Stochastic Processes

Sabyasachi Chatterjee

November 16, 2022

Contents

1 Introduction **4**

1.1 Deterministic vs Stochastic Model 4

1.2 Some Applications of Stochastic Models 4

1.3 Random Walk 5

1.4 Gambler's Ruin 6

1.5 Review of Basic Probability Facts 6

1.6 Lets solve Gambler's Ruin 9

2 Discrete Time Markov Chains **9**

2.1 Frog and Bog 9

2.2 Definition 10

2.3 Some Examples of Markov Chains 11

2.4 Matrix Computations 12

3 Long Run Behaviour of Finite Markov Chains **14**

3.1 Properties of Limiting Distribution 15

3.2 Non-existence of Limiting Distributions 17

3.3 Reducibility 19

3.4 Periodicity 20

3.5 Fundamental Theorem for Irreducible, Aperiodic MC 21

3.6 Long run behavior for reducible and/or periodic chains 21

3.7 Fundamental Theorem for Irreducible MC 22

4 Return Times and Absorption Probabilities **24**

4.1 Expected Number of Visits to a Transient State 24

4.2 Expected Time till Absorption to a Recurrent Class 25
4.3 Expected First Return Time 26
4.4 Probability of Eventually Entering a Given Recurrent Class 27

5 More Examples of Finite Markov Chains 29

6 Countably Infinite Markov Chains 31

6.1 Recurrence and Transience 32
6.2 Recurrence/Transience of Simple Random Walk on Lattice 34
6.3 Null and Positive Recurrence 35
6.3.1 Stationary Distribution and Limiting Distribution 35
6.4 Recap of Differences between Finite and (Countably) Infinite Markov Chains 37

7 Branching Process 38

7.1 Extinction Probability in a Branching Process 39

8 Time Reversible Markov Chains 43

8.1 Definition and Characterization 43
8.2 Discussion about Local Balance 44
8.3 Checking Local Balance suffices to show Stationarity 46

9 Markov Chain Monte Carlo 47

9.1 Law of Large Numbers 48
9.2 Binary Sequences with No Consecutive 1's 49
9.3 Metropolis Hastings Algorithm 51
9.3.1 Examples/Applications of Metropolis Hastings Algorithm 54
9.3.2 Continuous State Space 56
9.4 Gibbs Sampling 56
9.5 Examples/Applications of Gibbs Sampling 59

10 A Linear Algebraic Condition for Convergence 62

11 Poisson Process 63

11.1 Bernoulli Counting Process 64
11.2 Definition of Poisson Process 66
11.3 Inter-Arrival Times 67

11.4 Memorylessness of the Exponential Random Variable 68

11.5 Conditioning on the number of arrivals in a Poisson Process 69

11.6 Superposition 71

11.7 Thinning 73

11.8 Birthday Problem and Poisson Process Embedding 75

11.9 Spatial Poisson Process 76

11.10Non Homogenous Poisson Process 77

11.11A Paradox 78

12 Brownian Motion **80**

12.1 Some History 80

12.2 Introduction 80

12.3 What is Brownian Motion Really? 83

12.4 Brownian Motion as a limit of Random Walk 83

12.5 Gaussian Process 85

12.6 Transformations and Properties 86

12.7 Nowhere Differentiability 87

1 Introduction

1.1 Deterministic vs Stochastic Model

Suppose we are tracking the number of bacteria in a container. Let $y(0)$ be the initial number and $y(t)$ be the number of bacteria at time t. Suppose the bacteria population grows at a 20 percent rate. One way to express this mathematically is to say that the growth function $y(t)$ satisfies the differential equation

$$y'(t) = 0.2y(t).$$

The solution to the above differential equation of course is

$$y(t) = y(0)\exp(0.2t).$$

This is an example of a deterministic equation. If we know $y(0)$ then we exactly know $y(t)$ for any $t > 0$. However, in this case it might be reasonable to take into account the randomness of growth of bacteria. So here is a possible stochastic growth model.

At any point of time, if there are n bacteria, the waiting time for the next bacteria to arrive is a random variable which follows the Exponential distribution with mean $\frac{1}{0.2n}$ or rate $0.2n$. Now, for any time t, $y(t)$ is a random variable. This is an example of a stochastic model. Natural questions of interest here are a) What is the distribution of $y(t)$? b) What is the distribution of the doubling time? This process is called a birth process and we may come back to this process later in the course.

1.2 Some Applications of Stochastic Models

- **Pagerank Algorithm**: Given a search query, Google ranks webpages related to the query and shows it to us. A high level idea of how they do this is s follows. Suppose the query is ”chess”. First consider the set of all the webpages S which contains the word ”chess”. Now the key idea is to view this set as a directed graph. Each website represents a node in this graph. If webpage A links to webpage B then there is a directed edge going from A to B in the graph. Now consider the random websurfer model. Start from any webpage A, choose with equal probability from the set of webpages A links to and go there. Repeat this process. This is clearly a stochastic/random process. Let $p_x^{(k)}$ be the probability of being at webpage x after k hops. The long run probability of a webpage can be defined as

  $$\pi(x) = \lim_{k\to\infty} p_x^{(k)}.$$

  The webpages are now ranked based on the π values; higher the π value for a webpage the higher it is ranked. One can think and convince himself that $\pi(x)$ is high if a lot of webpages link to webpage x or even if a few very highly rated webpages link to webpage x. So you cannot create a billion junk websites and link all of them to a master junk webiste and increase its rank in this way.

- **SIR Model**: The SIR model is a stochastic model for the spread of infectious diseases. Let the entire population be divided into three groups: susceptible (S), infected (I) and recovered (R). At time t, we have three numbers for the three groups S_t, I_t, R_t. Let the unit of time be such that it matches the recovery period of the disease. That is if a person is infected at time t, then at time $t+$ the person recovers. Assume that at any time t, each susceptible perosn is in contact with all the infected people and any contact with an infected person independently results in a trasmission of the disease with probability z. Then the probability of a susceptible person at time t to still remain susceptible at time $t+1$ and not get infected is $(1-z)^{I_t}$. Therefore, we can write
$$I_{t+1} \sim Binomial(S_t, 1-(1-z)^{I_t}).$$

Moreover, $S_{t+1} = S_t - I_{t+1}$ and $R_{t+1} = R_t + I_t$. The three tuple of numbers $\{(S_t, I_t, R_t) : t \geq 0\}$ forms a stochastic process. Run this stochastic process on a computer with various initial values of S_0, I_0 and z and see what happens to I_t as a function of time. One question of interest is whether everybody becomes infected eventually or not.

Definition 1.1. *A stochastic process is a collection of random variables $\{X_t : t \in T\}$. The index set T may be finite or infinite. Technically, the stochastic process refers to the joint distribution of the collection of random variables $\{X_t : t \in T\}$.*

Definition 1.2. *The set of values that the random variables $\{X_t : t \in T\}$ is called the state space, usually denoted by $\mathcal{S}$.*

Whenever you encounter a stochastic process, you should ask yourself these questions. What is the state space? Is the state space finite, countably infinite or uncountable infinite? What is the index set T? Is the index set T finite, countably infinite or uncountable infinite?

1.3 Random Walk

Here is a question. A mosquito starts from a point called the origin. What would be the distance travelled by the mosquito after 1 hour? Such a question came up when the scientist Ronald Ross was studying how malaria spreads. If we assume that the mosquito is dumb (which may not be true) then we can make the following model for its movements. Every second, the mosquito goes in a certain direction for 1 millimetre. Now there are infinitely many directions the mosquito can go so to simplify matters, let assume that there are only four directions (North, South, East, West) available for the mosquito to go to. So the mosquito chooses a direction with probability 1/4 every second and repeats this process. This random motion of the mosquito is called the *Random Walk* in two dimensions.

Let's come down to one dimension. Let $X_1, X_2, \ldots,$ be a sequence of i.i.d Rademacher random variables which takes the value 1 with probability $1/2$ and -1 with probability $1/2$. Define the random variable for any $n \geq 0$;

$$S_n = X_1 + \cdots + X_n$$

with the convention that $S_0 = 0$. The stochastic process $\{S_n : n \geq 0\}$ is called a *simple random walk*.

A natural question is what is the distribution of S_t for some large time t? Central Limit Theorem says that we can approximate this distribution by $N(0, t)$. Therefore, we can say the following: With probability atleast 0.99, S_t lies between $\pm 3\sqrt{t}$.

The simple random walk satisfies the following two properties:

- Independent Increments: Take any time points $t1 < t_2 < t_3 < t_4$. The random variables $S_{t_2} - S_{t_1}$ and $S_{t_4} - S_{t_3}$ are independent.
- Stationary Increments: Take any time points $t1 < t_2$. The distribution of $S_{t_2} - S_{t_1}$ is the same as the distribution of $S_{t_2 - t_1}$.

Stochastic Processes satisfying the above two properties are *fundamental* in some sense. We will see later that two other fundamental stochastic processes, Poisson Process and Brownian Motion, satisfy these two properties.

1.4 Gambler's Ruin

Here is a famous problem associated with the random walk process. Suppose a gambler at each round either wins a dollar or loses a dollar with probability $1/2$ each. Suppose the gambler starts at k dollars. He stops when either he reaches his goal of N dollars or he goes bankrupt and loses all his money. The winnings of the gambler can be thought of as a random walk starting from $S_0 = k > 0$ and stopping when either the random walk hits 0 or N for the first time? We will see that if we run this process; there are only two possibilities. Either the random walk hits N or hits 0, it cannot vacillate forever between 0 and N. The question of interest is: what is the probability that the gambler will be ruined; i.e he will lose all his money?

1.5 Review of Basic Probability Facts

- $$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

  The events A, B are independent iff $P(A|B) = P(A)$.

- Suppose the events $B_1, \dots, B_k$ partition the sample space. Then we have

  $$\sum_{i=1}^{k} P(B_i)P(A|B_i) = P(A).$$

  This is called the law of total probability (LOTP).

- $$P(B|A) = \frac{P(A|B)P(B)}{P(A|B)P(B) + P(A|B^c)P(B^c)}$$

  . This is called the *Bayes Rule*.

- *Conditional Distribution of* Y *given* $X = x$.

  1. Both X, Y are discrete.

     $$P(Y = y|X = x) = \frac{P(Y = y, X = x)}{P(X = x)}.$$

  2. Both X, Y are continuous.

     $$f(y|X = x) = \frac{f(x, y)}{\int_{-\infty}^{\infty} f(x, y)dy}.$$

  3. X continuous, Y discrete. Think of the joint distribution in such a way that for every value of $Y = y$ there is a conditional pdf $f(x|Y = y)$.

     $$P(Y = y|X = x) = \frac{f(x|Y = y)P(Y = y)}{\sum_y f(x|Y = y)P(Y = y)}$$

  4. Y continuous, X discrete. For each value of $X = x$, there is a conditional pdf $f(y|X = x)$.

- *Conditional Expectation*

  1. Y is discrete.

     $$\mathbb{E}(Y|X = x) = \sum_y yP(Y = y|X = x).$$

  2. Y is continuous.

     $$\mathbb{E}(Y|X = x) = \int_{-\infty}^{\infty} yf(y|X = x).$$

- *Properties of Conditional Expectation*

  Conditional Expectation is also expectation of some distribution and hence retains all the nice properties of expectation.

  1. Linearity:

     $$\mathbb{E}(aY + bZ|X = x) = a\mathbb{E}(Y|X = x) + b\mathbb{E}(Z|X = x).$$

2. If g is a function then

   (a) Y is discrete.

   $$\mathbb{E}(g(Y)|X=x) = \sum_y g(y)P(Y=y|X=x).$$

   (b) Y is continuous.

   $$\mathbb{E}(g(Y)|X=x) = \int_{-\infty}^{\infty} g(y)f(y|X=x).$$

   This is often called the *law of the unconscious statistician (LOTUS)* which is being applied here to conditional expectation.

3. Independence: If $\mathbb{E}(Y|X=x) = EY$ for all $X=x$ then X is independent of Y.

4. If $Y = g(X)$ then $\mathbb{E}(Y|X=x) = g(x)$.

- Conditional Expectation given an event A:

  $$\mathbb{E}(Y|A) = \frac{\mathbb{E}(Y1_A)}{P(A)}.$$

  The value $\mathbb{E}(Y|A)$ is same as $\mathbb{E}(Y|1_A = 1)$.

- Law of Total Expectation: Suppose the events $B_1, \ldots, B_k$ partition the sample space. Then we have

  $$\sum_{i=1}^{k} P(B_i)\mathbb{E}(Y|B_i) = \mathbb{E}Y.$$

- Conditional Expectation as a Random Variable:

  1. $\mathbb{E}(Y|X)$ is a random variable.
  2. $\mathbb{E}(Y|X)$ is a random variable which is a function of X.
  3. $\mathbb{E}(Y|X)$ is a random variable which takes the value $\mathbb{E}(Y|X=x)$ on those points in the sample space where $X=x$.
  4. Law of Iterated Expectation

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

- Conditional Variance and its properties:

  1. Conditional variance is defined as follows:

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

  2. $$Var(aY+b|X=x) = a^2 Var(Y|X=x).$$

3. $Var(Y|X)$ is a random variable.
4. Law of Total Variance:

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

   Prove this!

1.6 Lets solve Gambler's Ruin

Let A be the event that the gambler is ruined. Let $p_k = P(A|X_0 = k)$. The key idea is to condition on the first step. We will use this idea repeatedly in this course. By using LOTP for conditional probability we obtain for $k = 1, \dots, N-1$,

$P(A|X_0 = k) = P(A|X_0 = k, X_1 = k+1)P(X_1 = k+1|X_0 = k) + P(A|X_0 = k, X_1 = k-1)P(X_1 = k-1|X_0$
$\frac{p_{k+1}}{2} + \frac{p_{k-1}}{2}$.

Therefore, we get the recurrence relation for $k = 1, \dots, N-1$,

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

Note that $p_0 = 1$ and $p_N = 0$. We can now solve the recurrence relation to obtain that $p_k = \frac{N-k}{N}$.

2 Discrete Time Markov Chains

2.1 Frog and Bog

Markov Chain is perhaps the simplest dependent sequence of random variables you can think of. Let's start with an example. Suppose there are two bogs, Bog 0 and Bog 1. A frog, starting from Bog 0 either stays at Bog 0 if there are enough insects to eat or jumps to Bog 1 in the hope of more food. Once the frog is in Bog 1 it stays there if it gets enough food or otherwise it jumps back to Bog 0. To model the motion of this frog probabilisitically, let's say that when the frog is in Bog 0 it stays with probability 0.9 and jumps with probability 0.1. Similarly, when the frog is in Bog 1 it stays with probability 0.8 and jumps with probability 0.2. Now we can define $\{0, 1\}$ valued random variables $X_0 = 0, X_1, \dots$ where X_i denotes the position of the frog after i hops. This stochastic process $\{X_i\}_{i=0}^{\infty}$ is an example of a Markov Chain (MC).

The key property defining a MC is that the value of the random variable X_{i+1} only depends on X_i and not on the entire past history $X_0, \dots, X_{i-1}, X_i$. We will make this precise in a bit. Note that the MC forms a *dependent* sequence of random variables. The next state highly depends on the current state and is more likely to remain the same as the current state. A natural question of interest might be what is the proportion of time the frog spends in Bog 1? It turns out that this proportion converges to 1/3. We will simulate and observe this in class. This is not obvious as to why this happens. An intuitive explanation is as follows. Suppose the frog is at Bog 0. It has a 1 in 10 chance of going to Bog 1. So on an average the frog will take 10 hops to go to Bog 1. From Bog 1, the chance of jumping back to Bog 0 is 1 in 5. On an average the frog will take 5 hops to return back. So it is plausible that for every 10 times the frog will be in Bog 0 it will be in Bog 1 5 times. Hence the ratio 1/3.

Andrey Markov, a Russian mathematician, in 1907 came up with the concept of Markov Chains because he wanted to demonstrate a sequence of dependent random variables which exhibit a law of large numbers. The usual law of large numbers require independence. For the above stochastic process of the frog and the bogs, we indeed see the phenomenon of law of large numbers because the fraction of times the frog spends in Bog 1 converges to 1/3.

2.2 Definition

For now, we will focus on discrete state space $\mathcal{S}$. When we say discrete we mean either $\mathcal{S}$ is finite or it is countably infinite like the set of all integers $\mathcal{Z}$.

Definition 2.1. *A Markov Chain is a stochastic process $X_0, X_1, X_2, \ldots$ where the following is true*

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

for all $n \geq 0, x_o, x_1, \ldots, x_{n-1}, i, j \in \mathcal{S}$.

This definition says that the conditional distribution of X_{n+1} given the entire past $X_0, \ldots, X_n$ only depends on X_n and not on how the MC evolved to X_n.

Definition 2.2. *A MC is called time homogenous if*

$$P(X_{n+1} = i | X_n = j) = P(X_1 = i | X_0 = j).$$

We will be focussing only on time homogenous MC in this course. This does not mean non time homogenous Markov Chains are not important. Infact, they can arise in several contexts but that is left for future courses.

The transition probabilities for a time homogenous MC can be naturally written down as a matrix which we will denote by P satisfying $P_{ij} = P(X_1 = i | X_0 = j)$. This matrix P satisfies two properties:

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

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

Any matrix which satisfies the above two properties is called a stochastic matrix.

2.3 Some Examples of Markov Chains

Example 1: IID Sequence The simplest MC is a pure i.i.d sequence $X_1, X_2, \ldots$. The transition probabilities satisfy $P_{ij} = P(X_1 = j)$. Often, iid sequences are used to model random samples in statistics. However, many random phenomena are not independent and one needs to model the dependence. This is precisely where Markov Chains can be useful.

Example 2: One Dimensional Random Walk Here the state space is $\mathbb{Z}$. The transition matrix P is infinite. For each integer i, we have

$$P_{ij} = \begin{cases} p & \text{if } j = i+1 \\ 1-p & \text{if } j = i-1 \\ 0 & \text{if } j \neq i \pm 1 \end{cases}$$

where $0 < p < 1$. When $p = 0.5$, the random walk is often called a simple random walk.

Example 2: Gambler's Ruin Here the state space is $\mathbb{Z} \cap [0, N] = [0 : N]$. The transition matrix $P_{N+1 \times N+1}$ is as follows for any $i \in [1 : N-1]$.

$$P_{ij} = \begin{cases} p & \text{if } j = i+1 \\ 1-p & \text{if } j = i-1 \\ 0 & \text{if } j \neq i \pm 1 \end{cases}$$

where $0 < p < 1$. We also have $P[0, 0] = 1$ and $P[N, N] = 1$. So if the MC goes to state 0 or N it stays there forever. We call such states absorbing states.

Birth and Death Chain The state space is $\mathbb{Z}$. The transition matrix P is infinite. For each integer i, we have

$$P_{ij} = \begin{cases} p_i & \text{if } j = i+1 \\ q_i & \text{if } j = i-1 \\ 1-p_i-q_i & \text{if } j = i \end{cases}$$

where $0 < p < 1$. This MC is used to model population size, number of customers in a queue etc.

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

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

Random Walk on a Directed Weighted Graph Consider a weighted directed graph (V, W). Every ordered pair of vertices (i, j) has a weight $W_{ij} \geq 0$ associated to it. This is the weight of the edge going from i to j. The weight w_{ij} can be different from

w_{ji}. We can think of the graph as a map where each node represents a city and each edge represents a one way road. The weight W_{ij} can be thought of as the inverse of the time it takes to traverse the road and go from city i to city j. So higher the weight, better the road is and lesser the time it takes. Note that these are one way roads so it might take different time to go from city i to city j than come back from city j to city i. If the weight $W_{ij} = 0$ then this means that there is no road yet built which allows you to go from city i to city j.

The state space is again the set of vertices V. The transition matrix $P_{|V| \times |V|}$ is as follows.

$$P_{ij} = \frac{1}{\sum_{k \in V} W_{ik}} W_{ij}.$$

Card Shuffling When we shuffle cards we are effectively running a MC. Here the state space is S_{52} which is the set of permutations of the set $[1:52]$. For different shuffling mechanisms, the transition matrix may be different. Let's consider the simplest (and a inefficient) shuffling scheme. Pick the top card and put it back at a random place. What is the transition matrix? Write it down when we have a deck of 3 cards instead of 52. What would be a realistic shuffling scheme? A question of interest is what is the distribution of the ordering of the cards after we have shuffled 20 times? 50 times? 100 times?

