convergence speed is measured under vector ℓ_2 norm (of course there are other different measurements). As a consequence, unless the transition diagram of the Markov chain is "not nice",

we expect to only observe a short burn-in period. In order to ensure $||\mu P^t - \pi||_2 \leq \varepsilon$ for some error tolerance $\varepsilon > 0$,

$$t \geq \frac{1}{r} log \frac{C}{\varepsilon} \tag{131}$$

suffices. Numerically, if we want to clearly know if the burn-in period has ended, statistical tests (e.g. Kolmogorov-Smirnov test etc.) are available to judge if the distribution of random samples is not changing a lot.

When it comes to the detail of MH, another topic is the **choice of reference distribution** q. Typically, we require that q has a heavier tail than the distribution we want to sample from. For example, using Gaussian reference to sample from Cauchy doesn't behave numerically well but using Cauchy reference to sample from Gaussian is acceptable. Intuitively, if the reference has a lighter tail, then it's rare for it to generate samples at the tail of the distribution we want to sample from, causing the lack of exploration. Notice that q is actually a conditional distribution since we need to use $q(x|y)$ and $q(y|x)$ in MH. As a result, one of the choices is to set $q(\cdot|y)$ as a distribution centered at y, e.g. set $q(\cdot|y)$ as the PDF of $N(y, 1)$.

Application: Bayesian Setting

The most direct application of MH is to sample from the posterior. Consider the example where $X_1, ..., X_n \sim N(\theta, 1)$ with a prior given as $\pi(\theta) = \frac{1}{\pi(1+\theta^2)}$ and we want to calculate the posterior mean of θ.

Bayes formula tells us

$$\pi(\theta|x_1, ..., x_n) \propto \pi(\theta) p(x_1, ..., x_n|\theta) \propto \frac{1}{1+\theta^2} e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2} \tag{132}$$

with the normalization constant $C = \int_{\mathbb{R}} \frac{1}{1+\theta^2} e^{-\frac{1}{2}\sum_{i=1}^n (x_i-\theta)^2} d\theta$ impossible to calculate analytically. At this point, we desperately need to sample from the posterior distribution without even knowing C, which can be done by MH by taking the reference distribution $q(\cdot|y)$ as a Cauchy distribution centered at y. After getting samples as the output of a Markov chain from MH, the law of large numbers of Markov chain tells us that the sample mean approximates the posterior mean of θ (although they are not *i.i.d.* samples).

Special Case: Gibbs Sampler

We want to sample from a bivariate target distribution with fully known joint likelihood $f_{U,V}(u, v)$. Adopting the same MCMC idea in MH, assume that random sample U_i, V_i has been generated such that $(U_i, V_i) \sim f_{U,V}$, how to form new random samples U_{i+1}, V_{i+1} such that $(U_{i+1}, V_{i+1}) \sim f_{U,V}$?

Intuitively, we would say: why not create sample $U_{i+1} \sim f_{U|V}(u|V_i)$ and then create sample $V_{i+1} \sim f_{V|U}(v|U_{i+1})$? Since the joint likelihood is known, the conditional likelihood can definitely be derived. This sampling scheme turns out to be correct and is just called **the Gibbs sampler**.

Let's first see an example for Gibbs sampler where the joint likelihood is given as

$$f_{U,V}(u, v) = \frac{n!}{(n-u)!u!} v^{u+\alpha-1}(1-v)^{n-u+\beta-1} \ (u \in \{0, 1, ..., n\}, v \in [0, 1]) \tag{133}$$

we know nothing about this joint density so it seems that we should be using the multivariate version of MH directly. However, if we try to calculate the conditional likelihood, life becomes much easier. It's clear that the marginals

$$f_U(u) = \frac{n!}{(n-u)!u!} Beta(u+\alpha, n-u+\beta) \ (u \in \{0,1,...,n\}) \tag{134}$$

$$f_V(v) = v^{\alpha-1}(1-v)^{\beta-1} \ (v \in [0,1]) \tag{135}$$

so the conditional likelihoods are

$$f_{U|V}(u|v) = \frac{n!}{(n-u)!u!} v^u (1-v)^{n-u} \tag{136}$$

$$f_{V|U}(v|u) = \frac{1}{Beta(u+\alpha, n-u+\beta)} v^{u+\alpha-1}(1-v)^{n-u+\beta-1} \tag{137}$$

those are distributions we are familiar with

$$U|_V \sim B(n,V), V|_U \sim Beta(U+\alpha, n-U+\beta) \tag{138}$$

so it's very easy to sample from the conditionals. As a result, Gibbs sampler helps us complete the sampling task easily and effectively. The random sample sequence $U_1, V_1, U_2, V_2, ...$ is generated iteratively from

$$U_{i+1} \sim B(n, V_i), V_{i+1} \sim Beta(U_{i+1}+\alpha, n-U_{i+1}+\beta) \tag{139}$$

At this point, we are clear with how Gibbs sampler works. Let's show that Gibbs sampler is actually a special case of Metropolis-Hastings. After U_i, V_i have been generated, we generate sample U_{i+1}, V_{i+1} according to the dynamics of Gibbs sampler so the Markov chain generated is $\{(U_n, V_n)\}$ but in an alternating way. Under MH framework, when we are generating U_{i+1}, the reference distribution is $Y \sim f_{U|V}(\cdot|V_i)$. On the other hand, when we are generating V_{i+1}, the reference distribution is $Z \sim f_{V|U}(\cdot|U_{i+1})$.

As a result, the acceptance probability for the proposal Y is calculated through

$$r = \min\left\{\frac{f_U(Y)\cdot f_{V|U}(V_i|Y)}{f_V(V_i)\cdot f_{U|V}(Y|V_i)}, 1\right\} = 1 \tag{140}$$

since $f_{U|V}f_V = f_{U,V} = f_{V|U}f_U$. Similarly, the acceptance probability for the proposal Z is also 1 so **Gibbs sampler is just a special case of MH that never rejects the proposal**.

Remark. *Gibbs sampler can be generalized to sampling random vector from any distribution as long as all "full conditionals" (leave-one-out)* $f_{U_1|U_2,...,U_n}, f_{U_2|U_1,U_3,...,U_n}, ...$ *are known and can be sampled from in an easy way.*

Week 7

Poisson Process

In order to build continuous-time discrete-state Markov chain, we first introduce Poisson process as an important model. **Poisson process with intensity** λ is defined as $\{N_t\}$ satisfying $N_0 = 0$, $N_{t+h} - N_t \sim P(\lambda h)$ with independent increments, i.e. $\forall 0 \leq t_1 < ... < t_n, N_{t_1} - N_0, N_{t_2} - N_{t_1}, ..., N_{t_n} - N_{t_{n-1}}$ are independent. Obviously, it's a continuous-time increasing stochastic process and N_t can only take values in $\mathbb{N}$.

Let's state several most important properties of Poisson process. The first one is the *i.i.d.* **exponential inter-arrival time** of Poisson process. Imagine N_t is the number of customers in a shop at time t, then Poisson process actually provides a **memoryless arrival model** for the customers. To see this point, denote $T_i = \inf\{t : N_t = i\}$ as the first time there are i arrivals, then

$$\{T_i \leq t < T_{i+1}\} = \{N_t = i\} \tag{141}$$

since $T_0 = 0$, it's clear that

$$\mathbb{P}(t < T_1) = \mathbb{P}(N_t = 0) = e^{-\lambda t}, T_1 \sim \mathcal{E}(\lambda) \tag{142}$$

and

$$\mathbb{P}(T_2 - T_1 > k, T_1 > t) = \int_t^\infty \mathbb{P}(T_2 > x + k | T_1 = x) \cdot f_{T_1}(x)\, dx \tag{143}$$

the problem turns into calculating the conditional probability $\mathbb{P}(T_2 > x + k | T_1 = x)$ where

$$\begin{aligned} \mathbb{P}(T_2 > x + k | T_1 = x) &= \mathbb{P}(N_{x+k} = 1 | T_1 = x) && (144)\\ &= \mathbb{P}(N_{x+k} - N_x = 0 | T_1 = x) && (145)\\ &= \mathbb{P}(N_{x+k} - N_x = 0) = e^{-\lambda k} && (146) \end{aligned}$$

since $\{T_1 = x\}$ is only related to $\{N_t\}_{t \in [0,x]}$, which is independent of $N_{x+k} - N_x$ by the independent increment property. As a result,

$$\mathbb{P}(T_2 - T_1 > k, T_1 > t) = \int_t^\infty e^{-\lambda k} \cdot \lambda e^{-\lambda x}\, dx = e^{-\lambda k} e^{-\lambda t} \tag{147}$$

proves $\mathbb{P}(T_2 - T_1 > k | T_1 > t) = e^{-\lambda k}$ does not contain t so $T_2 - T_1 \sim \mathcal{E}(\lambda)$ and is independent of T_1. A similar argument enables us to prove that interarrival times $T_1, T_2 - T_1, ..., T_k - T_{k-1}, ...$ are *i.i.d.* and follows $\mathcal{E}(\lambda)$.

