1 Probability revision

1.1 Common random variables

Bernoulli distribution *Ber(p)*:
$P(X = x) = p^x(1-p)^{1-x}, x \in \{0,1\}, \text{mean} = p, \text{variance} = p(1-p)$
Binomial distribution *Bin(n,p)*:
$P(X = x) = \binom{n}{x}p^x(1-p)^{n-x}, x \in \mathbb{Z} \cap [0,n], \text{mean} = np, \text{variance} = np(1-p)$
Poisson distribution *Pois(λ)*:
$P(X = x) = e^{-\lambda}\frac{\lambda^x}{x!}, x \in \mathbb{Z}^{\geq 0}, \text{mean} = \lambda, \text{variance} = \lambda$

1.2 Property of expectation

If $X \geq 0$, then $E[X] \geq 0$
If $X \geq 0$ and $E[X] = 0$, then $P(X = 0) = 1$
If a and b are constants, then $E[a+bX] = a+bE[X]$
If X and Y are random variables, then $E[X+Y] = E[X] + E[Y]$
If X is a non-negative random variable that takes integral value, then we have $E[X] = \sum_{i=1}^{\infty} P(X \geq i)$

1.3 Property of variance

If a and b are constants, then $Var(a + bX) = b^2Var(X)$
$Var(X) = E[X^2] - (E[X])^2$

1.4 Moment generating function

It is the unique identifier of a random variable.
$M_X(t) = E(e^{tX})$
Calculate k-th moment of X:
$E[X^k] = \dfrac{d^k}{dt^k}M_X(t)|_{t=0}$
Linear transformation:
$M_{aX+b}(t) = e^{bt}M_X(at)$

1.5 Independence

Discrete random variables $X_1, \ldots, X_n$ are independent if for any $x_1, \ldots, x_n, P(X_1 = x_1, \ldots, X_n = x_n) = \Pi_{i=1}^n P(X_i = x_i)$
If $f_1, \ldots, f_n$ are functions from $\mathbf{R}$ to $\mathbf{R}$, then $Y_i = f_i(X_i)$ are also random variables, and they preserve independence.
If X_i's are independent, then $E[\Pi_{i=1}^n X_i] = \Pi_{i=1}^n E[X_i]$
If $Z = \Sigma_{i=1}^n X_i$, then $Var(Z) = \Sigma_{i=1}^n Var(X_i)$, and $M_Z(t) = \Pi_{i=1}^n M_{X_i}(t)$

1.6 Conditional probability

Discrete: $p_{X|Y}(x|y) = \dfrac{P(X = x, Y = y)}{P(Y = y)}$
Continuous: $f_{X|Y}(x|y) = \dfrac{f(x,y)}{f_Y(y)}$
Multiplication law: $p_{X,Y}(x,y) = p_{X|Y}(x|y)p_Y(y) = p_{Y|X}(y|x)p_X(x)$
Law of total probability:$p_X(x) = \Sigma_y p_{X|Y}(x|y)p_Y(y)$
Bayes' formula: $p_{Y|X}(y|x) = \dfrac{p_{X|Y}(x|y)p_Y(y)}{\Sigma_y p_{X|Y}(x|y)p_Y(y)}$
Conditional independence: $P(X = x, Y = y|Z = z) = P(X = x|Z = z)P(Y = y|Z = z)$
Law of total expectation: $E[X] = E[E[X|Y]]$

2 Markov Chain preliminaries

2.1 Definition

A stochastic process with discrete state space S and discrete time set T satisfying the Markovian property.

2.2 Markovian property

Given X_n, what happened afterwards ($t > n$) is independent with what happend before ($t < n$). Mathematically, for any set of state $i_0, \ldots, i_{n-1}, i, j \in S$ and $n \geq 0$, $P(X_{n+1} = j|X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = P(X_{n+1} = j|X_n = i)$

2.3 One-step transition probability

Let $p_{i,j}^{n,n+1} = P(X_{n+1} = j|X_n = i)$ denote the one-step transition probability.

3 First step analysis

3.1 Initial distribution

Suppose the head of the Markov Chain X_0 follows a certain distribution π_0, then $X_n|X_0 \sim \pi_0$ will follow a distribution of $\pi_0\mathbf{P}^n$.