2.4 Matrix Computations

Given any possible trajectory $X_0 = x_0, X_1 = x_1, \dots, X_n = x_n$ we can write calculate its probability in terms of the initial distribution of X_0 and the transition probability matrix P.

Lemma 2.3 (Distribution of Entire Trajectory)**.** *Given a time homogenous MC $X_0, X_1, \dots$ with initial distribution $X_0 \sim \alpha$ and transition matrix P, we have for all $x_0, x_1 \dots, x_n \in \mathcal{S}$,*

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

Proof.

$$P(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) =$$
$$P(X_0 = x_0) P(X_1 = x_1 | X_0 = x_0) P(X_2 = x_2 | X_1 = x_1, X_0 = x_0) \dots P(X_n = x_n | X_{n-1} = x_{n-1} \dots X_0 = x_0) =$$
$$P(X_0 = x_0) P(X_1 = x_1 | X_0 = x_0) P(X_2 = x_2 | X_1 = x_1) \dots P(X_n = x_n | X_{n-1} = x_{n-1})$$

where the second inequality follows from the Markov property. □

Lemma 2.4 (A Consequence of Markov Property)**.** *Given a time homogenous MC $X_0, X_1, \dots,$ for any subsequence of times $t_1 < t_2 < \dots < t_n$ and for all $x_{t_0}, x_{t_1} \dots, x_{t_n} \in \mathcal{S}$,*

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

Proof. For the sake of writing the proof let $n = 4$ and t_1, t_2, t_3, t_4 be $1, 3, 5, 8$ respectively. We can write

$$P(X_8 = x_8 | X_5 = x_5, X_3 = x_3, X_1 = x_1) = \frac{P(X_8 = x_8, X_5 = x_5, X_3 = x_3, X_1 = x_1)}{P(X_5 = x_5, X_3 = x_3, X_1 = x_1)}.$$

Now let's expand the numerator.

$$P(X_8 = x_8, X_5 = x_5, X_3 = x_3, X_1 = x_1) = \sum_{x_7, x_6, x_4, x_2, x_0} P(X_8 = x_8, X_7 = x_7, \dots, X_0 = x_0) =$$

$$\sum_{x_7, x_6, x_4, x_2, x_0} P_{x_7, x_8} P_{x_6, x_7} \dots P_{x_0, x_1} P(X_0 = x_0) = [\sum_{x_7, x_6} P_{x_7, x_8} P_{x_6, x_7} P_{x_5, x_6}][\sum_{x_4, x_2, x_0} P_{x_4, x_5} P_{x_3, x_4} \dots P_{x_0, x_1} P(X_0 =$$

$$P(X_8 = x_8 | X_5 = x_5) P(X_5 = x_5, X_3 = x_3, X_1 = x_1).$$

Clearly, this argument would work for any general $t_1 < t_2 < \dots < t_n$ as well. □

The transition matrix P gives the one step transition probabilities $P(X_1 = j | X_0 = i)$. The n step transition probabilities are also obtainable from the matrix P^n where P^n is obtained by multiplying the matrix P with itself n times.

Lemma 2.5. *Given a time homogenous MC $X_0, X_1, \dots,$ we have for any $n \geq 0$,*

$$P(X_n = j | X_0 = i) = P^n_{ij}.$$

Proof. What are the two step transition probabilities $P(X_2 = j | X_0 = i)$?

$$P(X_2 = j | X_0 = i) = \sum_{k \in \mathcal{S}} P(X_2 = j | X_0 = i, X_1 = k) P(X_1 = k | X_0 = i) =$$

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

This means that the matrix P^2 gives the two step transition probabilities. Now by mathematical induction, we can show that the matrix P^n gives the n step transition probabilities. Assume that this is true till integer $n - 1$.

$$P(X_n = j | X_0 = i) = \sum_{k \in \mathcal{S}} P(X_n = j | X_0 = i, X_1 = k) P(X_1 = k | X_0 = i) =$$

$$\sum_{k \in \mathcal{S}} P(X_n = j | X_1 = k) P(X_1 = k | X_0 = i) = \sum_{k \in \mathcal{S}} P^{n-1}_{kj} P_{ik} = (P^n)_{ij}.$$

where in the second equality we have used Lemma 2.4. □

Chapman Kolmogorov Equations: In particular, the matrix identity $P^{m+n} = P^m P^n$ implies that for any i, j we have

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

Lemma 2.6 (Distribution of Partial Trajectory)**.** *Given a time homogenous MC $X_0, X_1, \dots$ with initial distribution $X_0 \sim \alpha$ and transition matrix P, we have for all $j \in \mathcal{S}$,*

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

Infact, for any subsequence of times $t_1 < t_2 < \cdots < t_n$, we have

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

Proof.

$$P(X_n = j) = \sum_{k \in \mathcal{S}} P(X_0 = k) P(X_n = j | X_0 = k) = \sum_{k \in \mathcal{S}} \alpha_k P^n_{kj} = (\alpha P^n)_j.$$

For the second part, write

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

where in the second equality we have again used Markov property as in Lemma 2.4. □

Takeway Message: In principle, probability of any event concerning the MC can be calculated by matrix computations.

3 Long Run Behaviour of Finite Markov Chains

A central question in the topic of Markov Chains is that what happens to the chain in the long run? Specifically, what is the distribution of X_n? We can raise this question when we shuffle cards, for gambler's ruin, for a random walk on a weighted directed graph and so on. Clearly, we need to examine what happens to the n step transition probabilities P^n_{ij}. For the next little while, we are going to focus on markov chains with finite state space.

To compute P^n one can always use the computer. Let's come back to our frog and bog example. The transition matrix in this case is

$$P = \begin{pmatrix} 0.9 & 0.1 \\ 0.2 & 0.8 \end{pmatrix}$$

It can be checked in a computer that for large n, we have

$$P^n \to \begin{pmatrix} 2/3 & 1/3 \\ 2/3 & 1/3 \end{pmatrix}$$

So, when n is large, we can say that the probability of the frog being in Bog 1 is close to 1/3 if n is large irrespective of where the frog started from. Since P^n converges to a stochastic matrix with the same rows, for any initial distribution α, we have the distribution of X_n given by αP^n converging to $(2/3, 1/3)$.

Definition 3.1. *A MC is said to have a limiting distribution λ if for all $i, j \in \mathcal{S}$ we have*

$$\lim_{n\to\infty} P^n_{ij} = \lambda_j.$$

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

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

Note that by definition, limiting distribution of a MC is unique.

Example: General Two State MC Consider a general two state MC with the following transition matrix

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

If $p+q = 1$ then the rows of P are the same and $P^n = P$. Hence the limiting distribution $\lambda = (1-p, p)$. So let's assume that $p+q \neq 1$. Let's compute P^n. We can write

$$\begin{aligned} &P^n_{11} = (P^{n-1}P)_{11} = P^{n-1}_{11}P_{11} + P^{n-1}_{12}P_{21} = \\ &P^{n-1}_{11}(1-p) + (1-P^{n-1}_{11})q = q + (1-p-q)P^{n-1}_{11} \end{aligned}$$

The above is a recurrence relation of the form $a_n = b + ca_{n-1}$. The solution to the above recurrence relation is $a_n = b\frac{1-c^{n-1}}{1-c} + c^{n-1}a_1$. Plugging in the appropriate values, we get

$$P^n_{11} = \frac{q}{p+q} + \frac{p}{p+q}(1-p-q)^n.$$

Similarly, other elements of P^n can be found to see that

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

3.1 Properties of Limiting Distribution

Lemma 3.2. *If λ is the limiting distribution for a MC with transition matrix P then λ satisfies the equation*

$$\lambda P = \lambda.$$

Proof.

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

□

Definition 3.3. *A distribution π which satisfies the equation*

$$\pi P = \pi$$

is called a stationary distribution for the MC.

If π is a stationary distribution and $X_0 \sim \pi$ then the distribution of X_1 is π and in fact for any $n \geq 1$, the distribution of X_n is also π. This is because $\pi P^n = \pi$. So if the MC has initial distribution π its distribution at any time n will remain π. This is why π is called a stationary distribution. Lemma 3.2 says that a limiting distribution λ for the MC has to also be a stationary distribution. The converse is not always true. For example, consider the MC with the transition matrix

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$

This is a MC chain which always flips states deterministically. There is no limiting distribution for this chain since the distribution of X_n will keep on alternating according to whether n is even or odd. However, $\pi = (1/2, 1/2)$ is a stationary distribution for this chain.

Also consider the MC with the transition matrix

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

This is a MC chain which just stays put in its initial state. Since $P = P^2 = P^n$ there is no limiting distribution but any distribution π is a stationary distribution for this chain.

We will see that for a large and generic class of Markov Chains there will be unique stationary distribution π which will also be the limiting distribution for the MC. This will give us a generic way to calculate the limiting distribution. All we have to do is to solve the linear system of equations $\pi P = \pi$.

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

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

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

Lemma 3.4. *If λ is the limiting distribution for a MC with transition matrix P then λ satisfies the equation $\lim_{n\to\infty} \mathbb{E}(F_{n,j} | X_0 = i) = \lambda_j$ for all $j, i \in \mathcal{S}$.*

Proof. We can write

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

Therefore, taking limits we can conclude that

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

In the above display, the second equality is true because of Cesaro's lemma from real analysis. Cesaro's lemma says that if a sequence of real numbers $a_1, a_2, \ldots$ converges to a number then the sequence of partial means of the same sequence $b_n = \frac{a_1+\cdots+a_n}{n}$ will also converge to the same number. □

3.2 Non-existence of Limiting Distributions

Limiting distributions may not exist for a given MC. Let's consider some examples of MC which illustrate the various settings under which a limiting distribution would not exist. Recall that we are only considering finite MC for now.

Example 1: Simple Random Walk with Reflecting Boundary

Consider the state space $\{0, 1, 2, 3, 4\}$ and the transition matrix

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

For large n, one can check that P^n looks different depending on whether n is odd or even. If n is even, then P^n is nearly

$$\begin{pmatrix} 1/4 & 0 & 1/2 & 0 & 1/4 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 1/4 & 0 & 1/2 & 0 & 1/4 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 1/4 & 0 & 1/2 & 0 & 1/4 \end{pmatrix}$$

If n is odd, then P^n is nearly

$$\begin{pmatrix} 0 & 1/2 & 0 & 1/2 & 0 \\ 1/4 & 0 & 1/2 & 0 & 1/4 \\ 0 & 1/2 & 0 & 1/2 & 0 \\ 1/4 & 0 & 1/2 & 0 & 1/4 \\ 0 & 1/2 & 0 & 1/2 & 0 \end{pmatrix}$$

There is a periodic nature to this MC. After even number of steps, the chain can only be of the same parity as the initial state. We will later see that the period of this chain is 2. There is no limiting distribution of this chain precisely because of this periodicity.

Question: Why does this chain not have a limiting distribution? Answer: Periodicity

Example 2: Simple Random Walk with Absorbing Boundary

Again consider the state space $\{0, 1, 2, 3, 4\}$ and the transition matrix

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

If n is large, we see that P^n is nearly

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 \\ 3/4 & 0 & 0 & 0 & 1/4 \\ 1/2 & 0 & 0 & 0 & 1/2 \\ 1/4 & 0 & 0 & 0 & 3/4 \\ 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

In this case, the random walker eventually settles at either 0 or 4. For example, starting from 1 with probability $3/4$ it eventually settles at 0 and with probability $1/4$ it eventually settles at 4. This means that with probability 1 the random walker will eventually leave the state 1 and never come back. The same conclusion holds for states $2, 3$. We will call such states as *transient.*

Question: Why does this chain not have a limiting distribution? Answer: Existence of Transient States

Example 3: Suppose $S = \{1, 2, 3, 4, 5\}$ and let the transition matrix be

$$P = \begin{pmatrix} 1/2 & 1/2 & 0 & 0 & 0 \\ 1/6 & 5/6 & 0 & 0 & 0 \\ 0 & 0 & 3/4 & 1/4 & 0 \\ 0 & 0 & 1/8 & 2/3 & 5/24 \\ 0 & 0 & 0 & 1/6 & 5/6 \end{pmatrix}$$

For large n, we see that P^n is nearly

$$\begin{pmatrix} 1/4 & 3/4 & 0 & 0 & 0 \\ 1/4 & 3/4 & 0 & 0 & 0 \\ 0 & 0 & .182 & .364 & .455 \\ 0 & 0 & .182 & .364 & .455 \\ 0 & 0 & .182 & .364 & .455 \end{pmatrix}$$

The chain splits into two non interacting or separate MC. One with state space $\{1, 2\}$ and the other with state space $\{3, 4, 5\}$. We call such a MC *reducible.*

Question: Why does this chain not have a limiting distribution? Answer: Reducibility

We will see that these are the only three ways a MC can fail to have a limiting distribution.

3.3 Reducibility

Definition 3.5. *States i and j communicate if there exists integers $m, n \geq 0$ such that $P^m_{i,j} > 0$ and $P^n_{i,j} > 0$.*

Therefore, we say that i and j communicate with each other if it is possible for the MC to go from i to j and also come back from j to i in a finite number of steps. This communication relation between pairs of states define an equivalence relation on the state space $\mathcal{S}$. This is because the following three properties hold.

1. i communicates with i.

2. If i communicates with j then j communicates with i.

3. If i communicates with j and j communicates with k then i communicates with k.

The first one is true because $P^0_{i,i} = 1$ for any state i. The second one is true by the symmetric definition. The third one is true because $P^{m+n}(i,k) \geq P^m(i,j)P^n(j,k) > 0$.

This equivalence relation partitions the state space $\mathcal{S}$ into equivalence classes called **communication classes**.

Definition 3.6. *If the MC has only one communicating class then the MC is called irreducible.*

Sometimes, we abuse terminology and say that a transition matrix P is irreducible which means that the MC with transition matrix P is irreducible.

In example 1, we see that there is only one communicating class. It is possible to go to any state from any state. Therefore, the MC here is irreducible. In example 2, the communication classes are $\{2,3,4\}, \{1\}, \{5\}$. If the random walker starts in the class $\{2,3,4\}$ then w.p 1 he/she eventually leaves the class forever. Such classes are called transient classes. The states of such a class are called transient states.

Definition 3.7. *A communicating class is called transient if starting from that class, with probability 1 the MC leaves that class and never returns. The states of such a class are called transient states. A communication class which is not transient is called a recurrent class. The states of such a class are called recurrent states.*

An important fact about a recurrent communication class is that if a MC starts in this class then it never leaves this class. Think why.

In general, a finite MC might have several recurrent classes and several transient classes. It must have atleast one recurrent class. In particular, if the MC is irreducible then all states are recurrent states. By reordering the states if necessary, we can write its transition matrix in the following block matrix form:

$$P = \left[\begin{array}{cccc|c} P1 & 0 & 0 & 0 & \\ 0 & P_2 & 0 & 0 & \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & P_r & \\ \hline \multicolumn{4}{c|}{S} & Q \end{array}\right]$$

Here, $P_1, \ldots, P_r$ are the transition matrices of the recurrent classes, S represents the one step transition probabilities from a transient state to a recurrent state and Q represents the one step transition probabilities from a transient state to another transient state.

In this case, we can write the n step transition matrix as

$$P^n = \left[\begin{array}{cccc|c} P_1^n & 0 & 0 & 0 & \\ 0 & P_2^n & 0 & 0 & \\ 0 & 0 & \ddots & 0 & 0 \\ 0 & 0 & 0 & P_r^n & \\ \hline \multicolumn{4}{c|}{S_n} & Q^n \end{array}\right]$$

Therefore, to analyze the long run behavior of the MC we need to understand what happens to $P_1^n, \ldots, P_r^n$ individually. This means we need to understand what happens to P^n for a given irreducible transition matrix P. We will later see what happens to S_n and Q^n as well.

3.4 Periodicity

Suppose P is the transition matrix for an irreducible MC. If it is reducible, we should consider each recurrent communication class separately. For a given state i, let us define the set J_i as follows

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

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

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

Here *gcd* stands for greatest common divisor. If we return back to example 1 of a RW with reflecting boundaries, we see that for state 1, the set J_i equals the set of all positive even integers. Hence, the period of state 1 is 2.

Lemma 3.8. *For an irreducible MC, all states have the same period.*

Proof. Let d be a common divisor of J_i. Consider any other state j. We will show that d also is a common divisor of J_j. This will mean that $gcd(J_i) = gcd(J_j)$ and hence the period of state i is the same as the period of state j. Since the MC is irreducible there exists integers $m, n > 0$ such that $P_{ij}^m > 0$ and $P_{ji}^n > 0$. This implies that $m + n \in J_i$ and hence d divides $m + n$. Now take any $l \in J_j$. Now we have $P^{m+n+l}(ii) \geq P_{ij}^m P_{jj}^l P_{ji}^n > 0$. This means that $m + n + l \in J_i$ and hence d divides $m + n + l$ as well. Since d divides both $m + n$ and $m + n + l$ it must divide l as well. Since l was an arbitrary element of J_j, d is a common divisor of J_j. □

Example: Consider a Random Walk on a undirected graph. The MC is irreducible iff the graph is connected. For a connected graph, every vertes has degree atleast 1. Hence it is possible for the MC to start from i and come back to i in an even number of steps. Hence, the period of the chain is either 1 or 2. The period is 2 iff the graph is bipartite, meaning that the set of vertices can be divided into two subsets and each edge in the graph goes from one subset to another. An example of a bipartite graph is a cycle graph of even length.

3.5 Fundamental Theorem for Irreducible, Aperiodic MC

Theorem 3.9 (Fundamental Theorem for Irreducible, Aperiodic Markov Chain)**.** *If P is the transition matrix for an irreducible, aperiodic (finite) Markov chain then there exists a unique stationary distribution or a unique solution to the equation $\pi = \pi P$ which satisfies the following two properties:*

1. *If α is any initial distribution then*
$$\lim_{n \to \infty} \alpha P^n = \pi.$$
In words, π is the limiting distribution of the Markov chain.

2. *$\pi(j) > 0$ for all $j \in \mathcal{S}$. In words, π gives positive probability to each of the states.*

Operational Implication of Fundamental Theorem: If a MC is irreducible, aperiodic then to find the limiting distribution it is enough to solve the linear system $\pi = \pi P$.

Remark 3.1. *An equivalent condition for P to be irreducible, aperiodic is that there exists $n \geq 1$ such that P^n has all entries positive.*

3.6 Long run behavior for reducible and/or periodic chains

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

Assume P is reducible with recurrent classes $R_1, \ldots, R_r$ and transient classes $T_1, \ldots, T_s$. Each recurrent class acts as a separate MC with transition matrix $P_1, \ldots, P_r$. Assume each P_k is aperiodic. Then by Theorem 3.9 there exists r different limiting distributions $\pi^1, \ldots, \pi^r$. The distribution π^k is supported on its own recurrent class; i.e $\pi^k(j) = 0$ if $j \notin R_k$. There are three cases to consider:

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

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

Now suppose P is irreducible but periodic with period $d > 1$. In this case, the state space partitions itself into d sets $A_1, A_2, \ldots, A_d$. The matrix P^n will keep on switching according to whether $n|d$ has remainder $0, 1, \ldots, d-1$. We can see this in Example 1 where the period is 2. Therefore, there cannot be a limit of P^n in this case. However, the expected long run proportions of time spent in each state still has a limit. For any state j, we can define the indicator random variable $I_k = 1(X_k = j)$. Now define
$$F_{n,j} = \frac{1}{n}\sum_{k=0}^{n-1} I_k$$

Then $\mathbb{E}(F_{n,j}|X_0 = i) = \frac{1}{n}\sum_{k=0}^{n-1} P_{ij}^k$. In this irreducible and periodic case, $\lim_{n\to\infty} \frac{1}{n}\sum_{k=0}^{n-1} P_{ij}^k$ still exists even though $\lim_{n\to\infty} P_{ij}^n$ does not exist. We will state a theorem about this shortly. Before that, let us define the random variable for a state $i \in \mathcal{S}$,
$$T_i = \min\{n \geq 1 : X_n = i\}$$

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

If a (finite) MC is irreducible then $P(T_i < \infty) = 1$ for all states i. It turns out that the mean of the first passage time is intimately related to stationary distributions.