The second one is the **thinning of Poisson process**. For Poisson process $\{N_t\}$ with intensity λ, if there's a classifier independent of the whole Poisson process that splits all arrivals into two different processes $\{P_t\}, \{Q_t\}$, then those two processes must be independent Poisson processes. To be more specific, consider a classifier that classifies customers into male and female customers, $\{P_t\}$ only records the arrival of male customers while $\{Q_t\}$ only

records the arrival of female customers. Assume that each customer has probability p of being male and $1-p$ of being female, let's investigate those two thinned processes.

Obviously, $P_0 = 0$, $P_{t+h} - P_t = k$ iff there are k male customers arriving between $[t, t+h]$ iff there are K customers arriving between $[t, t+h]$ and k of them are male

$$\mathbb{P}(P_{t+h} - P_t = k) = \sum_{K=k}^{\infty} \mathbb{P}(N_{t+h} - N_t = K) \cdot \binom{K}{k} p^k (1-p)^{K-k} \tag{148}$$

$$= \sum_{K=k}^{\infty} \frac{(\lambda h)^K}{K!} e^{-\lambda h} \cdot \binom{K}{k} p^k (1-p)^{K-k} \tag{149}$$

$$= \sum_{K=k}^{\infty} \frac{(\lambda h)^K}{k!(K-k)!} e^{-\lambda h} \cdot p^k (1-p)^{K-k} \ (l = K-k) \tag{150}$$

$$= \frac{(\lambda hp)^k e^{-\lambda h}}{k!} \sum_{l=0}^{\infty} \frac{(\lambda h)^l}{l!} \cdot (1-p)^l \tag{151}$$

$$= \frac{(\lambda hp)^k}{k!} e^{-\lambda hp} \tag{152}$$

proves $P_{t+h} - P_t \sim P(\lambda ph)$ and it obviously has independent increments since $\{N_t\}$ does so $\{P_t\}$ is a Poisson process with intensity $p\lambda$. Similarly, $\{Q_t\}$ is a Poisson process with intensity $(1-p)\lambda$.

To prove the independence of $\{P_t\}$ and $\{Q_t\}$, it suffices to prove that the increments $P_{t+h} - P_t, Q_{t+h} - Q_t$ are independent. By the same reasoning,

$$\mathbb{P}(P_{t+h} - P_t = k, Q_{t+h} - Q_t = j) = \mathbb{P}(N_{t+h} - N_t = j+k) \cdot \binom{j+k}{k} p^k (1-p)^j \tag{153}$$

$$= \frac{(\lambda h)^{j+k}}{(j+k)!} e^{-\lambda h} \cdot \binom{j+k}{k} p^k (1-p)^j \tag{154}$$

$$= \frac{(p\lambda h)^k ((1-p)\lambda h)^j}{j!k!} e^{-p\lambda h} e^{-(1-p)\lambda h} \tag{155}$$

$$= \mathbb{P}(P_{t+h} - P_t = k) \cdot \mathbb{P}(Q_{t+h} - Q_t = j) \tag{156}$$

proves the conclusion.

Remark. *This is a remarkable result since $P_t + Q_t = N_t$, the thinned process must add up to the original Poisson process but they are actually independent!*

Conversely, one can easily prove that the sum of two independent Poisson processes with intensity λ, μ adds up to a Poisson process with intensity $\lambda + \mu$.

Construction of Continuous-time Markov Chain

At this point, we introduce the motivation of the construction of continuous-time discrete-state Markov chain. Assume $\{X_n\}$ is a discrete-time discrete-state Markov chain, to include the continuous-time effect, the biggest

difficulty is to maintain Markov property, which is some kind of memoryless property. However, since Poisson process has memoryless arrival, it's immediate that we expect to see

$$Y_t = X_{N_t} \tag{157}$$

as a continuous-time Markov chain. The idea is simple, whenever it makes a state transition, it waits for a period of time exactly the same as the interarrival time in Poisson process. It turns out that this construction maintains the Markov property of this process. For simplicity, we **don't allow the underlying discrete-time Markov chain** $\{X_n\}$ **to have** $p_{ii} > 0$, i.e. any state cannot transit to itself. Under this setting, any transition of state happens at rate λ which is the intensity of the Poisson process.

However, one could easily find out some restrictions out of this construction. For example, the interarrival time are *i.i.d.* for Poisson process but that's not necessarily true for continuous-time Markov chain. Independence of interarrival time is actually enough to guarantee the Markov property. As a result, we generalize this construction to add a **holding rate** for each state (high holding rate results in low holding time on average), denoted

$$q : S \to (0, \infty) \tag{158}$$

such that the holding time of each state could be different.

The final model is organized as the following:

- Initial state X_0, the transition rule of $\{X_n\}$ and holding rates q are given.
- Sample $E_0 \sim \mathcal{E}(1)$, scale it with holding rate of state X_0 to get $T_0 = \frac{1}{q(X_0)} E_0 \sim \mathcal{E}(q(X_0))$
- After time T_0, make a state transition for the underlying chain from X_0 to X_1 so $\forall t \in [0, T_0), Y_t = X_0$ and $Y_{T_0} = X_1$.
- Sample $E_1 \sim \mathcal{E}(1)$ independent of E_0, scale it with holding rate of state X_1 to get $T_1 = \frac{1}{q(X_1)} E_1 \sim \mathcal{E}(q(X_1))$
- After time T_1, make a state transition for the underlying chain from X_1 to X_2 so $\forall t \in [T_0, T_0 + T_1), Y_t = X_1$ and $Y_{T_0+T_1} = X_2$.
- Repeat this procedure.

here we naturally call T_i **holding times** and they are independent (interarrival time in Poisson process). $S_j = T_0 + T_1 + ... + T_{j-1}$ is the time when the j-th jump happens (arrival time in Poisson process).

Birth-death chain (BDC) is the most useful example of continuous-time Markov chain. In the case of BDC, the state transition of $\{X_n\}$ either increases the state by 1 (birth) or decreases the state by 1 (death) so the only nontrivial transition probability of $\{X_n\}$ is $p_{i,i+1}, p_{i,i-1}$.

Denote $B(i), D(i)$ as the time until the next birth/death given that $Y_t = i$ (there are i individuals now), then $B(i), D(i)$ are independent exponentially distributed random variables (to maintain the Markov property) so that

$$B(i) \sim \mathcal{E}(\lambda_i), D(i) \sim \mathcal{E}(\mu_i) \tag{159}$$

where λ_i, μ_i are called **birth/death rates** with $q_i = \lambda_i + \mu_i$ as holding rates such that

$$\lambda_i = p_{i,i+1}q_i, \mu_i = p_{i,i-1}q_i \tag{160}$$

can be easily verified.

Remark. *Notice that the holding time $T_i = \min\{B(i), D(i)\} \sim \mathcal{E}(\lambda_i + \mu_i)$ (the time until the next birth or death happens, which brings with a state transition) since $B(i), D(i)$ are independent exponentially distributed. That's why $q_i = \lambda_i + \mu_i$ is the holding rate.*

Interpret Models as BDC

Let's try to identify some commonly appearing models as BDC and try to specify their birth/death rates. The correspondence is important to truly understand the generality of BDC and to apply conclusions of BDC for those models.

Before doing that, let's recall that continuous-time Markov chain $\{Y_t\}$ is **regular** if $\forall i \in S, \mathbb{P}_i(S_\infty = \infty) = 1$, i.e. starting from any state, there's finite number of state transition in finite time, i.e. the state transition does not happen too often. An equivalent condition for this is that $\forall i \in S, \mathbb{P}_i\left(\sum_n \frac{1}{q(X_n)} = \infty\right) = 1$ and a useful sufficient condition is that $X_0 = i$ and i is the recurrent state of $\{X_n\}$.

Let's first look at **M/M/1 queue** with memoryless arrival (rate λ), memoryless serving time (rate μ) and only one server in the system. Let Y_t denote the number of people in the system at time t, $Y_0 = 0$. It's clear that it's a BDC with birth rate $\lambda_i = \lambda$ and death rate $\mu_i = \mu$ so the holding rate $q_i = \lambda + \mu$, a regular process.

Then what if we increase the number of servers to N? The birth rate remains the same while the death rate shall be figured out by considering

$$D(i) = \min\{V_1, ..., V_{N\wedge i}\} \tag{161}$$

where actually $N \wedge i$ servers are working given that i people are in the system and V_k denotes the serving time of the k-th server. As a result, $V_k \sim \mathcal{E}(\mu)$ so $D(i) \sim \mathcal{E}((N\wedge i)\mu)$, providing us with $\mu_i = (N\wedge i)\mu$. Notice that when $N = \infty$, $\mu_i = i\mu$, a process with constant birth and linear death. For M/M/∞ queue, $q_i = \lambda + i\mu, p_{i,i+1} = \frac{\lambda}{\lambda+i\mu}, p_{i,i-1} = \frac{i\mu}{\lambda+i\mu}$ and there's no obvious way to judge if this process is regular. However, if we consider the recurrence of the underlying discrete-time Markov chain $\{X_n\}$, it becomes a discrete-time BDC we have discussed. Let's check that

$$\sum_{l=1}^{\infty}\prod_{i=1}^{l}\frac{p_{i,i-1}}{p_{i,i+1}} = \sum_{l=1}^{\infty}\prod_{i=1}^{l} i\frac{\mu}{\lambda} = \sum_{l=1}^{\infty}\left(\frac{\mu}{\lambda}\right)^l l! = \infty \tag{162}$$

proves that $\{X_n\}$ is recurrent (and it's irreducible) so this proves **M/M/∞ queue** is regular.