3.2 Terminologies

Absorbing state: A state i which satisfies that $p_{i,j} = 0\ \forall j \neq i$.
Stopping time: $T = \min\{n \geq 0 : X_n = i\}$

3.3 Trick: move one step forward

Consider the case of gambler's ruin with winning probability p. Intuition suggests that $P(X_T = 0|X_1 = 2) = P(X_T = 0|X_0 = 2)$
Define the process $Y_n = X_{n+1}$, hence $P(X_T = 0|X_1 = 2) = P(Y_{T-1} = 0|Y_0 = 2)$, then we have $Y_{T-1} = X_T = 0 \implies T_Y = T - 1$. Since $\{Y_n\}$ and $\{X_n\}$ have the same probabilistic structure, we have $P(X_T = 0|X_1 = 2) = P(Y_{T-1} = 0|Y_0 = 2) = P(Y_{T_Y} = 0|Y_0 = 2) = P(X_T = 0|X_0 = 2)$, which completes our proof. By induction, the identity could be generalised from 1 step to any finite k steps. Similar process could be applied to derive the relation $E[T|X_0 = i] + 1 = E[T|X_1 = i]$

4 Classification of states

4.1 Accessibility

Definition: State j is accessible from state i if $\mathbf{P}_{i,j}^m > 0$ for some $m > 0$. Note that state i is accessible from itself since we define $\mathbf{P}^0 = \mathbf{I}$. If state i is also accessible from state j, then they communicate with each other, and we denote by $i \leftrightarrow j$.

4.2 Communication class

Each communication class is an equivalence class defined by the communication relation. It contains all states that are communicating with each other.

4.3 Reducible chain

Definition: An MC is irreducible if all states communicate with each other, otherwise reducible.

4.4 Return probability

Probability that starting from state i, and revisit state i at the n-th step, which is $P_{i,i}^n$. If state i is transient, then $P_{i,i}^n \to 0$ when $n \to +\infty$, which means that in long term we will never revisit state i.

4.5 First return probability

The probability that starting from state i, the first revisit to state i occurs at the n-th step. Mathematically, we denote by $f_{ii}^n = P(X_1 \neq i, X_2 \neq i, \ldots, X_{n-1} \neq i, X_n = i|X_0 = i)$. We define $f_{ii}^0 = 0$. Since first revisit at different steps are disjoint events, and if a state is transient, we may not return to it in long run, we naturally have $P(\text{first revisit at step n} < +\infty) = \sum_{n=1}^{\infty} f_{ii}^n \leq 1$, and $f_{ii}^n \leq P_{ii}^n\ \forall n \in \mathbf{Z}^{>0}$.
Condition on the first revisit, we have $P_{ii}^n = \sum_{k=0}^n f_{ii}^k P_{ii}^{n-k}$.

4.6 Recurrent and transient states

Consider $f_{ii} = \sum_{n=0}^{\infty} f_{ii}^n = \lim_{N\to\infty}\sum_{n=0}^{N} f_{ii}^n$. It denotes the probability of revisiting i in the future.
Definition: A state i is recurrent if $f_{ii} = 1$, and transient if $f_{ii} < 1$.
If the flow will revisit state i for sure, then state i is recurrent. Otherwise, the flow will finally leave state i and never come back.
Let N_i denote the number of times that the flow revisits state i. $N_i = \sum_{n=1}^{\infty} I(X_n = i)$, then recurrence means $E[N_i|X_0 = i] = \infty$, and transient means $E[N_i|X_0 = i] < \infty$.
Theorem: For a state i, consider the expected number of visits to i, $E[N_i|X_0 = i]$, then if $f_{ii} < 1$(i.e.i is transient), then $E[N_i|X_0 = i] = \frac{f_{ii}}{1-f_{ii}}$. If $f_{ii} = 1$(i.e.i is recurrent), then $E[N_i|X_0 = i] = \infty$.
Note that $E[N_i|X_0 = i] = E[\sum_{n=1}^{\infty} I(X_n = i|X_0 = i)] = \sum_{n=1}^{\infty} E[I(X_n = i|X_0 = i)] = \sum_{n=1}^{\infty} P(X_n = i|X_0 = i) = \sum_{n=1}^{\infty} P_{ii}^n$.
Theorem: State i is recurrent iff $\sum_{n=1}^{\infty} P_{ii}^n = \infty$
For recurrent states, it means a significant probability to return. For transient states, the series will converge to 0, which means that the probability of coming back vanishes in the long run.