3.7 Fundamental Theorem for Irreducible MC

Let's now restate the fundamental theorem for irreducible MC. This is similar to Theorem5.5 stated in Lecture 5 except we do not assume aperiodicity.

Theorem 3.10 (Fundamental Theorem for Irreducible Markov Chain)**.** *If P is the transition matrix for an irreducible Markov chain then there exists a unique stationary distribution or a unique solution to the equation $\pi = \pi P$ which satisfies the following three properties:*

1.
$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=0}^{n-1} P_{ij}^k = \pi(j)$$
In words, $\pi(j)$ can be interpreted as the limit of the expectation of the proportion of time spent in state j till time $n-1$.

2. *$\pi(j) > 0$ for all $j \in \mathcal{S}$. In words, π gives positive probability to each of the states.*

3. *Consider the first return times T_j for states j. Then we have the following equality*
$$\pi(j) = \frac{1}{\mathbb{E}(T_j|X_0 = j)}.$$
In words, $\pi(j)$ can also be interpreted as the average waiting time for the first return time to state j when the chain starts at j.

Proof. We will assume that there exists a unique stationary distribution π that is positive and which is the limit of the expected proportion of time spent in each state. We will argue that $\mathbb{E}(T|X_0 = i) = \frac{1}{\pi(i)}$. Consider the time until the k th return to the state i. This time is given by a sum of i.i.d random variables $T_1 + \cdots + T_k$ each of which has the same distribution as T conditional on $X_0 = i$. For k large, the Law of Large Numbers say that $\frac{T_1+\cdots+T_k}{k}$ is very close to $\mathbb{E}(T|X_0 = i)$. Therefore, we have approximately k visits to i in $k\mathbb{E}(T|X_0 = i)$ many rounds. Therefore, the expected proportion of time the chain spends in state i is approximately $1/\mathbb{E}(T|X_0 = i)$. Therefore, $\pi(i)$ has to equal $1/\mathbb{E}(T|X_0 = i)$. This is not a fully rigorous proof but with this idea a formal proof can be made. □

Example: Two State MC Consider the transition matrix
$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix}$$

Here, by Theorem 3.10 we have
$$\mathbb{E}(T_0|X_0 = 0) = \frac{1}{\pi(0)} = \frac{p+q}{q}.$$
In this special case, we can actually find the entire distribution of T_0 conditional on $X_0 = 0$.

$$P(T \geq n|X_0 = 0) = P(X_1 = \cdots = X_{n-1} = 1|X_0 = 0) = p(1-q)^{n-2}$$

In general, given only the stationary distribution π, one can only talk about $\mathbb{E}T$ and not its entire distribution. In the above example, consider the case when $p = q$. Then $\mathbb{E}(T_0|X_0 = 0) = 2$. However, if p is close to 1 then $Var(T_0|X_0 = 0)$ is much smaller than when p is close to 0 where the variance can be made arbitrarily large. (**Check this!**)

4 Return Times and Absorption Probabilities

4.1 Expected Number of Visits to a Transient State

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

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

We also can write for any integer $n \geq 1$,

$$P^n = \begin{pmatrix} \tilde{P}^n & 0 \\ S_n & Q^n \end{pmatrix}$$

Consider Example 2 from Lecture 2 of SRW with absorbing boundaries. We can order the state space as $\{0, 4, 1, 2, 3\}$ and write its transition matrix as

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

where

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

Now since the states represented by Q are transient, we must have $\lim_{n\to} Q^n = 0$. Infact, one can show that $I - Q$ in this case is invertible and hence we can define $M = (I - Q)^{-1}$. This matrix M plays an important role in our current discussion.

Let i be a transient state and let us define

$$Y_i = \sum_{n=0}^{\infty} 1(X_n = i).$$

In words, Y_i is a random variable which counts the total number of visits to the state i. Since i is transient, $Y_i < \infty$ w.p 1.

Lemma 4.1. *Let Q denote the part of transition matrix indexed by the transient states. Define $M = (I - Q)^{-1}$. We have the following equality for any two transient states $i, j \in \mathcal{S}$,*

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

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

Proof. We can write

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

The second last equality follows because $P^n_{ji}=Q^n_{ji}$ when j,i are transient states. The last equality follows because

$$I+Q+Q^2+\cdots=(I-Q)^{-1}.$$

□

4.2 Expected Time till Absorption to a Recurrent Class

Let us define

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

In words, T_{abs} is the waiting time till the chain enters a recurrent class. Now suppose the MC starts from a transient state j. A natural question is what is $\mathbb{E}(T_{abs}|X_0=j)$? Note that we can write

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

Therefore, by taking conditional expectation both sides we obtain the following corollary of Lemma 4.1.

Corollary 4.2. *For any transient state $j\in\mathcal{S}$,*

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

Coming back to SRW with absorbing boundaries on $\{0,1,2,3,4\}$ we calculated the transition matrix Q before. We can now calculate

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

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

We could have also obtained the fact that $\mathbb{E}(Y_i|X_0=j)=M_{ij}$ for any two transient states i,j by conditioning on the first step as follows

$$\begin{aligned}\mathbb{E}(Y_i|X_0=j)&=1(i=j)+\sum_{k\ \textit{transient}}\mathbb{E}(Y_i|X_0=j,X_1=k)P(X_1=k|X_0=j)\\&=1(i=j)+\sum_{k\ \textit{transient}}\mathbb{E}(Y_i|X_1=k)Q_{jk}\end{aligned}$$

which can be written in the matrix form and then seen to be equivalent to the conclusion of Lemma 4.1.

4.3 Expected First Return Time

Suppose we have an irreducible MC with a transition matrix P. Recall the first return times T_i. We saw before that $\mathbb{E}(T_i|X_0 = i) = \frac{1}{\pi(i)}$. We now want to calculate $\mathbb{E}(T_i|X_0 = j)$ where $i \neq j$. One way to calculate this is the following. We first write the transition matrix P with i being the first state.

$$P = \begin{pmatrix} P_{ii} & R \\ S & Q \end{pmatrix}$$

We then modify the MC and make i an absorbing state. The transition matrix for this modified chain is

$$\tilde{P} = \begin{pmatrix} 1 & 0 \\ S & Q \end{pmatrix}$$

It is clear that the value of $\mathbb{E}(T_i|X_0 = j)$ will remain the same in both the MC. Since the original MC is irreducible, in this modified MC, the only recurrent state is $\{i\}$ and all the other states including j are transient. Therefore, $\mathbb{E}(T_i|X_0 = j) = \mathbb{E}(T_{abs}|X_0 = j)$ is the expected waiting time till absorption. Therefore, we can calculate it from Corollary 6.4 from Lecture 6.

Work out the example for SRW with reflecting boundary Suppose P is the matrix for SRW with reflecting boundary.

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

Suppose we want to calculate the expected time to hit 0 starting from 4. Then Q consists of the bottom right 4×4 submatrix and then we can calculate

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

Now we have

$$\mathbb{E}(T_0|X_0 = 4) = M_{41} + M_{42} + M_{43} + M_{44} = 16.$$

Another way to calculate $\mathbb{E}(T_i|X_0 = j)$ is to calculate it for all $j \in \mathcal{S}$. Denoting $a_j = \mathbb{E}(T_i|X_0 = j)$, we can condition on the first step of the MC to obtain for each $j \in \mathcal{S}$

the equation

$$a_j = P_{ji} + \sum_{k \neq i} P_{jk}(1 + a_k).$$

The above gives a linear system which can be solved to calculate $\mathbb{E}(T_i|X_0 = j)$ for all $j \in \mathcal{S}$.

Example: Suppose we flip a fair coin repeatedly until we have flipped four consecutive heads. What is the expected number of flips that are needed? Construct a MC with state space $\{0, 1, 2, 3, 4\}$.

To answer the above question, we can construct a MC with state space $\{0, H, HH, HHH, HHHH\}$. The state 0 represents the beginning state or the state you return to anytime when you get tails before reaching state $HHHH$. The state $HHHH$ is an absorbing state. The transition matrix is as follows:

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

The question now is the expected time till absorption. We can again use Corollary 6.4 in Lecture 6 to calculate this expected time.

4.4 Probability of Eventually Entering a Given Recurrent Class

Suppose there are aleast two recurrent classes in a MC. One natural question is **what is the probability that the MC eventually ends up in a given recurrent class starting from a transient state** j For example, this question comes up in the problem of Gambler's ruin. There the recurrent states are $\{0\}$ and $\{N\}$ and the question of interest is precisely the probability of ending up in one of the recurrent states. To answer this question, we can create a modified MC where each of the recurrent classes are seen as single states. Let these states be $r_1, \ldots, r_k$ with $P(r_i, r_i) = 1$. If we order the states so that the recurrent states $r_1, \ldots, r_k$ precede the transient states $t_1, \ldots, t_s$ then the transition matrix looks like

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

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

$\alpha_{t_i,r_j} = P(X_n = r_j \text{ eventually } |X_0 = t_i) = \sum_{x\in\mathcal{S}} P(X_1 = x|X_0 = t_i)P(X_n = r_j \text{ eventually } |X_1 = x) = \sum_{x\in\mathcal{S}} P(t_i, x)\alpha(x, r_j)$.

If $A_{s\times k}$ is a matrix consisting of the entries α_{t_i,r_j} then the above display can be written in a matrix form $A = S + QA$. This means that A can be calculated as follows

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

Gambler's Ruin: We will now consider the problem when the probability of winning a bet is $0 < p < 1$ and not equal to 0.5. The case when $p = 0.5$ has been done before. The MC is simply a random walk with absorbing boundaries on the state space $\{0, 1, \ldots, N\}$. Let α_j be the probability that the MC gets absorbed in state N starting from state j. Clearly, $\alpha(0) = 0, \alpha(N) = 1$. For any $0 < j < N$, we can condition on the first step to get

$$\alpha(j) = (1 - p)\alpha(j - 1) + p\alpha(j + 1).$$

This gives us $N - 1$ linear equations in $N - 1$ unknowns. It can be shown that the general solution of the above linear difference equations in the case $p \neq 0.5$ is

$$\alpha_j = c_1 + c_2\big(\frac{(1-p)}{p}\big)^j$$

The boundary conditions allow us to determine the constants c_1, c_2 so we get

$$\alpha_j = \frac{1 - \big(\frac{(1-p)}{p}\big)^j}{1 - \big(\frac{(1-p)}{p}\big)^N}.$$

Note that if $p \leq 0.5$, then $\lim_{N\to\infty} \alpha(j) = 0$. So if the house has very large resources then the gambler has very little chance of winning if the game is fair or unfair. However, if $p > 0.5$ then the game is in the gambler's favour and $\lim_{N\to\infty} \alpha(j) = 1 - \big(\frac{(1-p)}{p}\big)^j > 0$. In this case, there is a positive chance that the gambler will never lose all his resources and be able to play forever.

Suppose $p = 0.5$ now and let T be the time it takes for the RW to reach 0 or N. In principle, we can use Corollary 6.4 from Lecture 6 to calculate $\mathbb{E}(T|X_0 = j)$. We can also calculate it by conditioning on the first step. Let $G(j) = \mathbb{E}(T|X_0 = j)$. Then $G(0) = G(N) = 0$. Also for $0 < j < N$, we have

$$G(j) = 0.5(1 + G(j - 1)) + 0.5(1 + G(j + 1)).$$

One can show that all solutions of the above inhomogenous linear difference equation are of the form

$$G(j) = j^2 + c_ij + c_2.$$

Plugging in the boundary conditions we get

$$G(j) = j(N - j).$$

5 More Examples of Finite Markov Chains

- **SRW on a Undirected Graph**: Assume the graph $G = (V, E)$ is connected so the chain is irreducible. One can check that $\pi(v) = \frac{deg(v)}{2|E|}$ is a stationary distribution for the chain where $|E|$ is the number of edges in the graph and $deg(v)$ is the degree of the vertex v. Note that $\sum_{v \in V} deg(v) = 2|E|$. As we have seen before, the period of this MC is either 1 or 2. If the period is 1 then π is the limiting distribution for this chain. If the period is 2 then π can still be interpreted as the limiting expected fraction of time spent in each of the states.

- **Ehrenfest Chain** Imagine two dogs - Lisa and Cooper share a population of N fleas. At each unit of time, one of the fleas (randomly picked) jumps from the dog it is on to the other dog. Let X_n denote the number of fleas on Lisa after n jumps. The state space is $\mathcal{S} = \{0, 1, \ldots, N\}$. The transition matrix is given by

  $$P_{ij} = \begin{cases} i/N, \text{ if } j = i - 1 \\ (N - i)/N, \text{ if } j = i + 1 \\ 0, , \text{ if } j \notin \{i - 1, i + 1\} \end{cases}$$

  Exercise: Show that the binomial distribution $Bin(N, 1/2)$ with parameters N and $1/2$ is a stationary distribution for this MC.

  Remark 5.1. *Note that once we are given a candidate stationary distribution we just need to check whether it solves $\pi = \pi P$. However, even if we are not told the stationary distribution we can solve this equation to obtain the stationary distribution.*

  This chain does not have a limiting distribution since the chain is periodic with period 2. This chain is irreducible. Hence, Theorem 5.9 from lecture 5 applies here. We can now modify this chain to make it aperiodic.

  We pick a flea uniformly at random as before and then pick a dog uniformly at random as well for the flea to jump to. The transition matrix P for this modified Ehrenfest chain becomes

  $$P_{ij} = \begin{cases} i/(2N), \text{ if } j = i - 1 \\ (N - i)/(2N), \text{ if } j = i + 1 \\ 1/2, , \text{ if } j = i \\ 0, , \text{ if } j \notin \{i - 1, i, i + 1\} \end{cases}$$

Now the chain is aperiodic and irreducible. Therefore this chain has a limiting distribution. One can check that the stationary distribution is still $Bin(N, 1/2)$ and hence is also the limiting distribution by the fundamental theorem.

Remark 5.2. *Note that for any given flea, the moment we pick that flea once, after that the probability of it being on Lisa or Cooper at any given time is* $1/2$ *from then on, irrespective of where this flea started from. Now clearly there will be a time* T *(which is random) when all the fleas have been picked. After this* T*, it is clear that each flea (independently) has a probability* $1/2$ *of being either on Lisa or Cooper. So in this case the distribution of* X_{T+n} *for* $n \geq 0$ *is exactly Binomial*$(n, 1/2)$*.*

- **Wright Fisher Model in Genetics** Consider the following MC which models reproduction of cells. Suppose each cell contains N particles or genes, each of type A or type B. Suppose a given cell has j particles of type A and $N - j$ particles of type B. When the cell self replicates into two it has $2j$ particles of type A and $2(N - j)$ particles of type B. It then selects N out of $2N$ particles randomly to create a new cell. By using the hypergeometric distribution, we see that this gives rise to transition probabilities

$$P_{jk} = \frac{\binom{2j}{k}\binom{2(n-j)}{N-k}}{\binom{2N}{N}}$$

The MC has two absorbing states 0 and N. Eventually all the particles will either be of type A or type B. A natural question would be what is what would be the fraction of all cells which will have all particles of type A. This would be given by

$$\alpha(j) = P(\text{absorption in state N}).$$

Again, the answer turns out to be j/N. In particular, one can verify that this choice of $\alpha(j)$ satisfies for all $1 \leq j \leq N - 1$,

$$\alpha(j) = \sum_{k=0}^{N} P_{jk}\alpha(k).$$

- **PageRank**

Given a directed graph like the web graph, we have a transition matrix Q associated with the random walk on the graph. Is it irreducible? Perhaps not as there could be nodes which have no outgoing links. It may not be aperiodic as well. To make the chain irreducible and aperiodic we can modify the transition matrix as follows. Before every move the random surfer flips a coin with probability α of heads. If heads, then the random surfer chooses a random outgoing link, if tails then the random surfer chooses a random node out of all the N nodes with equal probability $1/N$. Then the new transition matrix becomes

$$G = \alpha Q + (1 - \alpha)J/N$$

where $J_{N\times N}$ matrix is the matrix of all ones. This chain is irreducible and aperiodic if $0 < \alpha < 1$. This means that there is a unique stationary distribution called the *pagerank* and the chain will converge to this distribution. The choice of α is important. On the other hand choosing α close to 1 respects the structure of the web graph. Choosing α closer to 0 makes the chain converge to the pagerank distribution much faster. Original recommendation of Brin and Page was $\alpha = 0.85$.

Solving $\pi G = \pi$ is hard since the matrices are so huge. However, we can compute tG^n for any initial distribution t for n large as this should converge to π.

$$tG = \alpha(tQ) + \frac{1-\alpha}{N} tJ.$$

Computing tG could be potentially easier as the Q is typically highly sparse and tJ is just the vector of all ones again. Thus one can iteratively compute tG^n until the sequence converges (although it may be hard to know that it has converged.)

- **Card Shuffling**

  Consider a deck of cards numbered $1, \ldots, n$. At each time we will shuffle the cards by drawing a card at random and then placing it at the top of the deck. At each time the ordering of the cards, which is represented by a permutation of $1, \ldots, n$, constitute the MC. The state space is S_n the set of all permutations of $1, \ldots, n.$. If λ is the current permutation for the MC there are n other permutations the MC can go to in the next step with probability $1/n$. This chain is irreducible and aperiodic. It can also be checked that the uniform distribution on S_n is the stationary distribution for this chain and hence the limiting distribution as well by the fundamental theorem. Therefore, if we start with any ordering of the cards, after enough moves the deck will be well shuffled.

  A much harder question is how many moves are enough for the deck to be well shuffled? We will not discuss how to solve this question as this will require advanced techniques. Questions like the expected number of moves from a given permutation to another permutation, can be answered in principle by matrix computatons. However, since the size of the matrix is $n! \times n!$ this is not feasible.

6 Countably Infinite Markov Chains

Now we will consider the case when the state space $\mathcal{S}$ is countably infinite. The transition matrix is now an infinite matrix. The Chapman Kolmogorov Equation below still holds:

$$P^{m+n}(x,y) = \sum_{z\in\mathcal{S}} P^m(x,z)P^n(z,y)$$

The only difference now is that the above sum is an infinite sum.

Some examples of countably infinite MC are

Example 1: Random Walk with Partially Reflecting Boundaries

Let $0 < p < 1$ and $\mathcal{S} = \{0, 1, 2, \dots\}$. The transition probabilities are given by $p(x, x-1) = 1-p, p(x, x+1) = p$ when $x > 0$ and $p(0,0) = 1-p$ and $p(0,1) = p$.

Example 2: Simple Random Walk on the Integer Lattice

The state space is $\mathbb{Z}^d$. Any point in $\mathbb{Z}^d$ has $2d$ neighbors and the RW moves to one of them with equal probability. Points x, y are neighbors in $\mathbb{Z}^d$ iff $|x - y| = 1$. Therefore, the transition probabilites are given by $P(x, y) = \frac{1}{2d}$ if $|x-y| = 1$ and 0 otherwise.

Example 3: Queueing Model

Let X_n be the number of customers waiting for a service at time n. During each time interval there is a probability p that a new customer arrives. Independently, with probability q the service for the first customer is completed and the customer leaves the queue. This is a MC on the state space $\mathcal{S} = \{0, 1, 2, \dots\}$. The transition probabilites are $p(x, x-1) = q(1-p), p(x, x-1) = p(1-q), p(x, x) = pq + (1-p)(1-q)$ if $x > 0$. Otherwise, $p(0, 0) = 1-p$ and $p(0, 1) = p$. This is an example of a birth and death chain.

As in the case of finite MC, we are interested in long run behavior of MC. Some of the ideas introduced so far apply equally well to the infinite case. For example, the notion of communication classes apply, we call a MC irreducible if all the states communicate with each other. All the examples above are irreducible. We can also talk of the period of a chain in the same way for a communicating class. Examples 1 and 3 above are aperiodic and 2 has period 2. The main difference with the finite case will be that an irreducible, aperiodic MC with infinite state space may not have a stationary distribution and thus will not have a limiting distribution as well.

6.1 Recurrence and Transience

Recall the definition of the first return times

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

Also define

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

Let us now define the notion of recurrent and transient classes more generally this time.