The last example to mention is the **continuous-time branching process with immigration**. Each particle acts independently, waits $\mathcal{E}(q)$ time from it appears, after which it splits into two particles with probability p or vanishes with probability $1 - p$. New particles immigrate into the system following a Poisson process with intensity λ. Y_t denotes the number of particles in the system at time t.

In this example, given $Y_t = i$, the holding time

$$T = \min\{L_1, ..., L_i, I\} \tag{163}$$

where $L_1, ..., L_i \sim \mathcal{E}(q)$ are the times particles have to wait until they make a decision and I is the next arrival time of immigrant. It's clear that $T \sim \mathcal{E}(\lambda + iq)$ so the holding rate $q_i = iq + \lambda$. Now $p_{i,i+1}$ is the probability that the next transition is a birth, i.e. an immigrant comes in or a particle splits, so

$$p_{i,i+1} = \frac{\lambda}{\lambda + iq} + p\frac{iq}{\lambda + iq}, p_{i,i-1} = 1 - p_{i,i+1} \tag{164}$$

so

$$\lambda_i = \lambda + ipq, \mu_i = i(1-p)q \tag{165}$$

similar to the example above, let's check if the underlying discrete-time Markov chain $\{X_n\}$ is recurrent

$$\sum_{l=1}^{\infty}\prod_{i=1}^{l}\frac{p_{i,i-1}}{p_{i,i+1}} = \sum_{l=1}^{\infty}\prod_{i=1}^{l}\frac{i(1-p)q}{\lambda + ipq} = \sum_{l=1}^{\infty}[(1-p)q]^l\prod_{i=1}^{l}\frac{i}{\lambda + ipq} \tag{166}$$

it seems hard to judge convergence from the first sight. It's clear that $\exists k \in \mathbb{N}, \lambda \leq kpq$ so

$$\frac{i}{\lambda + ipq} \geq \frac{1}{pq}\frac{i}{i+k}, \prod_{i=1}^{l}\frac{i}{\lambda + ipq} \geq (pq)^{-l}\prod_{i=1}^{l}\frac{i}{k+i} \tag{167}$$

with

$$\sum_{l=1}^{\infty}\prod_{i=1}^{l}\frac{p_{i,i-1}}{p_{i,i+1}} \geq \sum_{l=1}^{\infty}\left(\frac{1-p}{p}\right)^l\frac{k!l!}{(k+l)!} = \infty \tag{168}$$

iff $p < \frac{1}{2}$ (calculate the ratio of the $n+1$-th term and the n-th term in the series). At least, we can claim that **the process is regular if** $p < \frac{1}{2}$. Those examples show the connection between continuous-time and discrete-time BDC.

Lemma 16 (Exercise)**.** *Prove that for continuous-time branching process without immigration, i.e.* $\lambda = 0$*, the process is regular if* $p \leq \frac{1}{2}$*.*

Week 8

The Generator

As we have seen in the lecture, the (infinitesimal) generator of the continuous-time Markov chain $\{Y_t\}$ is defined as

$$G = \lim_{h \to 0} \frac{P_h - I}{h} \tag{169}$$

where $(P_t)_{ij} = \mathbb{P}(Y_h = j | Y_0 = i)$ is the transition probability matrix and I is the identity matrix. We have shown that

$$G_{ij} = \begin{cases} -q_i & \text{if } i = j \\ q_i p_{ij} & \text{else} \end{cases} \tag{170}$$

so the generator provides another perspective for the dynamics of the Markov chain.

To see this, for a small enough $h \to 0$,

$$p_{i,i}(h) = \mathbb{P}(Y_h = i | Y_0 = i) \tag{171}$$

if the chain stays at the same state after a small enough time h, it either does not transit or it transits for multiple times and then come back to the same state. However, transiting more than one time in time interval $[0, h]$ has probability

$$\begin{aligned} \mathbb{P}(S_2 < h | X_0 = i, X_{S_1} = j) &= \mathbb{P}\left(\frac{E_1}{q_i} + \frac{E_2}{q_j} < h\right) && (172) \\ &\leq \mathbb{P}\left(\frac{E_1}{q_i} < h\right) \mathbb{P}\left(\frac{E_2}{q_j} < h\right) && (173) \\ &= (1 - e^{-q_i h})(1 - e^{-q_j h}) = o(h) && (174) \end{aligned}$$

where we used the Taylor expansion $e^{-x} = 1 - x + o(x)$ $(x \to 0)$. At this point, we are able to proceed and see that

$$\begin{aligned} p_{i,i}(h) &= \mathbb{P}(Y_h = i, \text{no transition happens} | Y_0 = i) + o(h) && (175) \\ &= \mathbb{P}\left(\frac{E_1}{q_i} > h\right) + o(h) && (176) \\ &= e^{-q_i h} + o(h) = 1 - q_i h + o(h) && (177) \end{aligned}$$

using Taylor expansion once more. As a result,

$$p_{i,i}(h) = 1 + G_{ii} h + o(h) \tag{178}$$

and a similar argument shows

$$\begin{aligned}
\forall i \neq j, p_{i,j}(h) &= \mathbb{P}\left(Y_h = j | Y_0 = i\right) && (179)\\
&= \mathbb{P}\left(Y_h = j, \text{one transition happens} | Y_0 = i\right) + o(h) && (180)\\
&= p_{ij} \cdot \mathbb{P}\left(\frac{E_1}{q_i} < h\right) + o(h) && (181)\\
&= p_{ij} \cdot (1 - e^{-q_i h}) + o(h) && (182)\\
&= p_{ij} q_i h + o(h) && (183)\\
&= G_{ij} h + o(h) && (184)
\end{aligned}$$

where p_{ij} is the transition probability of the underlying discrete-time Markov chain $\{X_n\}$.

At this point, the interpretation of the generator should be a little bit clearer. The diagonal value of the generator matrix G_{ii} is the **negative flow rate out of state** i and the non-diagonal value G_{ij} is the **flow rate from state** i **into state** j. From the generator, we are actually viewing everything in terms of rate and it has completely captured the **drift effect** of the Markov chain in terms of **rate**.

At this point, the backward Kolmogorov equation

$$P_t' = GP_t \tag{185}$$

and the forward Kolmogorov equation

$$P_t' = P_t G \tag{186}$$

characterizes transition probability P in terms of G so knowing the generator matrix is equivalent to knowing the transition law of $\{Y_t\}$.

Lemma 17 (Exercise)**.** *Argue from definition that for time-homogeneous continuous-time Markov chain* $\{Y_t\}$*, its transition probability satisfies* $P_{t+s} = P_t P_s = P_s P_t$ *resulting in* $P_t G = G P_t$*.*

Just to mention here, the backward and forward Kolmogorov equations is very similar to the ODE

$$y' = ay \tag{187}$$

whose solution is $y(t) = Ce^{at}$. In analogue, one expects to see

$$P_t = e^{tG} \tag{188}$$

with the exponential of the operator to be well-defined for G with enough regularity. This shows another connection between the transition probability and the generator.

Remark. *At this point, one might be wondering why the backward and forward Komogorov equations are important since they seem to be similar to each other and does not have an intuitive explanation. This is mainly because we*

are currently in the setting of discrete-state Markov chain where P_t, G can still be represented as matrices (although it's possible countably infinite dimensional). When we are in the setting of continuous-time continuous-state Markov chain, the state space is uncountable so there's no possibility to organize everything in terms of matrices.

*In that case, the Markov chain $\{Y_t\}$ is typically defined through an SDE (stochastic differential equation) and its generator G is defined to totally capture the drift effect of the Markov process (the same motivation). The difference is that such G no longer has matrix representation (a mapping from a countable dimensional space to another countable dimensional space) but is actually an operator (a mapping from a function space to another function space). The calculation of the generator requires stochastic calculus taught in the higher-level courses so we won't be mentioning it here but the backward and forward Kolmogorov equation still exists. Instead of being matrix-valued ODE, since the state space is continuous, we would expect them to be **PDE**, which exactly gives birth to the Feynman-Kac formula that connects SDE and PDE.*

Lemma 18 (Exercise)**.** *From the fact that transition probability matrix P_t and generator matrix G has one-to-one correspondence, prove that $\{N_t\}$ is a Poisson process with intensity λ if and only if it is a continuous-time birth death chain satisfying the following conditions:*

$$N_0 = 0 \tag{189}$$

$$\forall t > 0, h > 0, \mathbb{P}\left(N_{t+h} - N_t \geq 2\right) = o(h) \ (h \to 0) \tag{190}$$

$$\forall t > 0, h > 0, \mathbb{P}\left(N_{t+h} - N_t = 1\right) = \lambda h + o(h) \ (h \to 0) \tag{191}$$

Hint: try to write out birth death rate of Poisson process and use the interpretation above for the generator matrix.