4.7 Summary of recurrence

Recurrent $\iff f_{ii} = 1 \iff \sum_{n=1}^{\infty} P_{ii}^n = \infty \iff E[N_i|X_0 = i] = \infty$
Transient $\iff f_{ii} < 1 \iff \sum_{n=1}^{\infty} P_{ii}^n < \infty \iff E[N_i|X_0 = i] < \infty$

4.8 States of the same class

Theorem: States in the same communication class are either all transient or all recurrent. We could interpret transient/recurrent status as a status for the whole class. An MC with finite states must have at least one recurrent class.

5 Long run performance

5.1 Periodicity

Definition: For a state i, consider $\{n \geq 1 : P_{ii}^n > 0\}$, we define the period of state i, $d(i)$, as the greatest common divisor of the set. If the set is empty, we define $d(i) = 0$. If $d(i) = 1$, we say that state i is aperiodic.
Theorem: If $i \leftrightarrow j$, then $d(i) = d(j)$.
Theorem: $\exists N$ such that $P_{ii}^{Nd(i)} > 0$, and $\forall n \geq N$, $P_{ii}^{nd(i)} > 0$.
Theorem: If $\exists m > 0$ such that $P_{ji}^m > 0$, then for sufficiently large n, we have $P_{ji}^{m+nd(i)} > 0$.

5.2 Regular Markov Chain

Definition: An MC is regular if $\exists k > 0$, such that all elements of $\mathbf{P}^k$ are strictly positive.
It means that the flow can achieve any state from any state at step k.
Theorem: If a Markov Chain is irreducible, aperiodic and with finite states, then it is regular.

5.3 Main theorem

Theorem: Suppose $\mathbf{P}$ is a regular transition probability matrix with states $S = \{1, 2, \ldots, N\}$, then
1. $\lim_{n\to\infty} p_{ij}^n$ exists.
2. The limit does not depend on i. Hence, we can denote it by $\pi_j = \lim_{n\to\infty} p_{ij}^n$.
3. $\sum_{k=1}^N \pi_k = 1$. We call it as the limiting distribution.
4. The limits $\pi = (\pi_1, \pi_2, \ldots, \pi_N)$ are the solutions of the system $\pi_j = \sum_{k=1}^N \pi_k P_{kj}, j = 1, 2, 3, \ldots, N, \sum_{k=1}^N \pi_k = 1$. In matrix form, it is to solve $\pi P = \pi, \sum_{k=1}^N \pi_k = 1$
5. The limiting distribution π is unique.

5.4 Stationary distribution

When the MC is not regular, it is possible that $|S| = \infty$ and $\pi = 0$, but it is also possible that we can still find a non-trivial π. It means, if the initial states have a distribution π, then after any steps, the chain also has a distribution π on the states, which is called the stationary distribution.
Definition: Consider a Markov Chain with state space $S = \{1, 2, \ldots\}$ and the transition probability matrix P. A distribution $(p_1, p_2, \ldots)$ on S is called a stationary distribution, if it satisfies that if $P(X_n = i) = p_i, i = 1, 2, \ldots$, then $P(X_{n+1} = i) = p_i, i = 1, 2, \ldots$
It means that if the initial states have a distribution π, then after any steps, the chain still has a distribution π on the states. Note that a irregular MC may have more than one stationary distributions. To solve for possible π, note that here we also have to solve the system $\pi P = \pi$

5.5 Basic limit theorem

Suppose now we still have a recurrent MC, but with infinite states. It has a chance that although we know the return probability $f_{ii} = 1$, but the expectation of return time is infinity, which in practice is still impossible to return.
We let $Ri = \min\{n \geq 1 : X_n = i\}$, then $m_i = E[R_i|X_0 = i] = \sum_{n=1}^{\infty} n f_{ii}^{(n)}$, even if the state is recurrent, this expectation could still diverge. In the case of divergence, we call the MC null recurrent, as infinite return time means zero limiting probability. In the case of convergence, we call the MC positive recurrent. If aperiodic, then there is a limiting distribution. If periodic, we can only consider distribution at step nd.
The formal results are summarised as follows: For a *positive recurrent, irreducible, and aperiodic MC*:

1. Limiting distribution π exists.
2. $\pi_j = \frac{1}{m_j}$
3. Such MC is called *ergodic* MC.

If MC is periodic, we still have $\lim_{n\to\infty} P_{jj}^{(nd)} = \frac{d}{m_j}$

6 Application: Branching process

6.1 Definition

Suppose there is one monkey at the starting point, at each generation, the monkey can give birth to ξ monkeys, and the old monkey will die. Let X_n denote the number of monkeys at n^{th} generation, then $X_0 = 1$. At n^{th} generation, monkey i will generate

ξ_i offsprings. Each of ξ_i follows identical and independent distribution F. Therefore, we can formulate $X_{n+1} = \sum_{i=1}^{X_n} \xi_i$, a random sum. $\{X_n\}$ is a stationary MC, and we only need the distribution of ξ to specify the process.

6.2 Possible variations

First, the parent monkey may not die, then treat the parent monkey as an offspring of itself and then let it enter the next generation.
Second, $X_0 > 1$. The branching trees of different ancestor monkeys are independent. The analysis result could be easily derived by a multiplication with the result of single ancestor.

6.3 Analysis with partial information

Suppose we do not know the distribution F, but we know that $E[\xi] = \mu$, $Var[\xi] = \sigma^2$. We know that for a random sum,

$$E[X_{n+1}] = E[X_n]E[\xi_i^{(n)}] = \mu E[X_n]$$

$$Var[X_{n+1}] = \mu^2 Var[X_n] + \sigma^2 EX_n$$

Since we derive a recurrence relation between X_{n+1} and X_n, by induction we go back to X_0(base case with constant expectation and zero variance) and we have

$$E[X_n] = \mu^n;\ Var[X_n] = \mu^{n-1}\sigma^2 k$$

where $k = \frac{1-\mu^n}{1-\mu}$ if $\mu \neq 1$, and n otherwise. If $X_0 = c$, multiply by c to get the result.

6.4 Probability generating function

For a discrete random variable X, the probability generating function is defined as

$$\phi_X(t) = E[t^X] = \sum_{k=0}^{\infty} P(X = k)t^k$$

$\phi_X(t)$ is a function about t without any randomness. We have $P(X = 0) = \phi_X(0), P(X = k) = \frac{1}{k!}\frac{d^k}{dx^k}\phi_X(t)|_{t=0}$. The distribution of X is fully characterized by $\phi_X(t)$. If X and Y are independent, then $\phi_{X+Y}(t) = \phi_X(t)\phi_Y(t)$

6.5 Analysis with full information

Suppose we now know the distribution F. Equivalently, we know $\phi_\xi(t)$. Similarly, we want to derive a recurrence relation first and then establish the general result by induction.
Given X_n, we can derive $\phi_{X_{n+1}}(t)$ as follows:

$$\begin{aligned}\phi_{X_{n+1}}(t) &= E[t^{X_{n+1}}] = E[t^{\sum_{i=1}^{X_n}\xi_i}] \\ &= E[\prod_{i=1}^{X_n} t^{\xi_i}] = \prod_{i=1}^{X_n} E[t^{\xi_i}] \\ &= [\phi_\xi(t)]^{X_n}\end{aligned}$$

Therefore, consider X_n as a distribution, we have

$$\begin{aligned}\phi_{X_{n+1}(t)} &= E[\prod_{i=1}^{X_n} t^{\xi_i}] = E[E[\prod_{i=1}^{X_n} t^{\xi_i}|X_n]] \\ &= E[[\phi_\xi(t)]^{X_n}] = \phi_{X_n}(\phi_\xi(t))\end{aligned}$$

Again, by induction, we have $\phi_{X_n}(t) = \phi_\xi^{(n)}(t)$.
If $X_0 = k$, we raise the function to power k to get the result, directly due to property of probability generating function.