Definition 6.1. *A state j is recurrent if $f_j = 1$ and transient if $f_j < 1$.*

Lemma 6.2. *Recurrence and Transience are class properties. This means that if one state is recurrent/transient then all other states in that communication class are recurrent/transient.*

Proof. Suppose i is recurrent and j communicates with i. We need to show that j is recurrent or $P(T_j < \infty | X_0 = j) = 1$. It is enough to show that $P(T_j < \infty | X_0 = i) = 1$ and

$P(T_i < \infty | X_0 = j) = 1$. This is because the MC is certain to go to j from i and also certain to go to i from j and hence starting from j it is certain to come back to j.

Let us show that $P(T_j < \infty | X_0 = i) = 1$. Imagine starting the chain in state i, so that $X_0 = i$. With probability one, the chain returns at some time $T_i < \infty$ to i. For the same reason, continuing the chain after time T_i, the chain is sure to return to i for a second time. In fact, by continuing this argument we see that, with probability one, the chain returns to i infinitely many times. Thus, we may visualize the path followed by the Markov chain as a succession of infinitely many cycles, where a cycle is a portion of the path between two successive visits to i. That is, we will say that the first cycle is the segment $X_1, ..., X_{T_i}$ of the path, the second cycle starts with $X_{T_{i+1}}$ and continues up to and including the second return to i, and so on. The behaviors of the chain in successive cycles are independent and have identical probabilistic characteristics. In particular, letting $I_n = 1$ if the chain visits j sometime during the nth cycle and $I_n = 0$ otherwise, we see that $I_1, I_2, \ldots$ is an iid sequence of Bernoulli trials. Let p denote the common success probability

$$p = P(\text{visit j in a cycle}) = P(\cup_{k=1}^{T_i} \{X_k = j\} | X_0 = i)$$

for these trials. Clearly if p were 0, then with probability one the chain would not visit j in any cycle, which would contradict the assumption that j communicates with i. Therefore, $p > 0$. Now observe that in such a sequence of iid Bernoulli trials with a positive success probability, with probability one we will eventually observe a success. In fact,

$$P(\text{chain does not visit j in the first n cycles}) = (1-p)^n \to 0$$

as $n \to \infty$. That is, with probability one, eventually there will be a cycle in which the chain does visit j. This shows that $P(T_j < \infty | X_0 = i) = 1$.

Now suppose to the contrary that $P(T_i = \infty | X_0 = j) > 0$. Combining this with the hypothesis that j communicates with i, we see that it is possible with positive probability for the chain to go from i to j in some finite amount of time, and then, continuing from state j, never to return to i. But this contradicts the fact that starting from i the chain must return to i infinitely many times with probability one. Thus, $P(T_i < \infty | X_0 = j) = 1$ and we are done. □

Question: How can we determine whether a chain is recurrent or transient? One way is given by the following lemma.

Lemma 6.3. *An irreducible MC is transient if and only if the expected number of returns to a state is finite; i.e*

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

Proof. Define the random variable Y_i which counts the total number of visits to i.

$$Y_i = \sum_{n=0}^{\infty} 1(X_n = i).$$

We can see that

$$\mathbb{E}(Y_i|X_0 = i) = \sum_{n=0}^{\infty} P(X_n = i|X_0 = i) = \sum_{n=0}^{\infty} P_{i,i}^n.$$

Now if i is recurrent then starting from i, we know that with probability one, the number of visits to i is infinite. This in particular implies that $\mathbb{E}(Y_i|X_0 = i) = \infty$.

Now for the other part, we will show that if i is transient then $\mathbb{E}(Y_i|X_0 = i) < \infty$. Suppose i is transient. Then $P(T_i = \infty|X_0 = i) = q > 0$. So each time the chain is at i, we can think of a bernoulli trial and define failure if it returns to i (which happens with probability $1 - q$) and success if it never returns (which happens with probability q). Then the number of visits Y_i is the same as the number of trials till we obtain the first success. Therefore the distribution of $Y_i|X_0 = i$ is Geometric with success probability q. This means that $\mathbb{E}(Y_i|X_0 = i) = \frac{1}{q} < \infty$. □

6.2 Recurrence/Transience of Simple Random Walk on Lattice

Is the d dimensional SRW recurrent or transient? To answer this question we can use Lemma 6.3. Lets first consider the $d = 1$ case. Suppose we consider the state 0. Clearly, you can only return back to 0 from 0 in an even number of steps and this can happen only if you take equal number of left and right steps. The number of such trajectories is $\binom{2n}{n}$ and each of them have probability $\frac{1}{2^n}$. Therefore, for any $n \geq 1$,

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

It is not so clear what happens if I sum over $P_{0,0}^{2n}$ now. To simplify the above expression we can use Stirling's formula which estimates $n! \sim \sqrt{2\pi n} n^n \exp(-n)$ where $a_n \sim b_n$ means that $\lim_{n\to} \frac{a_n}{b_n} = 1$. Using Stirling's formula to simplify the factorials one gets

$$P_{0,0}^{2n} \sim \frac{1}{\sqrt{\pi n}}.$$

Now, since $\sum_{n=1}^{\infty} \frac{1}{\sqrt{n}} = \infty$ one can expect that $\sum_{n=0}^{\infty} P_{0,0}^{2n} = \infty$. Then Lemma 6.3 says that the 1 dimensional RW is recurrent.

It turns out that in d dimensions one can show that

$$P_{0,0}^{2n} \sim \frac{1}{2^{d-1}}(\frac{d}{\pi n})^{d/2}.$$

Now recall that $\sum n^{-a} < \infty$ if and only if $a > 1$. Hence, we have

$$\begin{cases} \sum_{n=0}^{\infty} P_{0,0}^{2n} = \infty \text{ if } d = 1, 2 \\ \sum_{n=0}^{\infty} P_{0,0}^{2n} < \infty \text{ if } d > 2. \end{cases}$$

This gives the following fact which is often called Polya's Theorem. This can be said to be one of the famous results in probability.

Theorem 6.4 (Polya). *Simple Random Walk in $\mathbb{Z}^d$ is recurrent if $d = 1$ or $d = 2$ and is transient if $d \geq 3$.*

The mathematician Shizuo Kakutani explained this result by saying " A drunk man will find his way home but a drunk bird may get lost forever."

Exercise: Show that RW in one dimension with probability p of going right is transient when $p \neq 1/2$. Hint: Again use Stirling's formula and examine convergence of the sum of $P^{2n}_{0,0} = \binom{2n}{n} p^n (1-p)^n$.

6.3 Null and Positive Recurrence

Define the expected first return times

$$\mu_j = \mathbb{E}(T_j | X_0 = j).$$

If a state is transient then $P(T_j = \infty | X_0 = j) > 0$. Therefore, $\mu_j = \infty$. When a state is recurrent, we know that $P(T_j < \infty | X_0 = j) = 1$. However, this does not mean that $\mu_j < \infty$. Suppose the state j is recurrent. A big difference between finite and infinite state space is that when $\mathcal{S}$ is finite, then necessarily $\mu_j < \infty$. However, when $\mathcal{S}$ is countably infinite, both $\mu_j < \infty$ and $\mu_j = \infty$ are possible. We now make the following definition.

Definition 6.5. *A state j is positive recurrent if it is recurrent and $\mu_j < \infty$. A state j is null recurrent if it is recurrent and $\mu_j = \infty$.*

We will see that 1D SRW is null recurrent.

Lemma 6.6. *Positive Recurrence and Null Recurrence are class properties. This means that if one state is positive/null recurrent then all other states in that communication class are positive/null recurrent.*

6.3.1 Stationary Distribution and Limiting Distribution

Now we investigate when a limiting distribution might exist for an irreducible MC. A limiting distribution π is a probability distribution on $\mathcal{S}$ such that for each $x, y \in \mathcal{S}$,

$$\lim_{n\to\infty} P^n(y, x) = \pi(x).$$

If the chain is transient then $\lim_{n\to\infty} P^n(y, x) = 0$. Hence, no limiting distribution can exist. If the chain is recurrent, it is still possible that $\lim_{n\to\infty} P^n(y, x) = 0$. Infact, this precisely happens when the chain is null recurrent. We record this fact as a lemma.

Lemma 6.7. *For an irreducible MC, $\lim_{n\to\infty} P^n(y, x) = 0$ for each $x, y \in \mathcal{S}$ iff the chain is transient or null recurrent.*

Positive recurrent chains behave very similarly to finite Markov chains. We now state restate the fundamental theorem generalizing it to hold also for countably infinite chains.

Theorem 6.8 (Fundamental Theorem for General Discrete Markov Chains)**.** *An irreducible, positive recurrent MC has a unique stationary distribution π (which is positive everywhere) solving the equation*

$$\sum_{y\in\mathcal{S}} \pi(y)P(y,x) = \pi(x) \;\; \forall x,y \in \mathcal{S}. \tag{1}$$

Moreover, π satisfies for all $i,j \in \mathcal{S}$,

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

If in addition, the MC is aperiodic then the last conclusion can be strengthened to

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

In words, π is the limiting distribution for this chain.

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

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

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

Question: How can we determine whether a chain is positive recurrent? One way is to try to solve for a stationary distribution. If a stationary distribution exists then it is positive recurrent. If not, then it is transient or null recurrent.

Example: 1D SRW Let's try to solve for a stationary distribution π which should satisfy for all $x \in \mathbb{Z}$,

$$\frac{\pi(x+1)}{2} + \frac{\pi(x-1)}{2} = \pi(x).$$

The above equation means that there cannot be any local maxima of π and hence π needs to be a constant function. Therefore, there cannot be any stationary distribution. We know that this chain is recurrent. Hence, the chain must be null recurrent as there is no stationary distribution.

Example: SRW with Partially Reflecting Boundary

Let's try to solve for a stationary distribution π which should satisfy for all $x > 0$,

$$\pi(x+1)(1-p) + \pi(x-1)p = \pi(x).$$

When $x = 0$ it should satisfy $\pi(1)(1-p) + \pi(0)(1-p) = \pi(0)$. The general solution to the above is

$$\pi(x) = \begin{cases} c_1 + c_2 \frac{p}{(1-p)}^x, \text{ if } p \neq 1/2 \\ c_1 + c_2 x \text{ if } p = 1/2. \end{cases}$$

Plugging in $\pi(0) = \frac{(1-p)}{p}\pi(1)$ in the above equations we get

$$\pi(x) = \begin{cases} c_2 \frac{p}{(1-p)}^x, \text{ if } p \neq 1/2 \\ c_1 \text{ if } p = 1/2. \end{cases}$$

Now $\sum_x \pi(x)$ needs to be equal to 1. Therefore, we cannot find such a π when $p = 1/2$. So, suppose $p \neq 1/2$. Clearly, we would need $c_2 \neq 0$. If $p > 1/2$ then $\sum_x \frac{p}{(1-p)}^x = \infty$ and hence we cannot find a proper c_2. Infact, in this case the chain is transient (Why?) and hence cannot be positive recurrent. However, if $p < 1/2$ then the sum is finite and we can choose

$$\pi(x) = (\frac{p}{(1-p)})^x (\sum_x (\frac{p}{(1-p)})^x)^{-1} = \frac{(1-2p)p^x}{(1-p)(1-p)^x}.$$

In this case the chain is positive recurrent and the above is the unique stationary distribution for the chain which is also the limiting distribution by Theorem 6.8.

We can summarize our discussion by saying that SRW with partially reflecting boundary is positive recurrent if $p < 1/2$ and not positive recurrent if $p \geq 1/2$. It turns out that it is null recurrent if $p = 1/2$ and transient if $p > 1/2$.

6.4 Recap of Differences between Finite and (Countably) Infinite Markov Chains

1. An irreducible MC has to be recurrent if $\mathcal{S}$ is finite. An irreducible MC could be recurrent or transient if $\mathcal{S}$ is infinite. Equivalently, $P(T_j < \infty | X_0 = j) = 1$ if $\mathcal{S}$ is finite but not necessarily so if $\mathcal{S}$ is infinite.

2. An irreducible MC is recurrent means that $P(T_j < \infty | X_0 = j) = 1$. However, this does not say whether $\mathbb{E}(T_j | X_0 = j)$ is finite or infinite. An irreducible recurrent MC is necessarily positive recurrent if $\mathcal{S}$ is finite. Equivalently, $\mathbb{E}(T_j | X_0 = j) < \infty$. An irreducible recurrent MC can have $\mathbb{E}(T_j | X_0 = j) < \infty$ or $\mathbb{E}(T_j | X_0 = j) = \infty$ if if $\mathcal{S}$ is infinite. Accordingly, the chain is positive recurrent or null recurrent.

3. If $\mathcal{S}$ is finite then an irreducible MC always has a unique stationary distribution. In fact, it can be shown that any finite MC has a stationary distribution (not necessarily unique). If $\mathcal{S}$ is infinite then an irreducible MC need not even have a stationary distribution and consequently not have a limiting distribution. Specifically, transient or null recurrent irreducible MC does not have a stationary distribution.

4. One needs to add the qualifier positive recurrent (in addition to irreducibility and/or aperiodicity) for the fundamental theorem to hold in general when the state space $\mathcal{S}$ is allowed to be infinite.

7 Branching Process

The branching process model we will study was formulated in 1873 by Sir Francis Galton, who was interested in the survival and extinction of family names. It is a stochastic model for population growth. Let X_n denote the number of individuals at time n. At each time interval, each individual will produce a random number of offsprinngs and then die. The two main assumptions about this reproduction process are

1. Each individual produces offspring with the same probability distribution: there are given non negative numbers $p_0, p_1, \dots$ summing up to 1 so that the probability of an individual producing k children is p_k.
2. The individuals reproduce independently.

Here is a question: What is the probability that the population eventually becomes extinct? Galton brought the problem to his mathematician friend, Rev. H. W. Watson, who devised the method of analysis using probability generating functions that is still used today. However, a minor mathematical slip caused Galton and Watson to get the answer to the main question wrong. They believed that the extinction probability is 1 i.e, all families or populations are doomed to eventual extinction. We will see below that this is false: if the expected number of sons is greater than 1, the branching process model produces lines of descent that have positive probability of going on forever.

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

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

Clearly state 0 is absorbing. Therefore, for each $i > 0$, since $P(X_1 = 0 | X_0 = i) = p_0^i > 0$, the state i must be transient. Consequently, we know that with probability 1, each state $i > 0$ is visited only a finite number of times. From this, it can be shown that, with probability 1, the chain must either get absorbed at 0 eventually or approach ∞. *Can you show this?*

7.1 Extinction Probability in a Branching Process

Let μ denote the mean number of offsprings produced by an individual.

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

It is straightforward to calculate the mean number of individuals in generation n, that is $\mathbb{E}X_n$. We can set up the recurrence relation by conditioning on X_{n-1},

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

This means that

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

From here, one can deduce the following.

Lemma 7.1. *If $\mu < 1$, then probability of extinction is* 1.

Proof. The event that the population becomes extinct is the same as $\cup_{n=0}^{\infty}\{X_n = 0\}$. Note that the events $\{X_n = 0\}$ is an increasing sequence of events in the sense that $\{X_{n-1} = 0\} \subseteq \{X_n = 0\}$ for all n. Therefore, we have

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

We are using the fact that if A_n is an increasing sequence of events then $P(\cup A_n) = \lim_{n\to} P(A_n)$. Show this using the axiom of countable additivity!

Now, we have

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

The above inequality is basically an instance of Markov's inequality. Now, if $\mu < 1$, then $\lim_{n\to\infty} \mathbb{E}X_n = \mathbb{E}X_0 \lim_{n\to\infty} \mu^n = 0$. The last two displays finish the proof. □

If $\mu = 1$, the expected population size remains constant while if $\mu > 1$, the expected population size grows. From this information it is not so clear whether or not the population dies out with probability 1. This is because it is possible for X_n to be 0 with probability very near 1 yet $E(X_n)$ not be small.

While we are at it let's calculate the variance of X_n. We will again condition on X_{n-1} and use the law of total variance. Let's denote the variance of the number of offsprings produced by an individual by σ^2.

$$Var X_n = Var(\mathbb{E}X_n | X_{n-1}) + \mathbb{E}Var(X_n | X_{n-1}) = Var(\mu X_{n-1}) + \mathbb{E}(\sigma^2 X_{n-1}) = \mu^2 Var(X_{n-1}) + \sigma^2 \mu^{n-1} \mathbb{E}X_0.$$

We can now solve the above recurrence relation to obtain the formula (below I am assuming that $X_0 = 1$ with probability 1.)

$$VarX_n = \begin{cases} n\sigma^2 \text{ if } \mu = 1 \\ \sigma^2\mu^{n-1}(\mu^n - 1)/(\mu - 1) \text{ if } \mu = 1. \end{cases}$$

We see that the variance grows linearly when $\mu = 1$ and grows exponentially when $\mu > 1$. Such a large variance means that there is a possibility that X_n takes value 0 with positive probability even if $\mathbb{E}X_n$ is very large. We now investigate how to calculate the probability of extinction.

In order to avoid trivial cases, let us assume that

1. $p_0 > 0$.
2. $p_0 + p_1 < 1$.

Why are these cases trivial? If $p_0 = 0$ then probability of extinction is 0. If $p_0 > 0$ and $p_0 + p_1 = 1$ then the probability of extinction is 1. So let's operate under the above two assumptions.

Let $a_n(k) = P(X_n = 0|X_0 = k)$ and let $a(k)$ denote the probability that the population dies out eventually assuming that there are k individuals initially. Then by the logic given in the proof of Lemma 7.1,

$$a(k) = \lim_{n\to\infty} a_n(k).$$

Now, if $X_0 = k$ then the only way the population becomes extinct if all the k branches die out. Since the branches act independently, we must have

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

It suffices therefore to compute $a(1)$ which we will simply denote by ρ. Therefore,

$$\rho = P(\text{extinction}|X_0 = 1) = P(\cup_{n=0}^{\infty}\{X_n = 0\}|X_0 = 1) = \lim_{n\to\infty} P(X_n = 0|X_0 = 1).$$

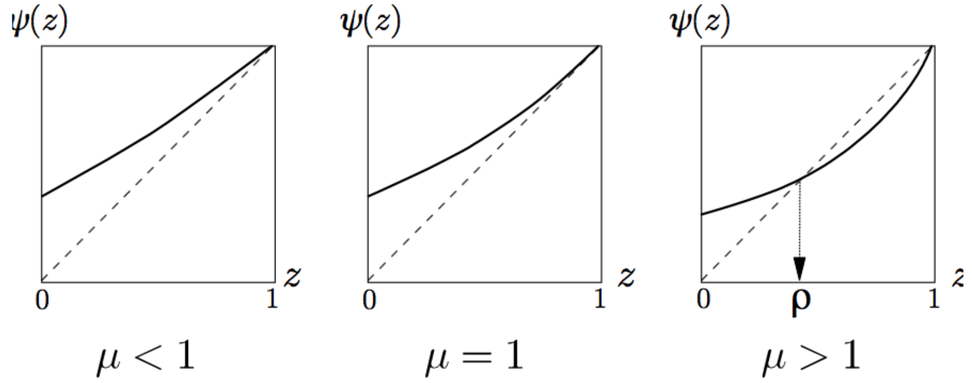
By conditioning on the first step, we can write

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

where $\psi : [0, 1] \to \mathbb{R}$ is given by $\psi(z) = p_0 + p_1 z + p_2 z^2 + \ldots$. Therefore, the desired probability ρ satisfies the equation $z = \psi(z)$. The function ψ is of sufficient interest to be given a name of its own.

Definition 7.2. *If a random variable X takes values in $\mathbb{Z}$, the probability generating function (pgf) of X is the function $\psi : [0, 1] \to \mathbb{R}$ given by*

$$\psi(x) = \psi_X(x) = \sum_{k=0}^{\infty} x^k P(X = k).$$



We now note some important properties of the function ψ.

1. $$\psi'(x) = \sum_{k=1}^{\infty} x^{k-1} k p_k.$$

   Hence $\psi' > 0$ on $(0,1)$ and hence the function ψ' is an increasing function.

2. $$\psi''(x) = \sum_{k=2}^{\infty} x^{k-2} k(k-1) p_k.$$

   Hence $\psi'' > 0$ on $(0,1)$ and hence the function ψ is a convex function.

3. $\psi(0) = p_0 > 0$.

4. $\psi(1) = 1$.

5. $\psi'(1) = \mu$.

These properties imply that the graph of ψ over $[0,1]$ must look like one of the three following pictures above, depending on the value of $\mu = \psi'(1)$.