Remark. *At this point, we should have seen at least three equivalent definitions of Poisson process. The first defines it as a process with independent stationary Poisson increments, the second defines it with birth death rate (transition probability of continuous-time Markov chain), the third defines it through the generator of continuous-time Markov chain. Different characterizations are easy to use under different situations and one has to make proper decisions which characterization to use. Two examples below illustrate why those characterizations might be useful.*

Lemma 19 (Exercise)**.** *Using the characterization of Poisson process above, prove that the sum of two independent Poisson process with intensity λ, μ respectively is still a Poisson process with intensity $\lambda + \mu$.*

Lemma 20 (Exercise)**.** *Prove the thinning of Poisson process, i.e. let $\{N_t\}$ be a Poisson process with intensity λ, and now there is a Bernoulli classifier independent of the whole Poisson process labelling all arrivals into category 0 with probability p and category 1 with probability $1-p$. $\{M_t\}$ denotes the number of arrivals in category 0 until time t, prove that M_t is still a Poisson process with intensity $p\lambda$.*

Hint: Under the framework of continuous- time birth death chain, let B_i^M be the time until next birth for process $\{M_t\}$ given that i individuals are in the system. It suffices to prove $B_i^M \sim \mathcal{E}(p\lambda)$. From the thinning procedure, we have $B_i^M = \sum_{j=1}^K B_i^j$ where $K \sim G(p)$ follows geometric distribution and $B_i^1, B_i^2, ...$ are i.i.d. random variables following $\mathcal{E}(\lambda)$ (birth time of Poisson process $\{N_t\}$).

Week 9

Sample Problems for the Final

In my opinion, the core topics we have covered throughout the quarter can be categorized into the following:

- Poisson Process
- Discrete-time discrete-state Markov chain
- GWB branching process
- Continuous-time discrete-state Markov chain (in particular birth-death chain)

I will provide two relevant problems for each of the topic listed above. You are welcome to first solve those problems on your own and then read my solution as reference. Please do not feel bad if you are finding those problems to be hard, since the problems in the final exam **will not be as hard as those**. Please just view the reviewing process as another chance to learn more things. I wish all of you good luck in the final.

Lemma 21 (Example)**.** *$N_1(t), ..., N_n(t)$ are independent Poisson process with respective intensities $\lambda_1, ..., \lambda_n$, now the aggregated Poisson process $N(t) = \sum_{i=1}^n N_i(t)$ has first arrival time T and J the index of the Poisson process responsible for the first arrival. In other words,*

$$T \overset{def}{=} \inf \{t \geq 0 : N(t) = 1\} \tag{192}$$

and $J = i$ happens if and only if

$$N_i(T) = 1 \tag{193}$$

which means the first arrival of the aggregated Poisson process actually comes from the process $N_J(t)$.

Show that T, J are independent and find their marginal distributions.

Proof. Let $T_1, ..., T_n$ be the first arrival times of $N_1(t), ..., N_n(t)$ respectively, it's clear that they are independent and

$$T_i \sim \mathcal{E}(\lambda_i) \tag{194}$$

as a result,

$$T = \min \{T_1, ..., T_n\} \sim \mathcal{E}(\lambda) \tag{195}$$

where $\lambda = \sum_{i=1}^n \lambda_i$.

When it comes to J, it's clear that J is a discrete r.v. with support $\{1, 2, ..., n\}$. Consider the joint distribution of (T, J) that

$$\begin{aligned}
\mathbb{P}(T > t, J = j) &= \mathbb{P}(\min \{T_1, ..., T_n\} > t, J = j) && (196)\\
&= \mathbb{P}(\min \{T_1, ..., T_n\} = T_j, T_j > t) && (197)\\
&= \mathbb{E}[\mathbb{P}(\min \{T_1, ..., T_n\} = T_j, T_j > t | T_j)] && (198)\\
&= \int_0^\infty \mathbb{P}(\min \{T_1, ..., T_n\} = T_j, T_j > t | T_j = x) \cdot f_{T_j}(x)\, dx && (199)\\
&= \int_t^\infty \mathbb{P}(\min \{T_1, ..., T_n\} = x | T_j = x) \cdot f_{T_j}(x)\, dx && (200)\\
&= \int_t^\infty \mathbb{P}(\min \{T_1, ..., T_{j-1}, T_{j+1}, ..., T_n\} \geq x | T_j = x) \cdot f_{T_j}(x)\, dx && (201)
\end{aligned}$$

where we used the law of iterated expectation. The independence of $T_1, ..., T_n$ allows us to get rid of the condition

$$\int_t^\infty \mathbb{P}\left(\min\left\{T_1, ..., T_{j-1}, T_{j+1}, ..., T_n\right\} \geq x | T_j = x\right) \cdot f_{T_j}(x)\, dx \tag{202}$$

$$= \int_t^\infty \mathbb{P}\left(T_1 \geq x, ..., T_{j-1} \geq x, T_{j+1} \geq x, ..., T_n \geq x\right) \cdot f_{T_j}(x)\, dx \tag{203}$$

$$= \int_t^\infty e^{-\lambda x} \cdot e^{\lambda_j x} \cdot f_{T_j}(x)\, dx \tag{204}$$

$$= \lambda_j \int_t^\infty e^{-\lambda x}\, dx \tag{205}$$

$$= \frac{\lambda_j}{\lambda} e^{-\lambda t} \tag{206}$$

it's clear that

$$\mathbb{P}\left(J = j\right) = \lim_{t \to 0} \mathbb{P}\left(T > t, J = j\right) = \frac{\lambda_j}{\lambda}, j \in \{1, ..., n\} \tag{207}$$

and $\mathbb{P}\left(T > t, J = j\right) = \mathbb{P}\left(T > t\right) \cdot \mathbb{P}\left(J = j\right)$ proves independence of T and J. □

Lemma 22 (Example)**.** *Let's consider the famous "coupon collector's problem" where we have n different kinds of pokemons indexed by $\{1, 2, ..., n\}$. Each time we encounter a pokemon, it's gonna be uniformly random among all possible kinds of pokemons and we always catch it. Let X_n denote the pokemon we catch on the n-th trial so that $X_1, X_2, ...$ are i.i.d. and uniformly distributed on $\{1, 2, ..., n\}$. Let T be the first time we have collected all n different kinds of pokemons at least once, i.e.*

$$T \overset{def}{=} \inf\{m \geq 1 : \{X_1, ..., X_m\} = \{1, ..., n\}\} \tag{208}$$

(1): Calculate $\mathbb{E}T$ (Hint: geometric distribution).

(2): Despite the fact that $\mathbb{E}T$ can be easily calculated, it's well-known that the distribution of T is hard to find. As an example, calculating $\mathbb{P}(T \leq m)$ (the probability that we have collected all different pokemons within m trials) is very hard. However, a smart approach is to calculate $\mathbb{P}(T \leq M)$ instead where $M \sim P(m)$ is a Poisson distributed random integer independent of all the trials. In other words, we change the deterministic integer m to a random integer M and calculate the probability that we have collected all different pokemons within a random number $M \sim P(m)$ of trials. Try to calculate $\mathbb{P}(T \leq M)$. (Hint: thinning of Poisson process)

(3): Argue why $M \sim P(m)$ is a reasonable approximation to m when $m \to \infty$. (Hint: central limit theorem)

*(4): Use the approximation in (2) and (3), prove $\forall x \in \mathbb{R}, \mathbb{P}(T \leq nx + n\log n) \to e^{-e^{-x}}$ $(n \to \infty)$. Check that $e^{-e^{-x}}$ is actually a CDF on $\mathbb{R}$ (of the so-called **Gumbel distribution**). We have actually proved*

$$\frac{T - n\log n}{n} \xrightarrow{d} Gumbel \ (n \to \infty) \tag{209}$$

*this is the **asymptotic Gumbel approximation of the full collection time**. From this one can reproduce a lot of well-known results, for example $\frac{T}{n\log n} \xrightarrow{p} 1$ $(n \to \infty)$ so asymptotically, collecting all n different kinds of pokemons require $n\log n$ trials.*

Proof. (1): We can decompose the stopping time T into the sum of stopping times. Let T_i be the first time we have collected i different pokemons at least once. Then $T = T_n$ so

$$T = T_1 + (T_2 - T_1) + ... + (T_n - T_{n-1}) \tag{210}$$

where $T_i - T_{i-1}$ is the time we have spent towards collecting the i-th new kind of pokemon we have never collected before. It's clear that $T_i - T_{i-1} \sim G(\frac{n-i+1}{n})$ since $T_i - T_{i-1}$ is the number of trials until we get the i-th new kind of pokemon but that happens with probability $\frac{n-i+1}{n}$ in each single trial (the $i-1$ pokemons we have already collected are not considered "new"!).