6.6 Extinction probability

We consider the case of extinction, which means $X_n = 0$ from a particular generation n (and of course afterwards).
We describe the problem as follows: Let $T = \min\{n : X_n = 0\}$. We are interested in $u_n^{(k)} = P(X_n = 0|X_0 = k) = P(T \leq n|X_0 = k)$. Whether the family will ultimately extinct can be defined by $u_\infty = \lim_{n\to\infty} u_n$. Note that a family of k ancestors going to extinct is equivalent to k families of one ancestor going to extinct, thus $u_n^{(k)} = u_n^k$. The equation still holds when $n \to \infty$ and taking limits of both sides, giving $u_\infty^{(k)} = u_\infty^k$.
By first step analysis, we have

$$\begin{aligned}u_n &= \sum_{l\in S} P(X_1 = l|X_0 = 1)u_{n-1}^{(l)} \\ &= \sum_{l\in S} P(\xi = l)u_{n-1}^{(l)} \\ &= E[u_{n-1}^{\xi}] = \phi_\xi(u_{n-1}).\end{aligned}$$

By induction, we then have $u_n = \phi_\xi^{(n)}(u_0) = \phi_\xi^{(n)}(0)$. According to first step analysis, $u_n = \phi_\xi(u_{n-1})$. Taking limit on both sides, we have $u_\infty = \phi_\xi(u_\infty)$, solve the equation, the solution in $[0, 1]$ is the extinction probability in long run. Geometrically, this is the intercept of $y = x$ and $y = \phi_\xi(x)$. Note that ϕ_ξ is an increasing and convex function on $(0, 1]$, and it passes through $(0, P(\xi = 0))$ and $(1, 1)$, if there is an intersection with x value in (0,1], then the extinction probability is less than 1. We summarise as follows:

1. If $P(\xi = 0) = 0$, then $u_\infty = 0$ (The number of offsprings is always positive, thus the population does not extinct.)
2. If $P(\xi = 0) > 0, E[\xi] \leq 1$, then $u_\infty = 1$.
3. If $P(\xi = 0) > 0, E[\xi] > 1$, then $u_\infty < 1$

7 Application: PageRank

7.1 Definition

We can model browsing webpages by Markov chain: Webpages form a network by hyperlinks. Each webpage denotes a node(state), and each hyperlink is an edge from one to another. We assume that if there are k links from webpage A to other pages, then each link has an equal probability to be accessed.

7.2 PageRank rationale

In the long run, the flow should converge to more important websites, thus the limiting probability(or approximately distribution after many steps) is a good indication of relative importance of websites. Suppose the initial distribution π_0 is uniform, after N steps, the distribution will be $\pi_N = \pi_0\mathbf{P}^N$. If limiting distribution exists, we can also compute and rank the webpages accordingly. **Variation:** To escape from non-ideal absorbing states, we could add perturbation at each step such as : $\pi_{n+1} = (1 - \lambda)\pi_n\mathbf{P} + \lambda\pi_0$

8 Application: MCMC Sampling

8.1 Definition

We wish to simulate sampling process from a given distribution using Markov Chain. We will design such a MC that in the long run, the limiting distribution converges to our target distribution. In this analysis, we assume our target π is Poisson(5).

8.2 Tasks

To design such a MC(which is specified by transition probability $\mathbf{P}$), we need to satisfy:

1. Global balanced equations: $\pi(j) = \sum_{k\in S}\pi(k)P_{kj}$
2. Local balanced equations: $\pi_i P_{ij} = \pi_j P_{ji}$

We design $\mathbf{P}$ such that local balanced equations are satisfied, then $\mathbf{P}$ will also satisfy global balanced equations.