Remark 7.1. *Probability Generating Functions characterize the distribution. This means that if two discrete random variables have their pgf the same then they have the same distribution. The other important fact about pgf is that pgf of the sum of two independent random variables $X+Y$ is the product $\psi_{X+Y}(x) = \psi_X(x)\psi_Y(x)$. This follows by noting that one can write $\psi_{X+Y}(x) = \mathbb{E}x^{X+Y} = \mathbb{E}x^X \mathbb{E}x^Y = \psi_X(x)\psi_Y(x)$.*

From the pictures we can see what happens. Since $\psi(1) = 1$, the equation $\psi(z) = z$ always has a trivial solution at $z = 1$. When $\mu \leq 1$, this trivial solution is the only solution, so that, since the probability ρ of eventual extinction satisfies $\phi(z) = z$, it must be the case that $\rho = 1$. When $\mu > 1$, there is one additional solution, indicated by the picture. This solution was missed by Watson and Galton (1875), leading them to believe that the probability of extinction would be 1 in this case as well. We will show that this was incorrect, and that the probability of extinction is the smaller solution of the equation $\psi(z) = z$.

So suppose $\mu > 1$. Defining r to be the smaller solution of $\psi(z) = z$ we want to show that $\rho = r$. Since $\phi(\rho) = \rho$ we know that ρ must be either r or 1. Denoting $q_n = P(X_n = 0|X_0 = 1)$, we know that q_n is a non decreasing sequence with $\lim_{n\to\infty} q_n = \rho$. Therefore, to show that $\rho = r$ it suffices to show that $q_n \leq r$ for all $n \geq 0$. We will do this by induction. observe that $q_0 = 0$, so that the statement holds for $n = 0$. Assume that $q_n \leq r$. Now we can write for any $n \geq 0$,

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

Therefore, since ψ is increasing we have $q_{n+1} = \psi(q_n) \leq \psi(r) = r$. We now collect what we have shown in the following theorem.

Theorem 7.3. *If $\mu < 1$ (sub-critical regime) or $\mu = 1$ (critical regime), the extinction probability $\rho = 1$, i.e, the population eventually dies out with probability one. If $\mu > 1$ (super-critical regime), then the extinction probability $\rho < 1$ and equals the unique root of the equation $z = \psi(z)$ on $0 < z < 1$.*

Example: Suppose each man has 3 children, with each child having probability 1/2 of being male, and different children being independent. What is the probability that a particular man's line of male descendants will eventually become extinct?

Here the distribution of male offsprings is the binomial distribution $Bin(3, 1/2)$, so that $\mu = 3/2 > 1$. Thus, we are in the supercritical case and we know that the probability ρ of extinction is less than 1. Here $p_0 = 1/8, p_1 = 3/8, p_2 = 3/8$ and $p_3 = 1/8$ so that the equation $\psi(r) = r$ becomes $1 + 3r + 3r^2 + r^3 = 8r$ or $r^3 + 3r^2 - 5r + 1 = 0$. Fortunately, $r = 1$ is a solution (as it must be!), so we can factor it out, getting the equation $(r-1)(r^2+4r-1) = 0$. Solving the quadratic equation gives $\rho = \sqrt{5} - 2 = 0.2361$. The man can rest assured that with probability $1 - \rho = 0.7639$ his glorious family name will live on forever.

Remark 7.2. *How to simulate probability of extinction? Just run several branching process Markov Chains till some fixed time n. Then at this round, look at the fraction of chains which are not extinct. This is an estimate of the actual probability. Of course, there is a bias to this estimate. However, as you will see if you simulate, if a chain goes extinct it will go extinct pretty quickly. So as long as n is reasonably large, the bias of your estimate should be small.*

8 Time Reversible Markov Chains

8.1 Definition and Characterization

Let $X_0, X_1, \ldots$ be a Markov chain having probability transition matrix P. Imagine that I recorded a movie of the sequence of states $(X_0, \ldots, X_n)$ and I am showing you the movie on my fancy machine that can play the tape forward or backward equally well. Can you tell by watching the sequence of transitions on the movie whether I am showing it forward or backward?

Of course, we are assuming that you know the transition matrix P ; otherwise, this would be an unreasonable request. There are cases in which distinguishing the direction of time is very easy. For example, if the state space is $1, 2, 3$ and $P(1,2) = P(2,3) = P(3,1) = 1$, observing just one transition of the chain is enough to tell you for sure the direction of time; for example, a movie in which we observe 3 followed by 2 must be running backward.

That one was easy. Lets consider another example: do you think a stationary Ehrenfest chain is time-reversible? Here the state space is $\{0, 1, ..., d\}$, say, and $X_0 \sim Bin(d, 1/2)$ the stationary distribution of the chain. It is clear in this case that you will not be able to tell for sure from observing any finite movie $(X_0, ..., X_n)$ which direction the movie is being shown. A sequence has positive probability if and only if its reversal also has positive probability. But we are asking whether or not you can get any sort of probabilistic hint about the direction in which the movie is being shown, and I am willing to show you as long a segment as you would like to request. So you can have plenty of data to look at. One might suspect that it should be possible to make this sort of distinction. For example, we know that the Ehrefest chain has a restoring force that pulls it toward the level d/2, where half of the fleas are in each of the two dogs. So, for instance, if we observe a long sequence that moves from (3/4)d down toward d/2, we might favor the explanation that the movie is being shown forward, since otherwise we are observing a long sequence moving against the restoring force. Did you buy that? I hope not, because in fact we will see that the Ehrenfest chain is time-reversible: no movie, no matter how long, will give you any probabilistic information that is useful in distinguishing the direction of time. [And the argument suggested above really didnt make much sense. What comes down must have gone up.]

Definition 8.1. *We say that a MC is time reversible if, for each $n \geq 1$, the distribution of $(X_0, \ldots, X_n)$ is the same as the distribution of $(X_n, \ldots, X_0)$. Equivalently, for any $x_0, \ldots, x_n \in \mathcal{S}$ we have*

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

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

Suppose a MC is time-reversible. As a particular consequence of the definition, we see that (X_0, X_1) must have the same dsitribution as (X_1, X_0). This, in turn, implies that the distribution of X_1 is the same as that of X_0. Thus, the initial distribution π_0 must satisfy $\pi_0 = \pi_0 P$ and hence is stationary. Not surprisingly, we have found that a time-reversible chain must be stationary. We will write π for the distribution of X_0 to emphasize that it is stationary. So $X_n \sim \pi$ for all n. The condition that the distribution of (X_0, X_1) is the same as the distribution of (X_1, X_0) also says that $P(X_0 = i, X_1 = j) = P(X_1 = i, X_0 = j)$ for all $i, j \in \mathcal{S}$ that is,

$$\pi(i)P(i,j) = \pi(j)P(j,i) \;\; \forall i, j \in \mathcal{S}. \tag{2}$$

We have shown that the above condition together with $X_0 \sim \pi$ is necessary for a chain to be reversible. In fact, these two conditions are also sufficient for reversibility.

Lemma 8.2. *The Markov chain $X_0, X_1, \dots$ is time-reversible if and only if the distribution π of X_0 satisfies the condition (2).*

Proof. Let's show the if part as we have already shown the only if part.

$$P(X_0 = x_0, X_1 = x_1, \dots, X_n = x_n) = \pi(x_0)P_{x_0,x_1}P_{x_1,x_2}\dots P_{x_{n-1},x_n} =$$
$$P_{x_1,x_0}\pi(x_1)P_{x_1,x_2}\dots P_{x_{n-1},x_n} = P_{x_1,x_0}P_{x_2,x_1}\pi(x_2)P_{x_2,x_3}\dots P_{x_{n-1},x_n} = \dots =$$
$$\pi(x_n)P_{x_n,x_{n-1}}\dots P_{x_1,x_0} = P(X_n = x_0, X_{n-1} = x_{n-1}, \dots, X_0 = x_n).$$

Notice how (2) allowed the π factor to propagate through the product from the left end to the right, reversing the direction of all of the transitions along the way. □

Remark 8.1. *The condition in 2 are together referred to as local or detailed balance equations.*

8.2 Discussion about Local Balance

We can visualize a MC in the following way. We think of a system of containers of fluid connected by tubes, one container for each state, and we think of probability as fluid flowing around the system. So, if the initial distribution is π^0 then the amount of fluid at state or container j is π^0_j. Now each pair of states is connected by a tube and the tubes are one way, meaning that for any pair of states i, j, one tube allows fluid to go from i to j and the other tube allows fluid to flow from j to i. After one unit of time, the fluid has flowed. For example, from state i the amount of fluid that has flowed to j is $\pi^0_i P_{ij}$. The amount of fluid now in each state corresponds to the distribution of the MC at time 1 which is $\pi^1 = \pi^0 P$. The amount of fluid in state or container j now is π^1_j. Now, saying the chain is stationary is equivalent to saying that the total fluid that goes out of any state is the same as the total fluid that comes in to that state. To make this concrete, lets define the notion of probability flow.

Definition 8.3. *For a distribution π on the state space $\mathcal{S}$ and any two subsets of the state space A, B define*

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

If π represents the amount of liquid at the various states, then $Flow(A, B)$ is the amount of fluid flowing from A to B in one unit of time.

Exercise: Show that if π is the stationary distribution of a MC then $Flow(A, A^c) = Flow(A^c, A)$ for any subset $A \subset \mathcal{S}$. Hint: It is enough to show for $A = \{i\}$ for all $i \in \mathcal{S}$.

However, (2) says something extra. It says that for any pair of states i, j, the amount of fluid that goes in from i to j is the same as the amount of fluid that comes in from j to i. This is the same thing as saying that $Flow(\{i\}, \{j\}) = Flow(\{j\}, \{i\})$. Thus, there is a kind of *local balance* in the system. Now, one should expect that if there is such a local balance then the chain is stationary globally as well.

Lemma 8.4. *If the local balance equations hold then π is stationary.*

Proof. We need to show $\pi P = \pi$. Let's fix a coordinate i.

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

□

So why is the Ehrenfest chain time-reversible? The Ehrenfest chain is an example of a birth and death chain, which is defined to be a Markov chain whose states consist of nonnegative integers and whose transitions increase or decrease the state by at most 1. That is, interpreting the current state of the chain as the population count of living individuals, the population can change by at most 1 in a transition, which might represent a birth, a death, or no change. The time reversibility of the Ehrenfest chain is an example of a more general fact.

Lemma 8.5. *All stationary birth and death chains are time reversible.*

Proof. We need to show (2) holds for all $i, j \in \mathbb{Z}_+$. The cases when $i = j$ or $|i - j| > 1$ are trivial. So it is enough to show take (2) when $j = i + 1$.

Since π is stationary we know that $Flow(A, A^c) = Flow(A^c, A)$ for any $A \subset \mathcal{S}$. Take $A = \{0, 1, 2, \ldots, i\}$. Then we have

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

where the last equality is because the MC is a birth and death chain and $P_{ij} = 0$ if $|j - i| > 1$. Now since $Flow(A, A^c) = Flow(A^c, A)$ therefore $\pi(i) P_{i,i+1} = \pi_{i+1} P_{i+1,i}$ for any $i \ge 0$.

□

Example: Random Walk on an Undirected Graph

Lemma 8.6. *Any stationary random walk on a weighted undirected graph is time reversible. On the other hand, any time reversible MC can be thought of as a random walk on a weighted undirected graph.*

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

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

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

The right hand side above is symmetric in u, v so local balance must hold.

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

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

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

□

We can summarize our discussion by saying that

1. Any MC can be thought of as a random walk on a weighted directed graph. (Why?)

2. If the MC is time reversible then it can be thought of as a random walk on a weighted undirected graph. An undirected weighted graph is a special case of a weighted directed graph where the weights are symmetric, i.e, for any pair of states i, j we have $W_{ij} = W_{ji}$.

8.3 Checking Local Balance suffices to show Stationarity

It is often easier to show that π satisfies local balance and hence is stationary in a given setting than directly showing π is stationary. For example, we now know that in an Ehrenfest chain, to show that the binomial distribution is stationary, we only need to show that local balance holds. We have already seen the example of a random walk on a weighted undirected graph. Let's see some more examples.

Example 1**:** Suppose the transition matrix P is symmetric, i.e, $P_{ij} = P_{ji}$. Then it is easy to check that when the state space $\mathcal{S}$ is finite, the uniform distribution on $\mathcal{S}$ satisfies local balance and hence is stationary. Let's consider the card shuffling MC with a random card being drawn and put back to top. A little bit of thought shows that the transition matrix here is symmetric. Therefore, the uniform distribution on the set of all orderings is stationary. This then implies that the uniform distribution on the set of all orderings is also the limiting distribution because the chain is irreducible, aperiodic. Any reasonable shuffling scheme should have a symmetric transition matrix and the same logic would apply.

Example 2**:** Consider a general birth and death chain on the state space $\mathbb{Z}_+$. Let $P_{i,i+1} = p_i$ for $i = 0, 1, \dots$ and $q_i = P_{i,i-1}$ for $i = 1, \dots$. We want to solve for a stationary distribution. By lemma 8.5 it is enough to check local balance for any $i, i+1$. So the equations for local balance are $x_i p_i = x_{i+1} q_{i+1}$ for $i = 0, 1, \dots$. Suppose $x_0 = 1$. Then $x_1 = p_0/q_1$ and $x_2 = (p_0 p_1)/(q_1 q_2)$. In general, we can write

$$x_k = \Pi_{i=1}^k \frac{p_{i-1}}{q_i}.$$

We can now normalize to obtain that the stationary distribution π must be equal to

$$\pi_k = \frac{x_k}{\sum_{j=0}^{\infty} x_j}$$

provided $\sum_{j=0}^{\infty} x_j < \infty$**.**

9 Markov Chain Monte Carlo

A fundamental problem that arises in a lot of scientific disciplines is how to simulate from a complex and high dimensional distribution. Markov Chain Monte Carlo (MCMC) is a methodology, which uses Markov Chains to simulate from seemingly intractable distributions. Given a probability distribution π, the goal of MCMC is to simulate a random variable X whose distribution is π. The distribution π may be discrete or continuous. In the beginning, we will consider discrete distributions. Often, we want to estimate an expectation of some function of X; i.e, $\mathbb{E}f(X)$ where $X \sim \pi$. We can do this by first simulating from π using MCMC and then using law of large numbers as we will explain.

The main idea of the MCMC algorithm is to construct an irreducible, aperiodic (positive recurrent if infinite) MC whose stationary distribution is π**. Then by the fundamental theorem, we know that** π **is also the limiting distirbution. Therefore, if we run the chain long enough the distribution of the samples have marginal distribution almost** π. Given a π, the main task therefore becomes how to construct a tractable Markov Chain (which we can easily simulate) with limiting distribution π.

9.1 Law of Large Numbers

One of the fundamental theorems of probability is the law of large numbers. Here is a statement of the strong law of large numbers.

Theorem 9.1 (Strong Law of Large Numbers for IID Sequences)**.** *Let $X_1, \ldots, X_n$ be i.i.d with common mean $\mu < \infty$. Then*

$$P(\lim_{n\to\infty} \frac{Y_1 + \cdots + Y_n}{n} = \mu) = 1$$

It turns out that the Law of Large Numbers can be extended to Markov Chains as well where the i.i.d assumption is not valid.

Theorem 9.2 (Strong Law of Large Numbers for Markov Chains)**.** *Let $X_0, X_1, \ldots, X_n$ be an irreducible, aperiodic, positive recurrent Markov Chain with stationary distribution π. Let r be a bounded, real valued function $\mathcal{S} \to \mathbb{R}$,*

$$P\big(\lim_{n\to\infty} \frac{r(X_1) + \cdots + r(X_n)}{n} = \mathbb{E}r(X)\big) = 1$$

where $\mathbb{E}r(X) = \sum_{j\in\mathcal{S}} r(j)\pi(j)$.

The Strong Law of Large Numbers for Markov Chains implies the corresponding Weak Law of Large Numbers.

Corollary 9.3. *Let $X_0, X_1, \ldots, X_n$ be an irreducible, aperiodic, positive recurrent Markov Chain with stationary distribution π. Let r be a bounded, real valued function. Then for any $\epsilon > 0$,*

$$P\big(|\frac{r(X_1) + \cdots + r(X_n)}{n} - \mathbb{E}r(X)| > \epsilon\big) \to 0$$

where $\mathbb{E}r(X) = \sum_{j\in\mathcal{S}} r(j)\pi(j)$.

Another implication of Strong Law of Large Numbers for Markov Chains is something we have already shown before. Take r to be the indicator function for any given state $j \in \mathcal{S}$. Then we have seen that the Fundamental Theorem says

$$\lim_{n\to\infty} \mathbb{E}\frac{r(X_1) + \cdots + r(X_n)}{n} = \lim_{n\to\infty} \mathbb{E}(\text{Proportion of Visits to State} j) = \pi(j).$$

This fact is actually implied by the above weak law in Corollary (9.3) which is in turn implied by the strong law (9.2).

The operational implication of the strong law (9.2) for us is that if we want to estimate $\mathbb{E}r(X)$ for some bounded function r and $X \sim \pi$ then we can run the MC for a long time and estimate by the sample mean of the MC. For example, let $A \subset \mathcal{S}$ be a subset and $\pi(A) = \sum_{j\in A} \pi(j)$. If we want to estimate $\pi(A)$, a natural estimate would be the proportion of times the MC visits the states in A out of n times where n is large.

9.2 Binary Sequences with No Consecutive 1's

Let us consider this example which illustrates the idea of MCMC and how it can be used to simulate from distributions that are otherwise hard to sample from. Define $B_n = \{0,1\}^d$ be the set of d length binary sequences. Let π_{B_d} be the uniform distribution on B_n. How to simulate from π_{B_d}? Let $X = (X_1, \ldots, X_d)$ be a random binary vector with distribution π_{B_d}. It is easy to see that $X_1, \ldots, X_d$ are i.i.d with distribution *Bernoulli*$(1/2)$. Therefore, we can just simulate d independent *Bernoulli*$(1/2)$ random variables which we know how to do.

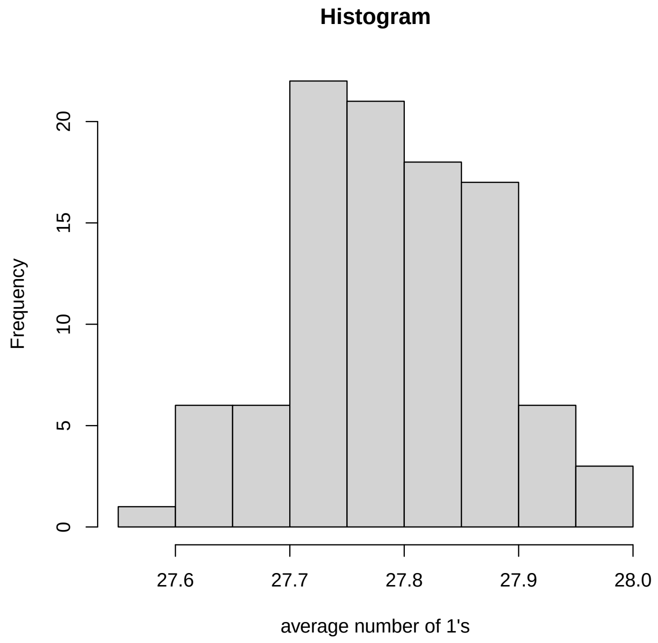
Now consider the set $G_d \subset B_d$ of binary d length sequences with no consecutive ones. Let π_{G_d} be the uniform distribution on G_d. **How do we simulate a uniformly random sequence from G_d or equivalently sample from π_{G_d}?** Another related question is **what is the expected number of 1's if all sequences in G_d are equally likely?** The answer of course is $\mu = \sum_{k=0}^{n} k\pi_k$ where π_k is the probability that a random sequence in G_d has exactly k 1's. In this case, it turns out that π_k can still be calculated by a combinatorial argument but it is not obvious. Of course, π_k can be estimated by Theorem 9.2 if we can sample from π_{G_d}. Let us see how we can do this.