As a result,

$$\mathbb{E}T = \mathbb{E}T_1 + \mathbb{E}(T_2 - T_1) + ... + \mathbb{E}(T_n - T_{n-1}) \tag{211}$$

$$= 1 + \frac{n}{n-1} + \frac{n}{n-2} + ... + \frac{n}{1} \tag{212}$$

so the expected time of full collection is

$$\mathbb{E}T = n\sum_{i=1}^{n}\frac{1}{i} \tag{213}$$

(2): Now that $M \sim P(m)$ can be seen as the value of the Poisson process with intensity m at time 1. If we add an independent classifier classifying the pokemon we collect into its index, each kind of pokemon appears with uniform probability $\frac{1}{n}$ so the thinning of Poisson process tells us that M_i, the number of the i-th kind of pokemon we have caught, satisfies

$$M_1 + ... + M_n = M, M_i \sim P\left(\frac{m}{n}\right) \tag{214}$$

and $M_1, ..., M_n$ are independent.

As a result,

$$\mathbb{P}\left(T \leq M\right) = \mathbb{P}\left(M_1 \geq 1, ..., M_n \geq 1\right) \tag{215}$$

since $T \leq M$ means that after M trials we have collected all kinds of pokemons, i.e. the number of each kind of pokemon we have caught is at least 1. Use independence,

$$\mathbb{P}\left(T \leq M\right) = \left[\mathbb{P}\left(M_1 \geq 1\right)\right]^n \tag{216}$$

$$= (1 - e^{-\frac{m}{n}})^n \tag{217}$$

(3): $M \sim P(m)$ can be written as a summation of r.v.

$$M \overset{d}{=} \xi_1 + ... + \xi_m \tag{218}$$

where $\xi_1, ..., \xi_m \overset{i.i.d.}{\sim} P(1)$. By central limit theorem,

$$\frac{M - m}{\sqrt{m}} \overset{d}{\to} N(0, 1) \ (m \to \infty) \tag{219}$$

that's to say, when $m \to \infty$, M is most likely to be between $[m - 3\sqrt{m}, m + 3\sqrt{m}]$. Since $\sqrt{m}$ has a much smaller order than m, M is a reasonable approximation to m.

(4): Set $m = nx + n\log n$, when $n \to \infty$, $m \to +\infty$ so M is a reasonable approximation to m. From the

calculation in (2),

$$\mathbb{P}\left(T \leq nx + n\log n\right) = \mathbb{P}\left(\frac{T - n\log n}{n} \leq x\right) \tag{220}$$

$$= (1 - e^{-\frac{nx+n\log n}{n}})^n \tag{221}$$

$$= \left(1 - e^{-x}\frac{1}{n}\right)^n \to e^{-e^{-x}} \ (n \to \infty) \tag{222}$$

if $F(x) = e^{-e^{-x}}$, $F(-\infty) = e^{-\infty} = 0, F(+\infty) = e^0 = 1$, it's continuous on $\mathbb{R}$ and is monotone increasing, a valid CDF.

To see the final result,

$$\forall x \in \mathbb{R}, \mathbb{P}\left(T \leq nx + n\log n\right) = \mathbb{P}\left(\frac{T - n\log n}{n} \leq x\right) \to F(x) \ (n \to \infty) \tag{223}$$

is the pointwise convergence of the CDF, which by definition implies convergence in distribution (if you don't know this part it's completely fine, we will learn it in more detail in the next quarter). □

Lemma 23 (Example)**.** *Chung-Fuchs theorem in Markov chain theory tells us that for random walk $\{S_n\}$ on $\mathbb{R}$, if the weak law of large number holds in the form $\frac{S_n}{n} \xrightarrow{p} 0$ $(n \to \infty)$, then $\{S_n\}$ is a recurrent Markov chain.*

(1): Use this fact to prove that if $\{S_n\}$ is a discrete-time discrete-state random walk on $\mathbb{R}$ with $S_0 = 0$ and i.i.d. increments $\xi_n = S_n - S_{n-1}$ being integrable, i.e. $\forall n, \mathbb{E}|\xi_n| < \infty$, then the asymptotic behavior of S_n must be in one of the following cases with probability 1:

1. $\forall n, S_n = 0$

2. $S_n \to +\infty$ $(n \to \infty)$

3. $S_n \to -\infty$ $(n \to \infty)$

4. $\limsup_{n\to\infty} S_n = +\infty, \liminf_{n\to\infty} S_n = -\infty$

(2): Without mentioning any extra technical details, we would expect the conclusion above to hold without the integrable condition, i.e. for increments ξ_1 following any probability distribution, and to hold also for continuous-state random walk, i.e. ξ_1 can have a continuous probability distribution. In particular, if ξ_i has a symmetric non-degenerated distribution w.r.t. the origin, case 4 always holds.

*Under this situation, let's consider a special kind of random walk where the increments $\xi_1, \xi_2, ...$ are i.i.d. following the α-**stable law** which is a continuous symmetric distribution w.r.t. the origin. Let $\phi(t) = \mathbb{E}e^{it\xi_1}$ denote the **characteristic function** of ξ_1 (which we will learn more in the next quarter), assume we already know the characteristic function $\phi(t) = e^{-|t|^\alpha}$ $(0 < \alpha < 1)$, judge the convergence of the following integral*

$$\int_{-\delta}^{\delta} \frac{1}{1-\phi(t)}\, dt \tag{224}$$

for some $\delta > 0$.

(3): A well-known result for the recurrence of random walk on $\mathbb{R}$ is that S_n is recurrent if and only if

$$\int_{-\delta}^{\delta} \frac{1}{1-\phi(t)}\, dt = \infty \tag{225}$$

for some $\delta > 0$ where ϕ is the characteristic function of the increment. Apply this result for the α-stable random walk in (2) and combine it with the conclusion from (1). What interesting results can you get for such kind of random walk?

Proof. (1): Actually we just need to discuss the sign of the expectation of the increments $\mu = \mathbb{E}|\xi_1|$. From strong law of large number, with probability 1,

$$\frac{S_n}{n} \to \mu \ (n \to \infty) \tag{226}$$

if $\mu > 0$, we are in case 2 and if $\mu < 0$ we are in case 3. If $\forall n, \xi_n = 0$ *a.s.*, we are in case 1 so it suffices to prove that if $\mu = 0$ but ξ_1 has a nondegenerated distribution (not almost surely zero), then case 4 must happen.

In this case, from the law of large number and Chung-Fuchs theorem, $\{S_n\}$ is recurrent, i.e. if we denote its

state space as $S \subset \mathbb{R}$,

$$\forall s \in S, \mathbb{P}\left(S_n = s \ i.o.\right) = 1 \tag{227}$$

now that if we can prove S is unbounded with probability 1, then with probability 1 we can find a subsequence of $\{S_n\}$ with limit s for $\forall s \in S$, which implies that $\limsup_{n\to\infty} S_n = +\infty, \liminf_{n\to\infty} S_n = -\infty$.

Finally, we prove that the state space S must be unbounded for such a random walk. Since S is at most countable, $\mu = 0$ and r.v. ξ is non-degenerated, $\exists a > 0, \mathbb{P}\left(\xi_1 = a\right) > 0$. By the independence,

$$\mathbb{P}\left(S_n = na\right) \geq \mathbb{P}\left(\xi_1 = a, ..., \xi_n = a\right) = \left[\mathbb{P}\left(\xi_1 = a\right)\right]^n > 0 \tag{228}$$

so $\forall n, na \in S$ proves S is unbounded from above. Similarly, S is unbounded from below which concludes the proof.

(2): The calculation is as follow

$$\int_{-\delta}^{\delta} \frac{1}{1-\phi(t)}\, dt = \int_{-\delta}^{\delta} \frac{1}{1-e^{-|t|^\alpha}}\, dt \tag{229}$$

$$= 2\int_{0}^{\delta} \frac{1}{1-e^{-t^\alpha}}\, dt \tag{230}$$

it's clear that $e^{-t^\alpha} \leq 1 - t^\alpha + \frac{t^{2\alpha}}{2}$ so

$$\int_{0}^{\delta} \frac{1}{1-e^{-t^\alpha}}\, dt \leq \int_{0}^{\delta} \frac{1}{t^\alpha - \frac{t^{2\alpha}}{2}}\, dt < \infty \tag{231}$$

since $0 < \alpha < 1$. This proves

$$\int_{-\delta}^{\delta} \frac{1}{1-\phi(t)}\, dt < \infty \tag{232}$$

(3): It's immediate that α-stable random walk is transient from the calculation in (2).

Use the result in part (1), since α-stable law is symmetric w.r.t. the origin and non-degenerated, case 4 is true, $\limsup_{n\to\infty} S_n = +\infty, \liminf_{n\to\infty} S_n = -\infty$ happens with probability 1.

