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.

$$\begin{aligned}
\mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)} \cap \bigcup_{i=1}^8 S_2^{(i)}\right) &= \mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)}\right) - \mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)} \cap \left[\bigcup_{i=1}^8 S_2^{(i)}\right]^c\right) &(15)\\
&= \mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)}\right) - \mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)} \cap \bigcap_{i=1}^8 \left[S_2^{(i)}\right]^c\right) &(16)\\
&= \mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)}\right) - \mathbb{P}\left(\bigcap_{i=1}^8 E_3^{(i)} \cap \bigcap_{i=1}^8 E_2^{(i)}\right) &(17)
\end{aligned}$$

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

$$\begin{aligned} G_T(s) &= \sum_{k=0}^{\infty} p_k \cdot s \cdot \mathbb{E}(s^{T_1} ... s^{T_k} \cdot \mathbb{I}_{T_1 < \infty} ... \mathbb{I}_{T_k < \infty}) && (27) \\ &= s \cdot \sum_{k=0}^{\infty} p_k \cdot \mathbb{E}(s^{T_1} \cdot \mathbb{I}_{T_1 < \infty}) ... \mathbb{E}(s^{T_k} \cdot \mathbb{I}_{T_k < \infty}) && (28) \\ &= s \cdot \sum_{k=0}^{\infty} p_k \cdot [G_{T_1}(s)]^k && (29) \\ &= s \cdot G(G_{T_1}(s)) && (30) \end{aligned}$$

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

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

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 $p_{ij}^n(1) = \mathbb{P}\left(X_{n+1} = j | X_n = i\right)$ 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 $\mathbb{P}\left(X_{n+1} = j | X_n = i\right)$ 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

$$\mathbb{P}\left(Z_n = (i_n, i_{n+1}) | Z_0 = (i_0, i_1), ..., Z_{n-1} = (i_{n-1}, i_n)\right) \tag{36}$$

$$= \mathbb{P}\left(X_n = i_n, X_{n+1} = i_{n+1} | X_0 = i_0, X_1 = i_1, ..., X_n = i_n\right) \tag{37}$$

$$= \mathbb{P}\left(X_{n+1} = i_{n+1} | X_0 = i_0, X_1 = i_1, ..., X_n = i_n\right) \tag{38}$$

$$= \mathbb{P}\left(X_{n+1} = i_{n+1} | X_{n-1} = i_{n-1}, X_n = i_n\right) \tag{39}$$

$$= \mathbb{P}\left(X_n = i_n, X_{n+1} = i_{n+1} | X_{n-1} = i_{n-1}, X_n = i_n\right) \tag{40}$$

$$= \mathbb{P}\left(Z_n = (i_n, i_{n+1}) | Z_{n-1} = (i_{n-1}, i_n)\right) \tag{41}$$

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 Cahpman-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

$$\begin{aligned}\mathbb{P}\left(Y_2=1,Y_4=1,....,Y_{2k}=1\right)&=\mathbb{P}\left(Y_1Y_3=1,Y_3Y_5=1,....,Y_{2k-1}Y_{2k+1}=1\right) &(46)\\ &=\mathbb{P}\left(Y_1=1,Y_3=1,...,Y_{2k+1}=1\right)+\mathbb{P}\left(Y_1=-1,Y_3=-1,...,Y_{2k+1}=-1\right) &(47)\\ &=\frac{1}{2^{k+1}}+\frac{1}{2^{k+1}}=\frac{1}{2^k}=\mathbb{P}\left(Y_2=1\right)\mathbb{P}\left(Y_4=1\right)...\mathbb{P}\left(Y_{2k}=1\right) &(48)\end{aligned}$$

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

$$\begin{aligned}\mathbb{P}\left(Y_1=1,Y_{2k}=1\right)&=\mathbb{P}\left(Y_1=1,Y_{2k-1}Y_{2k+1}=1\right) &(49)\\ &=\mathbb{P}\left(Y_1=1,Y_{2k-1}=1,Y_{2k+1}=1\right)+\mathbb{P}\left(Y_1=1,Y_{2k-1}=-1,Y_{2k+1}=-1\right) &(50)\\ &=\frac{1}{8}+\frac{1}{8}=\frac{1}{4}=\mathbb{P}\left(Y_1=1\right)\mathbb{P}\left(Y_{2k}=1\right) &(51)\end{aligned}$$

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

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

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

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

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

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

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

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

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

concludes the proof.

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

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

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,

$$\begin{aligned}\mathbb{P}\left(Y_{n+1} = y_{n+1}|Y_0 = y_0, ..., Y_n = y_n\right) &= \mathbb{P}\left(h(X_{n+1}) = y_{n+1}|h(X_0) = y_0, ..., h(X_n) = y_n\right) &(73)\\ &= \mathbb{P}\left(X_{n+1} = x_{n+1}|X_0 = x_0, ..., X_n = x_n\right) &(74)\\ &= \mathbb{P}\left(X_{n+1} = x_{n+1}|X_n = x_n\right) &(75)\end{aligned}$$

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

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

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.

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

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 starts from $X_{T_s} = s$. 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 $\frac{S_n}{\sqrt{n}} \overset{d}{\to} N(\mu, \sigma^2)$ 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 $S = \left\{1, 2, ..., 2^N - 1\right\}$ 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), $\mathbb{P}(X_{2k} = 0 \ i.o.) = 0$ 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

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

to get the expression of $u(k)$, it's again summing both sides for $i \in \{k, k+1, ..., j-1\}$

$$\begin{align} u(j) - u(k) &= -\sum_{i=k}^{j-1} \frac{1}{\sum_{l=i}^{j-1} \frac{P_l}{P_i}} \tag{116} \\ &= -\sum_{i=k}^{j-1} \frac{P_i}{\sum_{l=i}^{j-1} P_l} \tag{117} \end{align}$$

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

$$\begin{align} \mathbb{P}_0 (F_0 < \infty) &= p_0 \mathbb{P}_0 (F_0 < \infty | X_1 = 1) + r_0 \mathbb{P}_0 (F_0 < \infty | X_1 = 0) \tag{119} \\ &= p_0 \mathbb{P}_1 (F_0 < \infty) + r_0 \tag{120} \end{align}$$

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.*