Here is one way of sampling from π_{G_d} called the rejection sampling method which does not use Markov Chains. Sample a random sequence in B_n which is easy to do. If the sequence is in G_d accept the sample, otherwise reject the sample and try again. This approach gives what we want, i.e, the distribution of an accepted sample is π_{G_d}. (Show this!) However, a problem with this method would be that if d is not too small then you would have to reject a lot of samples before you accept one. When $d = 100$, it is true that $|B_d| = 2^{100} \sim 10^{30}$ and $|G_d| \sim 10^{21}$. Therefore, this algorithm would roughly accept once every 10^9 times which is pretty slow.

Let us now use the MCMC idea. We want to construct a MC whose limiting distribution is π_{G_d}. Let us define the state space to be G_d. Let us define the transition probabilities as follows. **Given any sequence $(x_1, \ldots, x_d)$ pick a coordinate at random (w.p $1/d$ each). If this coordinate is 1 then flip it to 0. If the coordinate is 0, then flip it to 1 if this results in a sequence in G_d, otherwise do not flip it.**

Observe the following facts about this MC.

1. The chain is irreducible. This is true because one can go from the all 0 sequence to any sequence in G_d and vice versa.
2. This chain is aperiodic because the period of the sequence $(1, 0, \ldots, 0)$ is 1.
3. For any $i \neq j \in G_d$ either we have $P_{ij} = P_{ji} = \frac{1}{n}$ or $P_{ij} = P_{ji} = 0$. Thus, the transition matrix P is symmetric.
4. The above three facts imply that the uniform distribution π_{G_d} is the stationary and limiting distribution of this chain.



We can now run the MC for a large enough n to simulate from π_{G_d}. The question is how large n should we take? For this, let's do a simulation. We want to estimate the expected number of 1's when we simulate from π_{G_d}. The true answer in this case is known and it is 27.7921 when $n = 100$. Now, let's simulate the above MC $10,000$ times starting from the all 0 sequence as the initial state to get a MCMC estimate. Let's repeat this experiment 100 times. We get the mean of these 100 estimates as 27.78949 and standard deviation of these 100 estimates as 0.08392706. Our answer is very close to the actual answer and running the MC for $10,000$ steps once just takes about 2 seconds! The figure above shows the histogram of these 100 estimates.

We now state the benefits of such a MCMC algorithm.

1. The state space of the MC is huge. The cardinality is about 10^{21} when $d = 100$. However, the number of steps required for the MC to get sufficiently close to the limiting distribution seems to be a small fraction of that. In this case, $10,000$ steps seems good enough.

2. The uniform distribution π_{G_d} assigns probability $1/c$ to each sequence in G_d where $c = |G_d|$. The actual value of c **is not needed** to run this algorithm. In this particular problem, it is possible by a clever argument to find $|G_d|$. However, one can imagine similar problems where it would be even harder to calculate the cardinality of the state space. For example, one can consider the set of all $d \times d$ matrices with every entry 1 surrounded by 0 from all sides and sampling from the uniform distribution on this set. It is relatively simple to construct a MC with the required limiting distribution

but finding this c is very difficult.

3. The algorithm is intuitive and easily and efficiently coded up.

9.3 Metropolis Hastings Algorithm

The Metropolis Hastings algorithm is a famous MCMC method for simulation due to Metropolis et al. (1953). W.K. Hastings extended the scope of the algorithm in 1970. Let's first describe the idea of this general algorithm. Suppose we have a finite state space $\mathcal{S}$ and a target distribution π on $\mathcal{S}$ which is positive everywhere. We want to sample from π.

We will start with any irreducible Markov chain on the state space $\mathcal{S}$ and then modify it into a new Markov chain that has the desired stationary distribution. This modification consists of introducing some selectiveness in the original chain: moves are proposed according to the original chain, but the proposal may or may not be accepted. For example, suppose the original chain is at a state called Boston and is about to transition to San Francisco. Then for the new chain, we either accept the proposal and go to San Francisco, or we turn down the proposal and remain in Boston. With a careful choice of the probability of accepting the proposal, this simple modification will guarantee that the new chain has the desired stationary distribution.

Let T be the transition matrix for any irreducible MC on $\mathcal{S}$. This T chain will be used as a proposal chain so it is important that we are able to simulate from the transition matrix T. Our ultimate goal is to sample from π. It would have been great if π satisfied the local balance equations $\pi_i T_{ij} = \pi_j T_{ji}$ for all $i \neq j$. In this case, we would know that π is stationary and hence limiting (if T is also aperiodic). Then our job would be done. It may not always be easy to come up with T so that π satisfies local balance.

Can we modify T now so that we can ensure local balance? So consider two different states i, j. We know that local balance does not hold. So let's assume w.l.g

$$\pi_i T_{ij} > \pi_j T_{ji}.$$

The main idea now is to make the effective T_{ij} smaller so that equality holds above. For this, when we run the original MC with transition T and when we are at state i, if we select the next state j to go to (which happens with probability T_{ij}) we wont immediately go to j. Instead, we will flip a coin with probability of heads A_{ij}. If it lands heads then we will go to j otherwise we will stay put at i. This modifies the original MC and the new transition probability of going from i to j becomes $T_{ij}A_{ij}$. For the local balance to hold, we will need

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

This suggests defining $A_{ij} = \frac{\pi_j T_{ji}}{\pi_i T_{ij}}$. Let's now write this formally.

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

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

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

Lemma 9.4. *Let P denote the modified transition matrix of the Metropolis-Hastings algorithm. Then π satisfies local balance with respect to P.*

Proof. The proof is basically given in our discussions above but lets prove it again. Take any two distinct states $i, j \in \mathcal{S}$. We know $P_{ij} = A_{ij}T_{ij}$. We can now write
$$\pi_i P_{ij} = \pi_i A_{ij} T_{ij} = \pi_i T_{ij} \min\{1, \frac{\pi_j T_{ji}}{\pi_i T_{ij}}\} = \min\{\pi_i T_{ij}, \pi_j T_{ji}\}.$$
Since the R.H.S above is symmetric in i, j therefore $\pi_i P_{ij}$ has to equal $\pi_j P_{ji}$ and hence we are done. □

Therefore, π is stationary and if the MC with the new transition dynamics is ergodic then π is limiting. If we start out with an irreducible chain then the final chain is also irreducible. Moreover, it will be aperiodic because we have introduced some laziness in the chain, i.e, there is positive probability of staying put in some states.

We now make some remarks about the Metropolis-Hastings algorithm.

1. The Metropolis-Hastings algorithm is an extremely general way to construct a Markov chain with a desired stationary distribution. In the above formulation, both π and T were very general, and nothing was stipulated about their being related (aside from being on the same state space). In practice, however, the choice of the proposal distribution is extremely important since it can make an enormous difference in how fast the chain converges to its stationary distribution. How to choose a good proposal distribution is a complicated topic and will not be discussed here.

2. Notice that only ratios of the form $\frac{\pi_i}{\pi_j}$ are needed to implement Metropolis Hastings. Thus, π only needs to be specified up to proportionality. For instance, if π is uniform on a set of size c, then $\frac{\pi_i}{\pi_j} = 1$ and the acceptance probability becomes $A_{ij} = \min\{1, \frac{T_{ji}}{T_{ij}}\}$. We do not need to know c which could be very hard to calculate.

3. If the proposal chain is ergodic so is the resulting Metropolis Hastings chain.

4. The generated sequence $X_0, X_1, \ldots$ produce approximate samples from π. However, if the chain requires a long time to get close to stationarity, there may be initial bias. *Burn-in* refers to the practice of discarding the initial iterations and using $X_m, X_{m+1}, \ldots, X_n$ for some m. The strong law for MC still gives $\lim_{n\to} \frac{1}{n-m+1} r(X_m) + \cdots + r(X_n) = \sum_{j \in \mathcal{S}} r(j)\pi_j$.

5. A major question in running a Markov chain $X_0, X_1, \ldots$ for a Monte Carlo computation is how long to run it. In part, this is because it is usually hard to know how close the chains distribution at time n will be to the stationary distribution. Another issue is that $X_0, X_1, \ldots$ are correlated in general. Some chains tend to get stuck in certain regions of the state space, rather than exploring the whole space. If a chain can get stuck easily, then X_n may be highly positively correlated with X_{n+1}. The autocorrelation at lag k is the correlation between X_n and the value k steps later, X_{n+k}, in the limit as n grows. It is desirable for the autocorrelation at lag k to approach 0 rapidly as k increases. High autocorrelation generally means high variances for Monte Carlo approximations. Analysis of how long to run a chain and finding diagnostics for whether the chain has been run long enough are active research areas. Some general advice is to run your chains for a very large number of steps and to try chains from diverse starting points to see how stable the results are.

Power Law Distribution Example: Power law distributions are positive probability distributions of the form π_i proportional to i^a for some constant $\alpha > 0$. Unlike distributions with exponentially decaying tails (e.g, Poisson, geometric, exponential) power law distributions have fat tails and are thus used to model heavy tailed data. A random variable X supported on $\{1, 2, \ldots, M\}$ has the Power Law distribution with parameter $a > 0$ if its PMF is

$$P(X = k) = \frac{k^{-\alpha}}{\sum_{j=1}^{M} j^{-a}}.$$

We can use the Metropolis-Hastings algorithm, after coming up with a proposal distribution. There are many possible proposal distributions, but one simple choice is the random walk on $\{1, 2, \ldots, M\}$ with reflecting boundaries. From state $1 < i < M$ move to state $i+1$ or $i-1$ with probability 1/2. From state M, stay there or move to $M-1$ with probability 1/2 and similarly for state 1.

Let P be the transition matrix of this chain. The stationary distribution for P is uniform because P is a symmetric matrix. Let X_0 be any starting state, and generate a chain $X_0, X_1, \ldots$ as follows. If the chain is currently at state i, then:

1. Generate a proposal state j according to the proposal chain P.

2. Accept the proposal with probability $\min\{i^a/j^a, 1\}$. If the proposal is accepted, go to j; otherwise, stay at i.

This chain is easy to implement and a move requires very little computation. Note that the normalizing constant $\sum_{j=1}^{M} j^{-a}$ is not needed to run this chain. Also note that the chain allows left moves to happen with probability 1 but controls the number of right moves by staying put with positive probability. In this way, the chain adjusts itself to put more mass to the smaller states.

Simulation Exercise: Simulate this MC a million times and record the proportion of states. These are MCMC estimates for the probabilities of the states. Compare these MCMC estimates with the actual probabilities.

9.3.1 Examples/Applications of Metropolis Hastings Algorithm

Cryptography

Markov chains have recently been applied to code-breaking; this example will introduce one way in which this can be done. A substitution cipher is a permutation g of the letters from a to z, where a message is enciphered by replacing each letter α by $g(\alpha)$. For example, if g is the permutation given by

$$abcdefghijklmnopqrstuvwxyz$$

$$zyxwvutsrqponmlkjihgfedcba$$

where the second row lists the values $g(a), g(b), ..., g(z)$ then we would encipher the word statistics as hgzgrhgrxh. (We could also include capital letters, spaces, and punctuation marks if desired.) The state space is all $26! \sim 4 \times 10^{26}$ permutations of the letters a through z. This is an extremely large space: if we had to try decoding a text using each of these permutations, and we could handle one permutation per nanosecond, it would still take over 12 billion years to work through all the permutations. So a brute-force investigation that goes through each permutation one by one is infeasible; instead, we will look at random permutations.

Consider the Markov chain that picks two different random coordinates between 1 and 26 and swaps those entries of the 2nd row, e.g., if we pick 7 and 20, then

$$abcdefghijklmnopqrstuvwxyz$$

$$zyxwvutsrqponmlkjihgfedcba$$

becomes

$$abcdefghijklmnopqrstuvwxyz$$

$$zyxwvugsrqponmlkjihtfedcba$$

The probability of going from g to h in one step is 0 unless h can be obtained from g by swapping 2 entries of the second row. Assuming that h can be obtained in this way, the probability is $\frac{1}{\binom{26}{2}}$, since there are $\binom{26}{2}$ such swaps, all equally likely. This Markov chain is irreducible, since by performing enough swaps we can get from any permutation to any other permutation. (Imagine rearranging a deck of cards by swapping cards two at a time; it is possible to reorder the cards in any desired configuration by doing this enough times.) Note that $p(g, h) = p(h, g)$, where $p(g, h)$ is the transition probability of going from g to h. Since the transition matrix is symmetric, the stationary distribution is uniform over all 26! permutations of the letters a through z.

Remark 9.1. *The above MC is irreducible but not aperiodic. Infact, the period of this chain is* 2*. Therefore the uniform distribution over all* 26! *permutations is not the limiting distribution. Can you guess the behaviour of* P^n *here (*P *is the transition matrix of the random transpositions MC) when* n *is large?*

Suppose we have a system that assigns a positive score s(g) to each permutation g. Intuitively, this could be a measure of how likely it would be to get the observed enciphered text, given that g was the cipher used. For example, this score could be the probability of any word assuming the word is generated from a MC. We could first construct a transition matrix denoting the probabilities of a letter β followed by α. A practical way fo doing this is to estimate these transition probabilities by going through a large english text. We could also assign the probability of the initial letter. With these in hand, we can calculate the probability of a word. For a given cipher, we can then define its score to be the probbility of the word that arises when we use this cipher to decode. We can now create a probability distribution π on the set of all 26! ciphers or permutations where π is proportional to the score s. In other words,

$$\pi(g) = \frac{s(g)}{\sum_g s(g)}.$$

So we reduce the task of decoding to sampling from π. The idea is that a sample from π would likely have a higher score and hence return a good cipher. For this, we can use a Metropolis Hastings algorithm. We will run the random transposition walk as described before on the set of permutations as our proposal chain. Since the transition matrix for our proposal chain is symmetric our acceptance function simply becomes

$$A_{ij} = \min\{1, \frac{\pi(j)}{\pi(i)}\}.$$

Now we can run this MC. Every time we make a random transposition to generate a proposal cipher or permutation. We then calculate the ratio $\frac{\pi(j)}{\pi(i)}$ which is simply $\frac{s(j)}{s(i)}$ which can be calculated easily. We then generate a coin flip with heads probability A_{ij} and move on to the next cipher. Note that because the proposal chain is irreducible, this modified chain is also irreducible. On the other hand, even though the proposal chain is periodic, the modified chain is aperiodic because of laziness. Therefore, π is indeed the limiting distribution of this modified chain.

9.3.2 Continuous State Space

MCMC can also be used to simulate from π when π has a probability density function (pdf). We have not yet learnt about continous state space stochastic processes. Intuitively, for a continuous state space Markov Process a transition function replaces the transition matrix, where P_{ij} is now the value of a conditional density function given $X_0 = i$. The Metropolis Algorithm is basically the same as in the discrete case, except that the transition probabilities are replaced by transition denisities. We present an example without going too much into technicalities.

Suppose we want to generate a standard normal random variable using only a uniform random number generator. The target density function is $\pi(t) = \frac{exp(-t^2/2)}{\sqrt{2\pi}}$. For the proposal distribution, we choose the uniform distribution of length 2 centered at the current state. From state s, the proposal chain moves to t, where t is uniformly distributed on $(s-1, s+1)$. The conditional density $T(s,t) = 1/2$ if $|s-t| \leq 2$ and 0 otherwise. The acceptance function then becomes

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

Simulation Exercise: Simulate this MC a million times. Plot the histogram and compare it with the Normal pdf.

Remark 9.2. *There are methods to sample exactly from the standard normal distribution without using MCMC. For any continuous random variable X with CDF F, the random variable $F^{-1}(U)$ has the same distribution as X when $U \sim Unif(0,1)$. For the standard normal the function F^{-1} is not available in closed form. There is another method called the Box Muller Transform which is capable of generating standard normals from uniform random numbers. The basic idea is as follows. X, Y are two independent standard normal random variables if and only if (R, Θ) are independent, Θ follows $Unif(0, 2\pi)$ and R^2 follows a Chi Squared distribution with degrees of freedom 2 which is the same as the Exponential Distribution with mean 2. Here (R, Θ) are the polar coordinates corresponding to the cartesian coordinates (X, Y). Therefore, to sample two independent standard normals it is enough to sample R and Θ. Sampling $\Theta \sim Unif(0, 2\pi)$ is easy and sampling $R = \sqrt{R^2} \sim \sqrt{Exponential(2)}$ is easy by the inverse CDF method.*

9.4 Gibbs Sampling

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

First we will run through how the Gibbs sampler works in the bivariate case, where the desired stationary distribution is the joint PMF of discrete r.v.s X and Y . There are several forms of Gibbs samplers, depending on the order in which updates are done. We will introduce two major kinds of Gibbs sampler: systematic scan, in which the updates sweep through the components in a deterministic order, and random scan, in which a randomly chosen component is updated at each stage.

Systematic scan Gibbs sampler

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

1. Draw a value x_{n+1} from the conditional distribution of X given $Y = y_n$, and set $X_{n+1} = x_{n+1}$.

2. Draw a value y_{n+1} from the conditional distribution of Y given $X = x_{n+1}$, and set $Y_{n+1} = y_{n+1}$.

3. Repeating steps 1 and 2 over and over, the stationary distribution of the chain $(X_0, Y_0), (X_1, Y_1), (X_2, Y_2), \ldots$ is p.

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

$$P(X' = x, Y = y) = P(X' = x|Y = y)P(Y = y) = p(x|y)p(y) = p(x, y).$$

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

Random Scan Gibbs sampler

As above, let X and Y be discrete r.v.s with joint PMF $p(x, y)$. We wish to construct a two-dimensional Markov chain (X_n, Y_n) whose stationary distribution is p. Each move of the random scan Gibbs sampler picks a uniformly random component and updates it, according to the conditional distribution given the other component:

1. Choose which component to update, with equal probabilities.

2. If the X-component was chosen, draw a value x_{n+1} from the conditional distribution of X given $Y = y_n$, and set $X_{n+1} = x_{n+1}, Y_{n+1} = y_n$. Similarly, if the Y-component was chosen, draw a value y_{n+1} from the conditional distribution of Y given $X = x_n$, and set $X_{n+1} = x_n, Y_{n+1} = y_{n+1}$.

3. Repeating steps 1 and 2 over and over, the stationary distribution of the chain $(X_0, Y_0), (X_1, Y_1), (X_2, Y_2), \dots$ is p.

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

The Gibbs sampler is less flexible than the Metropolis-Hastings algorithm in the sense that we dont get to choose a proposal distribution; this also makes it simpler in the sense that we dont have to choose a proposal distribution. The flavors of Gibbs and Metropolis-Hastings are rather different, in that Gibbs emphasizes conditional distributions while Metropolis-Hastings emphasizes acceptance probabilities. But the algorithms are closely connected, as we show below.

Theorem 9.5 (Random scan Gibbs as Metropolis-Hastings)**.** *The random scan Gibbs sampler is a special case of the Metropolis-Hastings algorithm, in which the proposal is always accepted. In particular, it follows that the stationary distribution of the random scan Gibbs sampler is as desired.*

Proof. We will show this in two dimensions, but the proof is similar in any dimension. Let X and Y be discrete r.v.s whose joint PMF is the desired stationary distribution. Lets work out what the Metropolis-Hastings algorithm says to do, using the following proposal distribution: from (x, y), randomly update one coordinate by running one move of the random scan Gibbs sampler. To simplify notation, write

$$P(X = x, Y = y) = p(x, y), P(Y = y|X = x) = p(y|x), P(X = x|Y = y) = p(x|y).$$

Lets compute the Metropolis-Hastings acceptance probability for going from (x, y) to (x', y'). The states (x, y) and (x', y') must be equal in at least one component, since the proposal says to update only one component. Suppose that $x = x'$ (the case $y = y'$ can be handled symmetrically). Then the acceptance probability is

$$A_{(x,y),(x,y')} = \frac{p(x, y')T_{(x,y'),(x,y)}}{p(x, y)T_{(x,y),(x,y')}} = \frac{p(x, y')p(y|x)1/2}{p(x, y)p(y'|x)1/2} = \frac{p(x)p(y'|x)p(y|x)}{p(x)p(y|x)p(y'|x)} = 1.$$

Thus, this Metropolis-Hastings algorithm always accepts the proposal! So its just running the random scan Gibbs sampler without modifying it. □

Remark 9.3. *The above theorem shows that the random scan gibbs sampler proposal chain satisfies local balance w.r.t p. In general, if we happen to choose a proposal chain so that it satisfies local balance w.r.t the target distribution π then we can still think of this as an instance of Metropolis Hastings algorithm with acceptance probability* 1.