There seems to be a "contradiction" between those results. Transience implies that such α-stable random walk only visits each state for finitely many times, however there is a fluctuation behavior, i.e. the value taken by this random walk fluctuates in a large extent between $-\infty$ and $+\infty$. As a result, the fluctuation of this random walk gets larger and larger as time goes by and there are great variations in its trajectories (the increments take extreme values quite often) such that it's transient despite the fact that it admits no particular "trend". □

Lemma 24 (Example)**.** *When it comes to the graph random walk, we always assume that an undirected connected graph G is given. There are n vertices $v_1, ..., v_n$ and each vertex v_i has degree d_i, i.e. there are d_i edges on v_i. Since the graph is connected, any two vertices has a path connecting each other and $\forall i, d_i \geq 1$. For simplicity, loops are not allowed in the graph, i.e. an edge cannot connect a vertex v_i to itself.*

The graph random walk $\{X_n\}$ starts at $X_0 = v_1$ and always transits along one of the associated edges to the next vertex with equal probability.

(1): Write down the state space S and the one-step transition probability.

(2): Is this Markov chain irreducible? Recurrent? Positive recurrent?

(3): Is there a stationary distribution? Is the stationary distribution unique? Find the stationary distribution π if it exists.

(4): Try to give an example of a graph G on which the graph random walk is not ergodic.

Proof. (1): Since the graph is connected, starting at v_1, there is always positive probability of going to any other vertex so $S = \{v_1, v_2, ..., v_n\}$.

The one-step transition probability

$$p_{v_i,v_j} = \mathbb{P}\left(X_{n+1} = v_j | X_n = v_i\right) \tag{233}$$

is non-zero iff there is at least one edge connecting v_i and v_j denoted $v_i \sim v_j$. In this case,

$$p_{v_i,v_j} = \frac{1}{d_i} \tag{234}$$

since there are d_i edges associated with v_i and each edge has equal probability of being chosen.

(2): Since the graph is connected, the graph itself is just the transition diagram of this Markov chain, there is only one communication class so it's irreducible.

The state space is finite so there exists at least one recurrent state so the whole chain is recurrent. From the finiteness of the state space, once more, the whole chain must be positive recurrent.

(3): Since the chain is irreducible and positive recurrent, the stationary distribution exists and is unique. In this problem, it's hard to write out the transition matrix so why don't we depart from the invariant measure and then normalize it to a stationary distribution.

By the definition of invariant measure μ,

$$\sum_{i=1}^{n} \mu_{v_i} p_{v_i,v_j} = \mu_{v_j} \tag{235}$$

and the LHS simplifies to

$$\sum_{i:v_i \sim v_j} \mu_{v_i} \frac{1}{d_i} \tag{236}$$

it's obvious that by setting $\mu_{v_i} = d_i$, we get on LHS $\sum_{i:v_i \sim v_j} 1 = d_j = \mu_{v_j}$. As a result, $\mu_{v_i} = d_i$ gives an invariant

measure. After normalization,

$$\pi_{v_i} = \frac{d_i}{\sum_{j=1}^{n} d_j} \tag{237}$$

gives the unique stationary distribution.

(4): For the graph random walk to be not ergodic, it's equivalent to require it to be not aperiodic.

An easy example is the star graph where the node in the middle is v_1 and has an edge with all other vertices $v_2, ..., v_n$. This is a graph with n vertices and $n-1$ edges. As a result, starting from v_1, it's only possible to return back to v_1 in $2, 4, 6, ...$ steps. Taking greatest common divisor to get the period 2 so this graph random walk is not aperiodic.

□

For branching process, especially the discrete-time branching process, there is not much to say. Keep in mind the phase transitions and the way to calculate the extinction probability. Refer to lemma 3 and lemma 4 for two important examples on discrete-time GWB branching process.

The following example is based on continuous-time branching process.

Lemma 25 (Example)**.** *Consider the continuous-time branching process with immigration and emigration denoted* $\{X_t\}$ *starting with* $X_0 = 1$*. Each individual, after waiting for* $\mathcal{E}(q)$ *time, either splits into two individuals with probability* p *or vanishes with probability* $1-p$*. The immigration is modelled by Poisson process with intensity* λ *and the emigration is modelled by Poisson process with intensity* μ*. The immigration and emigration process are independent of the individuals within the system.*

(1): Model this process as a continuous-time birth death chain. Find birth/death rate, holding rate and the transition probability of the underlying discrete-time Markov chain $\{Y_n\}$*.*

(2): Assume $\mu = \lambda$ *the intensity of immigration and emigration is the same and* $p = \frac{1}{2}$*, is this continuous-time Markov chain regular?*

(3): Provide a sufficient condition on p, q, λ, μ *for the continuous-time Markov chain to be regular. Try to interpret your result intuitively.*

Proof. (1):

When there are i individuals in the system, the holding time is

$$\min\{W_1, ..., W_i, I, E\} \tag{238}$$

where $W_1, ..., W_i \sim \mathcal{E}(q)$ are *i.i.d.* waiting times of each individual within the system, I is the time until next immigration, $I \sim \mathcal{E}(\lambda)$ and E is the time until next emigration, $E \sim \mathcal{E}(\mu)$. Since $W_1, ..., W_i, I, E$ are independent, it's clear that the holding time has distribution $iq + \lambda + \mu$, which is the holding rate

$$q_i = iq + \lambda + \mu \tag{239}$$

when it comes to the birth process, the time until next birth is

$$\min\{B_1, ..., B_i, I\} \tag{240}$$

where $B_1, ..., B_i$ are *i.i.d.* waiting times until next birth of each individual within the system. Since the waiting time $W_1, ..., W_i \overset{i.i.d.}{\sim} \mathcal{E}(q)$, it can be seen as the interarrival time of a Poisson process with intensity q. With an independent Bernoulli classifier $B(1, p)$ added to each Poisson arrival, the thinning of Poisson process tells us the occurrence of birth is a Poisson process with intensity pq. As a result, $B_1 \sim \mathcal{E}(pq)$, the same reasoning provides the birth rate

$$\lambda_i = ipq + \lambda \tag{241}$$

and the death rate

$$\mu_i = i(1-p)q + \mu \tag{242}$$

let $p_{i,i+1}, p_{i,i-1}$ denote the one-step transition probability of the underlying discrete-time Markov chain, then

$$p_{i,i+1} = \frac{\lambda_i}{q_i} = \frac{ipq+\lambda}{iq+\lambda+\mu}, p_{i,i-1} = \frac{\mu_i}{q_i} = \frac{i(1-p)q+\mu}{iq+\lambda+\mu} \tag{243}$$

(2):

This continuous-time Markov chain has infinitely many states and q_i is not bounded so the only way to prove regularity is to investigate the recurrence of the underlying discrete-time Markov chain $\{Y_n\}$, which is a discrete-time birth death chain.

It's clear that the discrete-time birth death chain is recurrent iff $\sum_{l=1}^{\infty}\prod_{i=1}^{l}\frac{p_{i,i-1}}{p_{i,i+1}} = \infty$.

Plug in the transition probability from part (1) to see

$$\sum_{l=1}^{\infty}\prod_{i=1}^{l}\frac{p_{i,i-1}}{p_{i,i+1}} = \sum_{l=1}^{\infty}\prod_{i=1}^{l}\frac{\frac{1}{2}iq+\mu}{\frac{1}{2}iq+\lambda} = \sum_{l=1}^{\infty}1 = \infty \tag{244}$$

proves the regularity of $\{X_t\}$.

(3):

The continuous-time Markov chain is regular if $\sum_{l=1}^{\infty}\prod_{i=1}^{l}\frac{p_{i,i-1}}{p_{i,i+1}} = \infty$, that's to say,

$$\sum_{l=1}^{\infty}\prod_{i=1}^{l}\frac{i(1-p)q+\mu}{ipq+\lambda} = \infty \tag{245}$$

it's clear that when $p < \frac{1}{2}$, regardless of μ, λ, $i(1-p)q+\mu$ will be larger than $ipq+\lambda$ for large enough i, so the chain is regular. When $p = \frac{1}{2}$, clearly the regularity depends on the value of μ, λ, when $\mu \geq \lambda$, the chain is regular.

As a result, a sufficient condition for the chain to be regular is that $p < \frac{1}{2}$ or $p = \frac{1}{2}, \mu \geq \lambda$.

Intuitively, a chain is regular if no explosion happens, i.e. the number of individuals in the system is not very large. In the case where $p < \frac{1}{2}$, the population within the system tends to vanish. In this case, immigration and emigration effects are negligible since the birth rate is dominated by the part ipq that grows as i grows. However, in the case where $p = \frac{1}{2}$, i.e. birth and death of individuals within the system is balanced, immigration and emigration effects play a role. When immigration rate is no larger than emigration rate, there is no explosion in the population.