8.3 Design the transition matrix

We begin with a general irreducible transition probability matrix Q, and adjust Q to form a desired P. The rationale is as follows: Given X_n, we sample the new random variable $t \sim Q_{X_n}$. Now the local balanced equations are $\pi_i Q_{ij} = \pi_j Q_{ji}$, which does not hold initially in general. We adjust by adding a term $\alpha(i, j) \in (0, 1]$, so that $\pi_i Q_{ij}\alpha(i, j) = \pi_j Q_{ji}\alpha(j, i)$.
We could define $\alpha(i, j) = \min\{\frac{\pi_j Q_{ji}}{\pi_i Q_{ij}}, 1\}$, then $P_{ij} = Q_{ij}\alpha(i, j)$ will be as desired.
In the actual execution, since the state space may be infinite, it may not be wise to compute full matrix $\mathbf{P}$, but to compute each P_{ij} whenever as required is good enough. We could interpret $\alpha(i, j)$ as a thinning factor for a process to jump from state i to j. The simulation of jumping with process $\mathbf{P}$ is implemented by rejection sampling on jumping with process $\mathbf{Q}$: A jump from i to j only occurs if the event of jump happens on $\mathbf{Q}$, and this jump is accepted.(controlled by α). If a jump occurs but failed, the process will be interpreted as jumping to the same state. This gives the actual implementation of Hastings-Metropolis algorithm.

8.4 Hastings-Metropolis algorithm

Note that if kernel function b exists, such that $\pi = cb$, we do not need to know π or c. Suppose we iteration N times.
The algorithm is implemented as follows:

1. Initialization: $n = 0, X(0) = k$
2. While $n < N$, generate $Y \sim Q_{X_n}$. Generate $U \sim Unif(0, 1)$
3. If $U < \alpha(X_n, Y) = \min\{\frac{b_Y Q_{Y,X_n}}{b_{X_n} Q_{X_n,Y}}, 1\}$, then $X(n + 1) = Y$, otherwise $X(n + 1) = X(n)$
4. Increase n by 1.

Note that to guarantee converge, we need to iterate for a large number of times, and truncate the results from earlier iterations. Analogous version for continuous distribution holds as well.

9 Poisson process

9.1 Introduction

We study a special case of continuous-time, discrete-state process: Poisson process. Examples include number of passengers arrived, or number of vehicles passing through a certain point by time t.

9.2 Poisson distribution

Poisson distribution is a good approximation for binomial distribution, since we could see the continuous time scale as breaking the time line into many small disjoint intervals, so small that no two events occur in one interval, then each interval is a Bernoulli variable, summing up to give binomial variable.

9.3 Properties of Poisson distribution

Suppose $X \sim Pois(\lambda), Y \sim Pois(\mu)$.
Summation of independent Possion variables: If X and Y are independent, then $X + Y \sim Pois(\lambda + \mu)$. If $Z|X \sim Binomial(X, r)$, then $Z \sim Pois(\lambda r)$.

9.4 Definition of Poisson process

A Poisson process with rate/intensity $\lambda > 0$, is an integer-valued stochastic process for which

1. for any time points $t_0 = 0 < t_1 < t_2 < \cdots < t_n$, the process increments $X_{t_i} - X_{t_{i-1}}$ are independent.
2. for $s \geq 0, t > 0$, $X(s + t) - X(s) \sim Pois(\lambda t)$
3. $X(0) = 0$

9.5 Law of rare events

Let ϵ_i be independent Bernoulli variables, each with parameter p_i. Let $S = \sum \epsilon_i$. Let $X \sim Pois(\sum p_i)$, then

$$|P(S = k) - P(X = k)| \ \leq \sum p_i^2$$

It shows that if a distribution satisfied the rarity condition(i.e. able to decompose as sum of many Bernoulli variables), then Possion distribution will be a good approximation.

9.6 Waiting time

Let $X(t)$ be a Poisson process with rate λ. Let W_n be the time of occurrence of the n-th event, it is called the waiting time of n-th event. We set $W_0 = 0$. The time between two occurrences $S_n = W_{n+1} - W_n$ are called sojourn times, which measures the duration that Poisson process sojourns in state n.
The waiting time W_n follows gamma distribution whole PDF is

$$f_{W_n}(t) = \frac{\lambda^n t^{n-1}}{(n-1)!}e^{-\lambda t}$$

In particular, W_1, the time to first event, follows exponential distribution with PDF

$$f_{W_1}(t) = \lambda e^{-\lambda t}$$

Since exponential variables are memoryless, we can interpret sojourn time S_n as if we are starting to observe the process when $X(t) = n$ until the first event occurs and $X(t) = n + 1$. Therefore, each of them follows independent and identical exponential distribution with parameter λ. This definition is useful for simulating Poisson process.