9.5 Examples/Applications of Gibbs Sampling

Lets study some concrete examples of Gibbs samplers.

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

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

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

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

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

   (e) Return to Step (b).

   Remark 9.4. *Recall that there is a simple exact method to sample standard Bivariate Normal with correlation ρ. First sample two i.i.d $Z_1, Z_2 \sim N(0, 1)$. Now let $X = Z_1$ and $Y = \rho Z_1 + \sqrt{1 - \rho^2} Z_2$. Why is this method valid?*

   (f) **Graph coloring**

   Let $G = (V, E)$ be a graph with n nodes. We have a set of k colors, e.g., if k = 7, the color set may be red, orange, yellow, green, blue, indigo, violet. A k-coloring of the network is an assignment of a color to each node, such that two nodes joined by an edge cannot be the same color. Graph coloring is an important topic in computer science, with wide-ranging applications. The figure above shows a 3 coloring of a graph.

   It is easy to see that a n coloring of the graph is possible. In general, a graph could be k colorable for k much less than n and there always exists a $k \leq n$ so that the graph is k colorable. For example, consider a world map with different countries. We can create a graph with V equal to the set of all countries and E consists of edges between any two countries which share a border. One would obviously like to color two countries differently if they share a border. **The question now is how many colors would one need?** One of the most

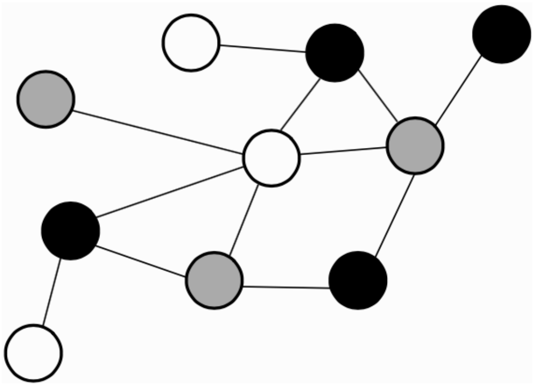


Figure 1: 3 coloring of a graph

famous theorems (the four color theorem) in mathematics says that four colors are enough.

Suppose that it is possible to k-color G. Form a Markov chain on the space of all k-colorings of G, with transitions as follows: starting with a k-coloring of G, pick a uniformly random node, figure out what the legal colors are for that node, and then repaint that node with a uniformly random legal color (note that this random color may be the same as the current color). This Markov chain is reversible, and its stationary distribution is the uniform distribution on the set of all k colorings of G. Let's see why this is true.

Let C be the set of all k-colorings of G, and let q_{ij} be the transition probability of going from i to j for any k-colorings i and j in C. We will show that $q_{ij} = q_{ji}$, which implies that the stationary distribution is uniform on C. For any k-coloring i and node v, let $L(i, v)$ be the number of legal colorings for node v, keeping the colors of all other nodes the same as they are in i. If k-colorings i and j differ at more than one node, then $q_{ij} = 0 = q_{ji}$. If $i = j$, then obviously $q_{ij} = q_{ji}$. If i and j differ at exactly one node v, then $L(i, v) = L(j, v)$, so

$$q_{ij} = \frac{1}{nL(i,v)} = \frac{1}{nL(j,v)} = q_{ji}.$$

So, the transition matrix is symmetric. This shows that the uniform distribution is a stationary distribution for this chain. Is the chain irreducible? It turns out that the answer is yes if k is not too small. So let's assume that k is also large enough so that this chain is irreducible. Clearly, the chain is aperiodic. Therefore, the uniform distribution on the set of all possible k colorings is also a limiting distribution.

How is this an example of Gibbs sampling? Think of each node in the graph as a discrete random variable that can take on k possible values. These nodes have a joint distribution, and the constraint that connected nodes cannot have the same color imposes a complicated dependence structure between nodes.

We would like to sample a random k-coloring of the entire graph; that is, we want to draw from the joint distribution of all the nodes. Since this is difficult, we instead condition on all but one node. If the joint distribution is to be uniform over all legal graphs, then the conditional distribution of one node given all the others is uniform over its legal colors. Thus, at each stage of the algorithm, we are drawing from the conditional distribution of one node given all the others: we are running a random scan Gibbs sampler!

2. **Ising Model**

The Ising Model was originally proposed in physics as a model for magnetism. It also arises in image processing. Consider a graph $G = (V, E)$ in which each node v is assigned a value/spin of ± 1. A configuration $\sigma \in \{\pm 1\}^{|V|}$ is an assignment of spins to each vertex and let $\Omega = \{\pm 1\}^{|V|}$ be the space of all such configurations.

Consider a friendship graph like the facebook graph. Let's denote each person's political preference by ± 1 (republican or democrat). It is likely that we would see a lot more agreements between friends than disagreements. We can also consider a black and white image on a $2d$ grid/lattice graph where the white or black squares tend to cluster together. What type of probability distribution would give more probability to such configurations? The Ising Model is a family of distributions that weighs configurations based on the number of agreements among neighbors.

Another way to motivate the Ising Model is that we can think of this as the discrete analog of the multivariate normal distribution. The multivariate normal distribution allows for correlations between the different components of a random vector which take real values. However, suppose I want to define a joint distribution or a random vector where each coordinate takes values ± 1 and yet different components are correlated. The graph G defines local neighborhoods of dependence.

The Ising Model distribution on Ω w.r.t to the graph G is given by

$$\pi(\sigma) = \frac{\exp(\beta \sum_{(i,j)\in E} \sigma_i \sigma_j)}{\sum_{\sigma \in \Omega} \exp(\beta \sum_{(i,j)\in E} \sigma_i \sigma_j)}.$$

The parameter $\beta = 0$ corresponds to the uniform distribution on Ω. If $\beta > 0$, this corresponds to weighting configurations with more agreements and it is the opposite for $\beta < 0$. Now suppose I want to sample from π. Sampling from π is hard because of the normalizing factor. However, the conditional distributions are simple to simulate. Let us denote the conditional distribution of $\sigma_i | \sigma_{-i}$ to be π_i. Then we can write

$$\pi_i(\sigma_i = \pm 1 | \sigma_{-i}) \propto \exp(\pm 1 \, \beta \sum_{j:(i,j)\in E} \sigma_j).$$

Therefore, we have

$$\pi_i(\sigma_i = 1|\sigma_{-i}) = \frac{\exp(\ \beta \sum_{j:(i,j)\in E} \sigma_j)}{\exp(\ \beta \sum_{j:(i,j)\in E} \sigma_j) + \exp(-\beta \sum_{j:(i,j)\in E} \sigma_j)}.$$

and

$$\pi_i(\sigma_i = -1|\sigma_{-i}) = \frac{\exp(-\beta \sum_{j:(i,j)\in E} \sigma_j)}{\exp(\ \beta \sum_{j:(i,j)\in E} \sigma_j) + \exp(-\beta \sum_{j:(i,j)\in E} \sigma_j)}.$$

So we just have to sum up $\sum_{j:(i,j)\in E} \sigma_j$ in order to run the Gibbs Sampler when we are at state i. Often, there are only a few neighbors of i so the above sum involves only a few terms.

Simulation Exercise: Run the Gibbs Sampler to simulate a $n \times n$ black and white image (viewed as a $2d$ lattice graph) for β positive, zero and negative.

Remark 9.5. *As with Metropolis-Hastings, Gibbs sampling also applies to continuous distributions, replacing conditional PMFs with conditional PDFs.*

10 A Linear Algebraic Condition for Convergence

All practical users of MCMC must confront the issue of how long to run the chain in order to reach convergence to the stationary distribution. We will now see that the second largest eigenvalue of the transition matrix P is a major player in this story.

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

We can check that

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

Since the chain is reversible, we obtain

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

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

1. A real symmetric matrix can be orthogonally diagonalized. This means that there exists an orthonormal real matrix S and a diagonal real matrix D such that $A = SDS^T$. This is called the spectral decomposition theorem in linear algebra. Moreover, the eigenvalues of A are the entries of the diagonal matrix D. Therefore, A has real eigenvalues.

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

3. (Lemmas 3.14 and 3.15 from the book)

   Since P is ergodic, there exists an integer $N > 0$ such that P^N has all entries strictly positive. For stochastic matrices with this property, there is a single largest eigenvalue 1 in absolute value and all the others are strictly less in absolute value. This means that the eigenvalues can be written in decreasing order

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

Now we can write

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

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

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

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

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

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

Remark 10.1. *In principle, for any MCMC method we can just compute its spectral gap to know how fast it will converge. In practice, this is often not possible as the state space is too large to compute eigenvalues of the transition matrix.*

11 Poisson Process

Poisson processes serve as a simple model for events occurring in time or space: in one dimension, cars passing by a highway checkpoint; in two dimensions, flowers in a meadow;

in three dimensions, stars in a region of the galaxy. Poisson processes are a primary building block for more complicated processes in time and space. A Poisson process is a special type of *counting process.* Given a stream of events that arrive at random times starting at $t = 0$, let N_t denote the number of arrivals that occur by time t, that is the number of events in $[0, t]$. For each $t \geq 0$, N_t is a random variable. The collection of random variables $(N_t)_{t\geq 0}$ is a continuous time, integer valued stochastic process, called a counting process. It is clear that N_t is non decreasing as a function of t.

Definition 11.1. *A counting process $(N_t)_{t\geq 0}$ is a collection of non negative integer valued random variables such that if $0 \leq s \leq t$, then $N_s \leq N_t$.*

So far we have only considered discrete time stochastic processes. A counting process forms an uncountable collection so it is a continuous time stochastic process.

There are several ways to characterize the Poisson process or any counting process for that matter. One can focus on

1. the number of events that occur in fixed intervals
2. when events occur, and the time between those events
3. the probabilistic behaviour of individual events on infinitesimal intervals.

This will lead to three equivalent definitions of a Poisson Process (PP).

Remark 11.1. *Throughout we will abuse notation and denote a counting process by $N(t)$ or N_t.*

11.1 Bernoulli Counting Process

Let us construct perhaps the simplest and most natural counting process on a time interval $[0, 1]$ (any interval $[0, a]$ can be handled similarly) by coin tossing. Pick a small number $\delta > 0$ which is our grid resolution. Now consider the grid $G_\delta = [\delta, 2\delta, \ldots, (N-1)\delta, N\delta]$ where $N = \frac{1}{\delta}$. For each interval $[i\delta, (i+1)\delta]$ in the grid, define an independent bernoulli random variable B_i with success probability λ/N representing whether an event has happened within the time interval $[i\delta, (i+1)\delta]$ or not. This creates N i.i.d Bernoulli random variables. It is clear that the total number of successes (or the total number of occurrences) follows a $Bin(N, \lambda/N)$ distribution. Hence the expected number of occurences is λ which can be thought of as the rate of occurrences. Note that λ does not depend on the resolution.

Now define the counting process

$$N(t) = \sum_{1\leq i\leq N:(i+1)\delta\leq t} X_i.$$

In words, the counting process $N(t)$ counts the number of successes/occurrences till time t. We call this process as the **Bernoulli process** on $[0, 1]$ with resolution δ. A natural question now is whether the Bernoulli process converges in some sense as the resolution $\delta \to 0$? *The answer is yes and the limiting process is the Poisson Process.*

Let us first take a step back and review where the Poisson distribution comes from. Why would the Poisson distribution with PMF $P(X = k) = \exp(\lambda)\frac{\lambda^k}{k!}$ arise in anyone's head? Let's look at the total number of successes/occurences at time 1 or the random variable $N(1)$. We know that the distribution of $N(1) \sim Bin(N, \lambda/N)$ where $N = \frac{1}{\delta}$. Now for each δ this is a separate distribution. If we take $\delta \to 0$ what happens to this distribution? Does this sequence of distributions converge to another distribution?

Lemma 11.2. *If λ_n is a sequence of positive numbers with $\lim_{n\to\infty} \lambda_n = \lambda$ then*

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

I leave the above as an exercise.

Theorem 11.3 (Law of Small Numbers)**.** *If $N \to \infty$ and $p \to 0$ in such a way that $Np \to \lambda$, then the Binomial distribution with parameters (N, p) converges to the Poisson λ distribution.*

Proof. Suppose $X \sim Bin(N, p)$ is a random variable. Fix any integer $0 \le k \le N$. Then

$$P(X = k) = \binom{N}{k} p^k (1 - p)^{n-k}.$$

We can write

$$\lim_{N\to\infty} \binom{N}{k} p^k (1-p)^{N-k} = \lim_{N\to\infty} \frac{1}{k!} N(N-1)\dots(N-k+1)p^k(1-p)^{N-k} =$$
$$\frac{1}{k!}\lim_{N\to\infty} N^k p^k (1-p)^{N-k} = \frac{1}{k!}\lim_{N\to\infty} (Np)^k (1 - \frac{\lambda}{N})^N = \frac{1}{k!}\lim_{N\to\infty} (Np)^k \lim_{N\to\infty} (1 - \frac{\lambda}{N})^N = \frac{1}{k!}\lambda^k \exp(-\lambda).$$

where in the last equality we have used Lemma 11.2. Lo and behold, we get the PMF of the Poisson distribution. □

Remark 11.2. *The Poisson distribution arises as a limit of Binomial random variables with $N \to \infty$ and $p \to 0$ in such a way that the expectation Np approaches a finite positive number λ. This viewpoint of the Poisson distribution would help us in understanding several properties of the Poisson distribution and the Poisson process. In the above proof, we showed that the convergence happens in the sense that that PMF of the sequence of Binomial distributions at any fixed k converges to the PMF of the Poisson distribution at k. Stronger notions of convergence can be shown. For example, in this setting one can also show that $TV(Bin(N, p), Poi(\lambda)) \to 0$ where TV is the total variation distance.*

11.2 Definition of Poisson Process

Definition 11.4 (Definition 1 of PP). *A PP with parameter λ is a counting process $(N_t)_{t\geq 0}$ with the following properties:*

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

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

Example: Joe receives text messages starting at 10 am at the rate of 10 texts per hour according to a Poisson process. Find the probability that he will receive exactly 18 texts by noon and 70 texts by 5 pm.

Solution: We need to compute

$$P(N(2) = 18, N(7) = 70) = P(N(2) = 18, N(7) - N(2) = 52) =$$
$$P(N(2) = 18)P(N(5) = 52) = P(Poi(20) = 18)P(Poi(50) = 52).$$

Proposition 11.5 (Translated PP). *Let N_t be a PP with parameter λ. For $t_0 > 0$, let*

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

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

Proof. Check that this satisfies the 4 requirements in Definition 1. □

Example: On election day, people arrive at a voting center according to a PP. On average, 100 voters arrive every hour. If 150 people arrive during the first hour, what is the probability that at most 350 people arrive during the third hour?

Solution: We need to compute

$$P(N_3 \leq 350 | N_1 = 150) = P(N_3 - N_1 \leq 200 | N_1 = 150) = P(N_3 - N_1 \leq 200) = P(Poi(200) \leq 200).$$

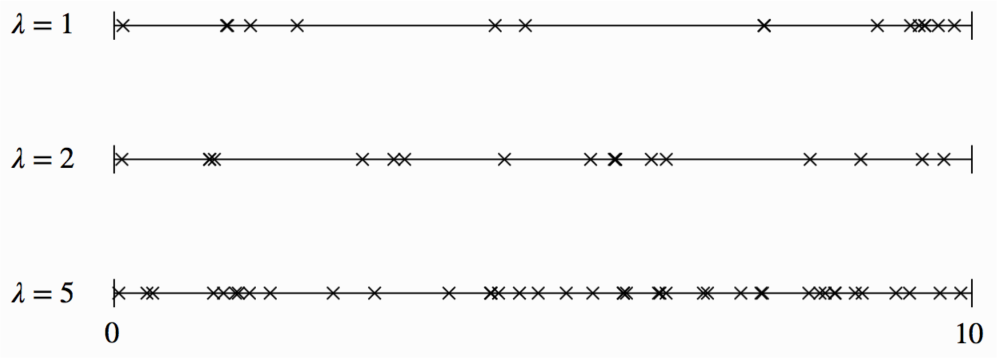


Figure 2: Simulated Poisson process in one dimension, for $\lambda = 1, 2, 5$.

11.3 Inter-Arrival Times

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

$$P(X > t) = P(N_t = 0) = \exp(-\lambda t).$$

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

Definition 11.6 (Definition 2 of PP)**.** *Let $X_1, X_2, \ldots$ be a sequence of i.i.d exponential random variables with parameter λ or mean $1/\lambda$. For $t > 0$, let*

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

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

We will see the equivalence of the two definitions later.

Remark 11.3. *The above definition says that a PP is a counting process for which interarrival times are i.i.d exponential random variables. Let $S_n = X_1 + \cdots + X_n$ for $n = 1, 2, \ldots$. We call $S_1, S_2, \ldots$ as arrival times of the process, where S_k is the kth arrival. Furthermore, $X_k = S_k - S_{k-1}$ is the kth interarrival time between the $(k-1)$th arrival and the kth arrival, with $S_0 = 0$.*

Remark 11.4. *The above definition leads to a direct method for simulating a Poisson Process in $[0, t]$.*

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

$$f_{S_n}(t) = \frac{\lambda^n t^{n-1} \exp(-\lambda t)}{(n-1)!}, \text{ for } t > 0.$$

Example: The time when goals are scored in hockey are modeled as a PP. For such a process, assume that the average time between goals is 15 minutes.

1. In a 60 minute game, find the probability that a fourth goal occurs in the last 5 minutes?

   Solution: $S_4 \sim Gamma(4, 1/15)$. So we just need to compute $P(55 \leq Gamma(4, 1/15) \leq 60)$.

2. Assume that at least 3 goals have been scored in the game. What is the mean time of the third goal? **Solution:** $S_3 \sim Gamma(3, 1/15)$. We need to compute

$$\mathbb{E}(S_3|S_3 \leq 60) = \frac{1}{P(S_3 \leq 60)} \int_0^{60} t f_{S_3}(t).$$

11.4 Memorylessness of the Exponential Random Variable

Definition 11.7 (Memorylessness)**.** *A positive random variable X possesses the memoryless property if for every $x \geq 0$ and $t > 0$,*

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

or equivalently,

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

Lemma 11.8. *If X is a continuous random variable then it satisfies memorylessness if and only if it is an Exponential random variable with some parameter $\lambda > 0$.*

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

Remark 11.5. *The only discrete random variable which has the memoryless property is the geometric distribution. This is not a surprise as the exponential distribution can be thought of as a continuous version of the Geometric distribution. Can you show how? (Exercise!)*

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

11.5 Conditioning on the number of arrivals in a Poisson Process

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

First, let us consider a Bernoulli process with a large N and a small p. Conditioning on $X_1 + \cdots + X_N = k$ what is the joint distribution of $(X_1, \ldots, X_N)$? It should be uniform over all binary vectors with k ones and $n-k$ zeroes. (In fact this holds irrespective of what N and p is.) The following theorem can be thought of as a limiting version of this fact.

Theorem 11.9. *Given that $N(1) = k$, the k points are uniformly distributed on $[0,1]$. That is, for any partition $J_1, \ldots, J_m$ of $[0,1]$ into non overlapping intervals,*

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

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

Remark 11.6. *The above theorem is saying that the distribution of the random vector $(N(J_1), \ldots, N(J_m))$ is distributed as Multinomial with number of trials n and probabilities $(|J_1|, \ldots, |J_m|)$.*

Proof. The random variables $N(J_i)$ are independent Poisson r.v.s with means $\lambda|J_i|$ by the definition of a Poisson process. Hence, for any nonnegative integers $k_1, \ldots, k_m$ that sum to k,

$$P(N(J_i) = k_i \forall i \in [1:m]) = \Pi_{i=1}^m (\lambda|J_i|)^{k_i} \frac{\exp(-\lambda|J_i|)}{k_i!} = \lambda^k \exp(-\lambda) \Pi_{i=1}^m \frac{|J_i|^{k_i}}{k_i!}$$

Dividing this by

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

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

Remark 11.7. *This suggests another way to simulate a Poisson point process of rate λ in $[0,a]$ for any integer $a \geq 1$. First, construct the counts $N[0,1], N[1,2], N[2,3], \ldots$ by i.i.d. sampling from the Poisson distribution with mean λ. Then, independently, throw down $N[i,i+1]$ points at random in the interval $[i,i+1]$ according to the uniform distribution.*

Corollary 11.10. *Let $S_1, S_2, \ldots$ be the occurrence/arrival times in a Poisson process $N(t)$ of rate λ. Then conditional on the event $N(1) = m$, the random variables $S_1, \ldots, S_m$ are* *distributed in the same manner as the* ***order statistics*** *of a sample of m i.i.d. uniform* $[0,1]$ *random variables.*

Example (Users on a website): Users visit a certain website according to a Poisson process with rate λ_1 users per minute, where an arrival at a certain time means that at that time someone starts browsing the site. After arriving at the site, each user browses the site for an $Expo(\lambda_2)$ amount of time (and then leaves), independently of other users. Suppose that at time 0, no one is using the site. Let N_t be the number of users who arrive in the interval $(0,t]$, and let C_t be the number of users who are currently browsing the site at time t.