□

Lemma 26 (Example)**.** *$\{X_t\}$ is a continuous-time birth death chain on $\mathbb{N}$ with birth rate λ_n and death rate μ_n. Assume $X_0 = i \geq 0$ and T_{i+1} is the first hitting time to the value $i+1$ of the process. Derive a recursive formula calculating $m_i \overset{def}{=} \mathbb{E}_i T_{i+1} = \mathbb{E}(T_{i+1}|X_0 = i)$.*

Proof. The recursive formula is derived by applying the Markov property. However, we have to set up the condition in a smart way. One typical choice is to condition on the first holding time H_1 so that we can discuss if the chain goes up or goes down at the next transition

$$\mathbb{E}_i T_{i+1} = \mathbb{E}_i[\mathbb{E}_i(T_{i+1}|H_1)] \tag{246}$$

Keep in mind that the holding rate $q_i = \lambda_i + \mu_i$ so $H_1 \sim \mathcal{E}(q_i)$ since $X_0 = i$. Conditioning on seeing the transition happens at a certain time, the transition is either a birth or a death with probability $\frac{\lambda_i}{\lambda_i+\mu_i}$ and $\frac{\mu_i}{\lambda_i+\mu_i}$ respectively.

At this point, it suffices to calculate

$$\mathbb{E}_i(T_{i+1}|H_1 = t) \tag{247}$$

apply the law of total probability and the Markov property when $i \geq 1$

$$\begin{aligned}
\mathbb{E}_i(T_{i+1}|H_1 = t) &= \mathbb{P}_i\left(X_{H_1} = i-1|H_1 = t\right)\mathbb{E}_i(T_{i+1}|H_1 = t, X_{H_1} = i-1) && (248)\\
&\quad + \mathbb{P}_i\left(X_{H_1} = i+1|H_1 = t\right)\mathbb{E}_i(T_{i+1}|H_1 = t, X_{H_1} = i+1) && (249)\\
&= \frac{\mu_i}{\lambda_i+\mu_i}\mathbb{E}_i(T_{i+1}|H_1 = t, X_{H_1} = i-1) + \frac{\lambda_i}{\lambda_i+\mu_i}t && (250)\\
&= \frac{\mu_i}{\lambda_i+\mu_i}(t + \mathbb{E}_{i-1}T_i + \mathbb{E}_i T_{i+1}) + \frac{\lambda_i}{\lambda_i+\mu_i}t && (251)
\end{aligned}$$

the last equation might seem confusing so we make some comments on that. If the state transition from i to $i+1$ happens, the value of T_{i+1} is observed. On the other hand, if the state transition from i to $i-1$ happens, in order to see the chain hit $i+1$, we must first see the chain return back to state i (since it's a birth death chain, each transition changes the state value by one!) and then restart the Markov chain at state i. In other words, t is the time we have spent transiting to state $i-1$, $\mathbb{E}_{i-1}T_i$ is the time we have to wait until the chain goes back to state i and $\mathbb{E}_i T_{i+1}$ is the time we have to wait after restarting the chain at state i. (Refer to the remark for a possible mistake you might make)

At this point, we see that

$$\begin{aligned}
m_i = \mathbb{E}_i T_{i+1} &= \int_0^\infty \mathbb{E}_i(T_{i+1}|H_1 = t)\cdot f_{H_1}(t)\,dt && (252)\\
&= \int_0^\infty (\mu_i(t + m_{i-1} + m_i) + \lambda_i t)\cdot e^{-q_i t}\,dt && (253)\\
&= \frac{1}{q_i} + \frac{\mu_i}{q_i}(m_{i-1} + m_i) && (254)
\end{aligned}$$

after simplification

$$m_i = \frac{1}{\lambda_i} + \frac{\mu_i}{\lambda_i} m_{i-1} \tag{255}$$

the recursive formula still need an initial condition. It's not hard to see that when $i = 0$,

$$m_0 = \mathbb{E}_0 T_1 = \frac{1}{\lambda_0} \tag{256}$$

since T_1 is the time until next birth when $X_0 = 0$, $T_1 \sim \mathcal{E}(\lambda_0)$. As a result,

$$\begin{cases} m_0 = \frac{1}{\lambda_0} \\ m_i = \frac{1}{\lambda_i} + \frac{\mu_i}{\lambda_i} m_{i-1} \ (i \geq 1) \end{cases} \tag{257}$$

□

Remark. *A common mistake one might make is the following. When $i \geq 1$,*

$$\mathbb{E}_i(T_{i+1}|H_1 = t) = \mathbb{P}_i\left(X_{H_1} = i-1 | H_1 = t\right) \mathbb{E}_i(T_{i+1}|X_{H_1} = i-1, H_1 = t) \tag{258}$$

$$+ \mathbb{P}_i\left(X_{H_1} = i+1 | H_1 = t\right) \mathbb{E}_i(T_{i+1}|X_{H_1} = i+1, H_1 = t) \tag{259}$$

$$= \frac{\mu_i}{\lambda_i + \mu_i}(t + \mathbb{E}_{i-1} T_{i+1}) + \frac{\lambda_i}{\lambda_i + \mu_i} t \tag{260}$$

the equation is correct but far from what we want since there's no way to represent $\mathbb{E}_{i-1} T_{i+1}$ using m_i.

The trick here is to consider ***the excursion w.r.t. state*** *i since any possibility of going from state $i-1$ to state $i+1$ shall pass state i for birth death chain (increments must take value ± 1)!*

Lemma 27 (Example)**.** $\{X_n\}$ *is discrete-time Markov chain on* $\mathbb{Z}$ *with no absorbing states,* $X_0 = 0$*. Define* $T_0 = 0$*,*

$$T_m = \inf\left\{n \geq T_{m-1} : X_n \neq X_{T_{m-1}}\right\} \ (m \in \mathbb{N}, m \geq 1) \tag{261}$$

as the time when Markov chain X_n *changes its state for the* m*-th time.*

(1): Show that $\forall m, T_m$ *are stopping times.*

(2): Show that $Z_n = X_{T_n}$ *is a Markov chain, derive its transition probability.*

Proof. (1): By the definition,

$$\{T_1 = k\} = \{X_0 = X_1 = ... = X_{k-1} = 0, X_k \neq 0\} \tag{262}$$

is an event only depending on $X_1, ..., X_k$ so T_1 is a stopping time. Notice that $T_0 < T_1 < T_2 < ...$ so

$$\{T_2 = k\} = \bigcup_{j=1}^{k-1} \{T_1 = j, T_2 = k\} = \bigcup_{j=1}^{k-1} \{X_0 = ... = X_{j-1} = 0, X_j \neq 0, X_j = X_{j+1} = ... = X_{k-1}, X_k \neq X_{k-1}\} \tag{263}$$

still only depends on $X_1, .., X_k$ so T_2 is a stopping time. An induction argument concludes the proof.

(2): Let's first prove Markov property

$$\begin{aligned}
&\forall i_0, ..., i_n, \mathbb{P}\left(Z_n = i_n | Z_{n-1} = i_{n-1}, ..., Z_0 = i_0\right) && (264)\\
&= \mathbb{P}\left(X_{T_n} = i_n | X_{T_{n-1}} = i_{n-1}, ..., X_{T_0} = i_0\right) && (265)\\
&= \mathbb{P}\left(X_{T_n} = i_n | X_{T_{n-1}} = i_{n-1}\right) && (266)\\
&= \mathbb{P}\left(Z_n = i_n | Z_{n-1} = i_{n-1}\right) && (267)
\end{aligned}$$

by strong Markov property of $\{X_n\}$ since $T_n, T_{n-1}, ..., T_0$ are stopping times and there is no absorbing states so $T_{n-1} < \infty$ happens with probability one.

For its transition probability,

$$\begin{aligned}
p_{ij}^Z &= \mathbb{P}\left(Z_{n+1} = j | Z_n = i\right) && (268)\\
&= \mathbb{P}\left(X_{T_{n+1}} = j | X_{T_n} = i\right) && (269)
\end{aligned}$$

is zero if $i = j$. In the case where $i \neq j$,

$$\begin{aligned}
&\mathbb{P}\left(X_{T_{n+1}} = j | X_{T_n} = i\right) && (270)\\
&= \mathbb{P}\left(X_{T_n+1} = ... = X_{T_{n+1}-1} = i, X_{T_{n+1}} = j | X_{T_n} = i\right) && (271)\\
&= \sum_{l=1}^{\infty} \mathbb{P}\left(T_{n+1} - T_n = l, X_{T_n+1} = ... = X_{T_n+l-1} = i, X_{T_{n+1}} = j | X_{T_n} = i\right) && (272)\\
& && (273)
\end{aligned}$$

here $\left\{X_{T_n+1} = ... = X_{T_n+l-1} = i, X_{T_{n+1}} = j\right\}$ is already implying $\{T_{n+1} - T_n = l\}$ so

$$\mathbb{P}\left(X_{T_{n+1}} = j | X_{T_n} = i\right) \tag{274}$$

$$= \sum_{l=1}^{\infty} \mathbb{P}\left(X_{T_n+1} = ... = X_{T_n+l-1} = i, X_{T_{n+1}} = j | X_{T_n} = i\right) \tag{275}$$

$$= \sum_{l=1}^{\infty} (p_{ii}^X)^{l-1} p_{ij}^X \tag{276}$$

$$= \frac{p_{ij}^X}{1 - p_{ii}^X} \tag{277}$$

$$= \frac{p_{ij}^X}{\sum_{k \neq i} p_{ik}^X} \tag{278}$$

finally we see the transition probability

$$p_{ij}^Z = \begin{cases} \frac{p_{ij}^X}{\sum_{k \neq i} p_{ik}^X} & \text{if } i \neq j \\ 0 & \text{else} \end{cases} \tag{279}$$

□

Lemma 28 (Example)**.** *When it's sunny, wildfire occurs with rate* 0.5 *per day, when it's cloudy, wildfire occurs with rate* 0.1 *per day. Model the weather as a two-state continuous-time Markov chain with sunny weather lasting on average for* 2 *days and cloudy weather lasting on average for* 1 *day. Wildfires happens at arrival time* $T_1, T_2, ...$ *of a continuous-time counting process and the weather is assumed to be independent of the wildfire occuring.* F_t *denote the number of wildfires happening up to time* t*.*

(1): Compute $\lim_{t\to\infty} \frac{\mathbb{E}F_t}{t}$*.*

(2): Explain why $X_t = (S_t, F_t)$ *is a Markov process and write down its generator matrix.*

(3): Suppose it's sunny now, find the expected time until next wildfire $\mathbb{E}T_1$*. Compare with the answer from (1).*

Proof. (1): The weather at time t is S_t with state space $S = \{s, c\}$ for sunny and cloudy. It's clear that if it's sunny now, the time until the next cloudy day follows $\mathcal{E}(q_s)$ with mean $\frac{1}{q_s} = 2$. Similarly, $\frac{1}{q_c} = 1$ so the holding rates are

$$q_s = 0.5, q_c = 1 \tag{280}$$

now the transition law of F_t depends on the value of S_t.

It's also clear that only the behavior of $\mathbb{E}F_t$ when t is large enough matters. So we care about the limiting distribution of F_t which depends on the limiting distribution of S_t. Since $\{S_t\}$ is irreducible, by ergodic theorem if it has a stationary distribution, it must also be the limiting distribution.

Write out the generator matrix of $\{S_t\}$,

$$G = \begin{bmatrix} -0.5 & 0.5 \\ 1 & -1 \end{bmatrix} \tag{281}$$

compute $\pi G = 0$ to get $\pi_s = \frac{2}{3}, \pi_c = \frac{1}{3}$, so the limiting distribution of S_t is known, i.e. when t is large enough, $\frac{2}{3}$ of the time it's sunny and $\frac{1}{3}$ of the time it's cloudy.

As a result, when t is large enough, $\frac{2}{3}$ of the time F_t is a Poisson process with intensity 0.5 and $\frac{1}{3}$ of the time F_t is a Poisson process with intensity 0.1

$$\lim_{t\to\infty} \frac{\mathbb{E}F_t}{t} = \lim_{t\to\infty} \frac{\frac{2}{3}t \cdot 0.5 + \frac{1}{3}t \cdot 0.1}{t} = \frac{1}{3} + \frac{1}{30} = \frac{11}{30} \tag{282}$$

(2): If $X_t = (s, n)$, since $\{F_t\}$ is a Poisson process (its intensity depends on $\{S_t\}$), it's only possible to transit to (c, n) or $(s, n+1)$. The holding time is the minimum of the holding time of state s in $\{S_t\}$ and the holding time of state n in $\{F_t\}$. Those two holding times are independent, the former holding time has distribution $\mathcal{E}(0.5)$ and the latter one has distribution $\mathcal{E}(0.5)$ so taking minimum gives the holding time of state (s, n) with distribution $0.5 + 0.5 = 1, q_{(s,n)} = 1$. Similarly, we can derive the holding rate of all states

$$q_{(s,n)} = 0.5 + 0.5 = 1, q_{(c,n)} = 1 + 0.1 = 1.1 \tag{283}$$

for the underlying discrete-time Markov chain $\{Y_n\}$ with state space$\{s, c\} \times \mathbb{N}$, the transition probability is

$$p_{(s,n),(c,n)} = 0.5, p_{(s,n),(s,n+1)} = 0.5, p_{(c,n),(s,n)} = \frac{10}{11}, p_{(c,n),(c,n+1)} = \frac{1}{11} \tag{284}$$

now we are clear with the dynamics of $\{X_t\}$, it's clear that all holding times are independent and are still exponentially distributed (memoryless), that's why it's still a Markov chain.

By definition, write down the generator matrix of $\{X_t\}$, notice that states in the row and column appear in the order of $(s,0),(s,1),(s,2),...,(c,0),(c,1),(c,2),...$

$$G = \begin{bmatrix} -1 & \frac{1}{2} & 0 & ... & \frac{1}{2} & 0 & 0 & ... \\ 0 & -1 & \frac{1}{2} & ... & 0 & \frac{1}{2} & 0 & ... \\ ... & ... & ... & ... & ... & ... & & \\ 1 & 0 & 0 & ... & -1.1 & 0.1 & 0 & ... \\ 0 & 1 & 0 & ... & 0 & -1.1 & 0.1 & ... \\ ... & ... & ... & ... & ... & ... & & \end{bmatrix} \tag{285}$$

(3): Now it's sunny so $X_0 = (s,0)$, consider the first holding time of $\{X_t\}$ denoted R_1, then $R_1 \sim \mathcal{E}(1)$ if $X_0 = (s,0)$ and $R_1 \sim \mathcal{E}(1.1)$ if $X_0 = (c,0)$. Condition on R_1 to calculate $\mathbb{E}_{(s,0)}T_1$.

$$\mathbb{E}_{(s,0)}T_1 = \mathbb{E}_{(s,0)}[\mathbb{E}_{(s,0)}(T_1|R_1)] \tag{286}$$

it suffices to calculate $\mathbb{E}_{(s,0)}(T_1|R_1 = r)$, by the law of total probability and Markov property

$$\begin{aligned} \mathbb{E}_{(s,0)}(T_1|R_1 = r) &= \mathbb{P}_{(s,0)}\left(X_{R_1} = (s,1)|R_1 = r\right)\mathbb{E}_{(s,0)}(T_1|X_{R_1} = (s,1), R_1 = r) && (287)\\ &+ \mathbb{P}_{(s,0)}\left(X_{R_1} = (c,0)|R_1 = r\right)\mathbb{E}_{(s,0)}(T_1|X_{R_1} = (c,0), R_1 = r) && (288)\\ &= \frac{1}{2}r + \frac{1}{2}[r + \mathbb{E}_{(c,0)}T_1] && (289) \end{aligned}$$

naturally, set up another equation for $\mathbb{E}_{(c,0)}T_1$

$$\begin{aligned} \mathbb{E}_{(c,0)}(T_1|R_1 = r) &= \mathbb{P}_{(c,0)}\left(X_{R_1} = (c,1)|R_1 = r\right)\mathbb{E}_{(c,0)}(T_1|X_{R_1} = (c,1), R_1 = r) && (290)\\ &+ \mathbb{P}_{(c,0)}\left(X_{R_1} = (s,0)|R_1 = r\right)\mathbb{E}_{(c,0)}(T_1|X_{R_1} = (s,0), R_1 = r) && (291)\\ &= \frac{1}{11}r + \frac{10}{11}[r + \mathbb{E}_{(s,0)}T_1] && (292) \end{aligned}$$

apply the law of iterated expectation to get

$$\mathbb{E}_{(s,0)}T_1 = \int_0^\infty \left(\frac{1}{2}r + \frac{1}{2}[r + \mathbb{E}_{(c,0)}T_1]\right)e^{-r}\,dr = 1 + \frac{1}{2}\mathbb{E}_{(c,0)}T_1 \tag{293}$$

$$\mathbb{E}_{(c,0)}T_1 = \int_0^\infty \left(\frac{1}{11}r + \frac{10}{11}[r + \mathbb{E}_{(s,0)}T_1]\right)1.1e^{-1.1r}\,dr = \frac{10}{11} + \frac{10}{11}\mathbb{E}_{(s,0)}T_1 \tag{294}$$

solve this linear system to get

$$\mathbb{E}_{(s,0)}T_1 = \frac{8}{3}, \mathbb{E}_{(c,0)}T_1 = \frac{10}{3} \tag{295}$$

Compare with the answer from part (1), in long term, the rate of wildfire is $\frac{11}{30}$ so the expected waiting time until the next wildfire can be estimated by $\frac{30}{11} \approx 2.727$. If it's sunny now, the expected waiting time until the next wildfire is $\frac{8}{3} \approx 2.667$ which is smaller since wildfire is more likely to happen in sunny days. If it's cloudy now, the expected waiting time until the next wildfire is $\frac{10}{3} \approx 3.333$ which is larger since wildfire is less likely to happen in cloudy days.

□