1. Let X be the time of arrival and Y be the time of departure for a user who arrives at a Uniform time point in $[0,t]$ (viewed as points on the timeline). Find the joint PDF of X and Y.

2. Let p_t be the probability that a user who arrives at a Uniform time point in $(0,t]$ is still browsing the site at time t. Find p_t.

3. Find the distribution of C_t in terms of λ_1, λ_2 and t.

4. Littles law is a very general result, which says the following: *The long-run average number of customers in a stable system is the long-term average arrival rate multiplied by the average time a customer spends in the system.* Explain what happens to $\mathbb{E}(C_t)$ for t large, and how this can be interpreted in terms of Littles law.

Solution:

1. We have $X \sim Unif(0,t)$. Given $X = x$, Y is an $Expo(\lambda_2)$ shifted to start at x, i.e.,$(Y-x)|(X=x) \sim Expo(\lambda_2)$. So the joint PDF of X and Y is

$$f(x,y) = \frac{\lambda_2}{t}\exp(-\lambda_2(y-x))$$

for $0 < x < t$ and $x < y$.

2. With the previous notation we want to find $p_t = P(Y > t)$. This can be done by integrating the joint PDF over all (x,y) with $y > t$:

$$P(Y>t) = \frac{1}{t}\int_0^t\int_t^\infty \lambda_2\exp(-\lambda_2(y-x)) = \frac{1}{t}\int_0^t \exp(\lambda_2 x)\big(\int_t^\infty \lambda_2\exp(-\lambda_2 y)dy\big)dx =$$
$$\frac{\exp(-\lambda_2 t)}{t}\int_0^t \exp(\lambda_2 x)dx = \frac{\exp(-\lambda_2 t)}{\lambda_2 t}(\exp(\lambda_2 t - 1)) = \frac{1-\exp(-\lambda_2 t)}{\lambda_2 t}.$$

3. By Theorem 11.9 given $N(t) = n$ we have that the n arrival times in $(0, t]$ are i.i.d and uniform in that interval. Therefore, $C_t | N_t \sim Bin(N_t, p_t)$ and $N_t \sim Poi(\lambda_1 t)$. From here, we can conclude that $C_t \sim Poi(\lambda_1 p_t t)$ by the thinning property of a Poisson random variable to be discussed later.

4. As $t \to \infty$, $\mathbb{E}C_t \to \frac{\lambda_1}{\lambda_2}$. This agrees exactly with Littles law since it says that the long-run average number of users 'in the system' (currently browsing the site) is the rate at which users arrive λ_1 times the average time a user browses in a session $\frac{1}{\lambda_2}$.

11.6 Superposition

The second property of Poisson processes is superposition: if we take two independent Poisson processes and overlay them, we get another Poisson process. This follows from the fact that the sum of independent Poissons is Poisson.

Lemma 11.11. *If $Y_1, Y_2, \ldots, Y_n$ are independent Poisson random variables with means $\lambda_1, \lambda_2, \ldots, \lambda_n$ then*

$$\sum_{i=1}^{n} Y_i \sim Poi(\sum_{i=1}^{n} \lambda_i).$$

Proof. There are various ways to prove this, none of them especially hard. For instance, you can use probability generating functions (Exercise.) Alternatively, you can do a direct calculation of the probability mass function when $n = 2$, and then induct on n (Exercise.) But the clearest way to see that this theorem must be true is to use the Law of Small Numbers. Consider, for definiteness, the case $n = 2$. Consider independent Bernoulli trials X_i, with small success probability p. Let $N_1 = \lfloor \frac{\lambda_1}{p} \rfloor$ and $N_2 = \lfloor \frac{\lambda_2}{p} \rfloor$. Clearly we have $\sum_{i=1}^{N_1} X_i \sim Bin(N_1, p)$, $\sum_{i=N_1+1}^{N_2} X_i \sim Bin(N_2, p)$ and $\sum_{i=1}^{N} X_i \sim Bin(N, p)$ where $N = N_1 + N_2$. The Law of Small Numbers implies that when p is small and N_1, N_2 and N are correspondingly large, the three sums above have distributions which are close to Poisson, with means λ_1, λ_2 and λ respectively. □

Theorem 11.12 (Superposition Theorem)**.** *Let $\{N_1(t), t > 0\}$ and $\{N_2(t), t > 0\}$ be independent Poisson processes with rates λ_1 and λ_2 respectively. Then the combined process $N(t) = N_1(t) + N_2(t)$ is a Poisson process with rate $\lambda_1 + \lambda_2$.*

Proof. Lets verify the properties in the definition of Poisson process.

For all $t \geq 0$, $N_1(t) \sim Pois(\lambda_1 t)$ and $N_2(t) \sim Pois(\lambda_2 t)$, independently, so $N(t) \sim Pois(\lambda_1 t + \lambda_2 t)$ by Lemma 11.11 The same argument applies for any interval of length t, not just intervals of the form (0, t].

Arrivals in disjoint intervals are independent in the combined process because they are independent in the two individual processes, and the individual processes are independent of each other. □

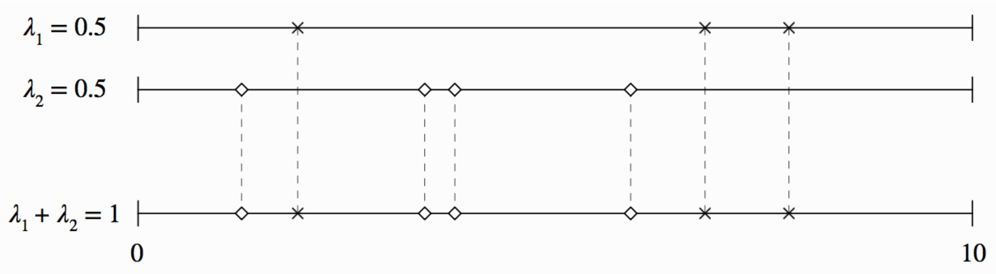


Figure 3: Superposition of two independent Poisson Process consisting of crosses and diamonds. Lets call the crosses type 1 events and the diamonds type 2 events. A natural question to ask is: what is the probability of observing a type 1 event before a type 2 event?

Remark 11.8. *The interarrival times in the superposed Poisson process are i.i.d* $Expo(\lambda_1 + \lambda_2)$*. On the other hand, the first arrival time is the minimum of the waiting times of the two processes. This shows that if* $X \sim Expo(\lambda_1)$ *and* $Y \sim Expo(\lambda_2)$ *independent of* X *then* $\min\{X, Y\} \sim Expo(\lambda_1 + \lambda_2)$.

Remark 11.9. *The most transparent way to generate from the superposition of two Poisson processes is exactly as one would expect: generate from the individual Poisson processes, then superpose them.*

Figure 12.2 depicts a superposed Poisson process consisting of crosses and diamonds. Lets call the crosses ”type 1 events” and the diamonds ”type-2 events”. A natural question to ask is: what is the probability of observing a type 1 event before a type 2 event?

Theorem 11.13 (Probability of type 1 event before type 2 event)**.** *If independent Poisson processes of rates* λ_1 *and* λ_1 *are superposed, the probability of a type* 1 *event before a type* 2 *event in the combined Poisson process is* $\frac{\lambda_1}{\lambda_1+\lambda_2}$.

Proof. Let T be the time until the first type 1 event and let V be the time until the first type 2 event. We seek $P(T < V)$. We know $T \sim Expo(\lambda_1)$ and $V \sim Expo(\lambda_2)$, so by applying scale transformations, $\tilde{T} = \lambda_1 T$ and $\tilde{V} = \lambda_2 V$ are i.i.d. $Expo(1)$. Letting $U = \frac{\tilde{T}}{\tilde{T}+\tilde{V}}$ we have

$$P(T \leq V) = P(\frac{\tilde{T}}{\lambda_1} \leq \frac{\tilde{V}}{\lambda_2}) = P(\frac{\tilde{T}}{\tilde{T}+\tilde{V}} \leq \frac{\lambda_1}{\lambda_2}\frac{\tilde{V}}{\tilde{T}+\tilde{V}}) = P(U \leq (1-U)\frac{\lambda_1}{\lambda_2}) = P(U \leq \frac{\lambda_1}{\lambda_1+\lambda_2}).$$

Now, since $\tilde{T}$ and $\tilde{V}$ are i.i.d $Expo(1)$ it follows that $U \sim Unif(0,1)$. Therefore we are done. □

Exercise: Let X, Y be i.i.d $Expo(\lambda)$. Show that $\frac{X}{X+Y} \sim Unif(0,1)$.

The above result applies to the first arrival in the combined Poisson process. After the first arrival, however, the same reasoning applies to the second arrival: by the memoryless property, the time to the next type 1 event is $Expo(\lambda_1)$ and the time to the next type 2 event is $Expo(\lambda_2)$, independent of the past. Therefore the second arrival is a type 1 arrival with probability $\frac{\lambda_1}{\lambda_1+\lambda_2}$, independent of the first arrival. Similarly, all of the arrival types can be viewed as i.i.d. coin tosses with probability $\frac{\lambda_1}{\lambda_1+\lambda_2}$ of landing Heads.

Corollary 11.14. *If independent Poisson processes of rates λ_1 and λ_1 are superposed, the arrival types are i.i.d and the probability of a type* 1 *arrival is* $\frac{\lambda_1}{\lambda_1+\lambda_2}$.

This yields an alternative path to simulate a superposition of two Poisson processes: we can first generate an $Expo(\lambda_1+\lambda_2)$ r.v. to decide when the next arrival occurs, and then independently flip a coin with probability $\frac{\lambda_1}{\lambda_1+\lambda_2}$ of heads to decide what kind of arrival it is.

Example (Competing risks): The lifetime of Freds refrigerator is $Y_1 \sim Expo(\lambda_1)$, and the lifetime of Freds dishwasher is $Y_2 \sim Expo(\lambda_2)$, independent of Y_1. Show that $\min\{Y_1, Y_2\}$ the time of the first appliance failure, is independent of $I(Y_1 < Y_2)$, the indicator that the refrigerator failed first.

Solution: This problem doesnt mention Poisson processes anywhere, but we will embed the r.v.s Y_1 and Y_2 into a Poisson process that we ourselves invent, in order to take advantage of the properties. of Poisson processes. So lets pretend there is an entire Poisson process of refrigerator failures with rate λ_1 and a Poisson process of dishwasher failures with rate λ_2. Then we can interpret Y_1 as the waiting time for the first arrival in the refrigerator process and Y_2 as the waiting time for the first arrival in the dishwasher process.

Furthermore, $\min\{Y_1, Y_2\}$ is the waiting time for the first arrival in the superposition of the two Poisson processes, and $I(Y_1 < Y_2)$ is the indicator of this arrival being a type 1 event. We know $\min\{Y_1, Y_2\} \sim Expo(\lambda_1 + \lambda_2)$ and $P(Y_1 < Y_2) = \frac{\lambda_1}{\lambda_1+\lambda_2}$. Now consider the conditional probability

$$P(Y_1 < Y_2 | \min\{Y_1, Y_2\} > t) = P(Y_1 < Y_2 | Y_1 > t, Y_2 > t).$$

Given $Y_1 > t$ and $Y_2 > t$ by memorylessness, the additional waiting times after t are also independent exponentials and hence the above conditional probability would again equal $\frac{\lambda_1}{\lambda_1+\lambda_2}$. This shows that the waiting times and event types in a superposed Poisson process can be generated completely independently! Knowing the refrigerator was the first appliance to fail does not provide any probabilistic information about the time of the first appliance failure.

11.7 Thinning

The third property of Poisson processes is thinning: if we take a Poisson process and, for each arrival, independently flip a coin to decide whether it is a type-1 event or type-2 event, we end up with two independent Poisson processes. This is the converse of superposition.

Lemma 11.15 (Thinning Property of Poisson)**.** *Suppose that $N \sim Poi(\lambda)$, and that $X_1, X_2, \dots$ are independent, identically distributed Bernoulli(p) random variables independent of N. Let $S_n = \sum_{i=1}^{n} X_i$. Then S_N has the Poisson distribution with mean λp. Similarly, $N - S_N$ has also the Poisson distribution with mean $\lambda(1-p)$. Moreover, S_N and $N - S_N$ are independent.*

Remark 11.10. *This is called the Thinning Property because, in effect, it says that if for each occurence counted in N you toss a p coin, and then record only those occurences for which the coin toss is a Head, then you still end up with a Poisson random variable.*

Proof. You can prove the first assertion directly, by evaluating $P(S_N = k)$ (exercise), or by using generating functions. We will use the Law of Small Numbers to see why this must be true. Let n be a large integer. Define $X_1, \dots, X_n$ i.i.d Bernoulli with success probability $\frac{\lambda}{n}$. Now define $Y_1, \dots, Y_n$ i.i.d Bernoulli with success probability p. Now consider the i.i.d sequence $Z_1, \dots, Z_n$ where $Z_i = X_i Y_i$.

Now, we know that the distribution of $N = X_1 + \cdots + X_n \sim Bin(n, \frac{\lambda}{n})$ which is very close to $Poi(\lambda)$. Now for those $X_i = 0$, multiplying by Y_i does not make a difference. However, for those $X_i = 1$, multiplying by Y_i effectively means we toss a p coin and we keep the value 1 if the coin lands heads. Therefore, the distribution of S_N should be very close to the distribution of $\sum_{i=1}^{n} Z_i$. Now the Z_i's are i.i.d Bernoulli with success probability $\frac{\lambda p}{n}$. Therefore, the distribution of $\sum_{i=1}^{n} Z_i \sim Bin(n, \frac{\lambda p}{n})$ should be close to $Poi(\lambda p)$. This proves the first assertion and the second assertion can be proved similarly.

Now we will see why S_N and $N - S_N$ should be independent. We have

$$P(S_N = m, N - S_N = f) = P(S_N = m, N - S_N = f, N = m + f) =$$
$$P(S_N = m, N - S_N = f | N = m + f)P(N = m + f) = P(S_N = m | N = m + f)P(N = m + f) =$$
$$\frac{(m+f)!}{m!f!} p^m (1-p)^f \exp(-\lambda) \frac{\lambda^{m+f}}{(m+f)!} = \big(\exp(-\lambda p)\frac{(\lambda p)^m}{m!}\big)\big(\exp(-\lambda q)\frac{(\lambda q)^m}{f!}\big)$$

where in the last line $q = 1 - p$. □

Theorem 11.16 (Thinning)**.** *Let $N(t)$ be a Poisson process with rate λ, and we classify each arrival in the process as a type 1 event with probability p and a type 2 event with probability $1 - p$, independently. Then the type 1 events form a Poisson process with rate λp, the type 2 events form a Poisson process with rate $\lambda(1-p)$, and these two processes are independent.*

Proof. Lets verify that the type 1 process, which well denote by $N_1(t)$, satisfies the properties in the definition of Poisson process.

Lets show that the number of arrivals for the type 1 process in an interval of length t is distributed $Poi(pt)$. For all $t \geq 0$, $N(t) \sim Poi(t)$ by definition, and $N_1(t)$ is the thinned version of $N(t)$. Hence, by lemma 11.15, $N_1(t) \sim Poi(pt)$. The same reasoning applies for any interval of length t, not just intervals of the form (0, t].

Arrivals in disjoint intervals are independent in the type 1 process because they are a subset of the arrivals in the full process, and we know the full process satisfies independence of disjoint intervals. Therefore, $N_1(t)$ is a Poisson process with rate p. The same reasoning applies for showing that the type 2 process, $N_2(t)$ is a Poisson process with rate $(1-p)$. The two processes are independent because for all $t > 0$, $N_2(t)$ is independent of $N_1(t)$. Actually, to show that the two processes are independent , one needs to show that $N_1(t), N_2(s)$ are independent for any $t \neq s$. This can now be seen by using the independent increments property. (How?)

□

11.8 Birthday Problem and Poisson Process Embedding

The classic birthday problem asks, "How many people must be in a room before the probability that some share a birthday, ignoring year and leap days, is at least 50 percent?". The probability that two people have the same birthday is 1 minus the probability that no one shares a birthday which is

$$p_k = 1 - \Pi_{i=1}^{k} \frac{366 - i}{365}.$$

One finds that $p_{22} = 0.476$ and $p_{23} = 0.507$. Thus, 23 people are needed.

Consider a sequential variant of the birthday problem where people enter a room one by one. Let K be the number of people in the room when for the first time two people share the same birthday? We want to calculate the mean and standard deviation of K.

Consider a continuous-time version of the previous question. People enter a room according to a Poisson process (N_t) with rate $\lambda = 1$. Each person is independently marked with one of 365 birthdays, where all birthdays are equally likely. The procedure creates 365 thinned Poisson processes, one for each birthday. Each of the 365 processes are independent, and their superposition gives the process of people entering the room.

Let $X_, X_2, \ldots$ be the interarrival sequence for the process of people entering the room. The X_i are i.i.d. exponential random variables with mean 1. Let T be the first time when two people in the room share the same birthday. Then, we can write $T = \sum_{i=1}^{K} X_i$. Therefore, we must have

$$\mathbb{E}T = \mathbb{E}(\mathbb{E}T|K) = \mathbb{E}K.$$

For each $k = 1, \ldots, 365$, let Z_k be the time when the second person marked with birthday k enters the room. Then, the first time two people in the room have the same birthday is $T = \min_{1 \leq k \leq 365} Z_k$. Each Z_k, being the arrival time of the second event of a Poisson process, has a Gamma distribution with parameters $n = 2$ and $\lambda = 1/365$ and are independent of each other.

We can now actually find the CDF of T.

$$P(T > t) = P(\min_{1 \leq k \leq 365} Z_k > t) = P(Z_1 > t, \dots, Z_{365} > t) = \Pi_{i=1}^{365} P(Z_i > t) = P(Z_1 > t)^{365}.$$

We can now find $\mathbb{E}T$ by the following formula for expectation.

$$\mathbb{E}T = \int_0^\infty P(T > t) dt.$$

A numerical software package finds that $\mathbb{E}T = 24.617$ and standard deviation of K is about 27.91.

11.9 Spatial Poisson Process

Poisson processes in multiple dimensions are defined analogously to the 1D Poisson process: we just replace the notion of length with the notion of area or volume. For concreteness, we will now define 2D Poisson processes, after which it should also be clear how to define Poisson processes in higher dimensions.

Definition 11.17. *(2D Poisson Process) Events in the 2D plane are considered a 2D Poisson process with intensity λ if*

1. *the number of events in a region A is distributed Pois(λ area(A));*
2. *the numbers of events in disjoint regions are independent of each other.*

As one might guess, conditioning, superposition, and thinning properties apply to 2D Poisson processes. Let N(A) be the number of events in a region A, and let $B \subset A$. Given $N(A) = n$, the conditional distribution of $N(B)$ is Binomial:

$$N(B)|N(A) = n \sim Bin(n, \frac{Area(A)}{Area(B)}).$$

Conditional on the total number of events in the larger region A, the probability of an event falling into a subregion is proportional to the area of the subregion; thus the locations of the events are conditionally Uniform, and we can generate a 2D Poisson process in A by first generating the number of events $N(A) \sim Pois(\lambda \, area(A))$ and then placing the events uniformly at random in A.

As in the 1D case, the superposition of independent 2D Poisson processes is a 2D Poisson process, and the intensities add. We can also thin a 2D Poisson process to get independent 2D Poisson processes.

Example: (Nearest star). Stars in a certain universe are distributed according to a 3D Poisson process with intensity λ. If you live in this universe, what is the distribution of the distance from you to the nearest star?

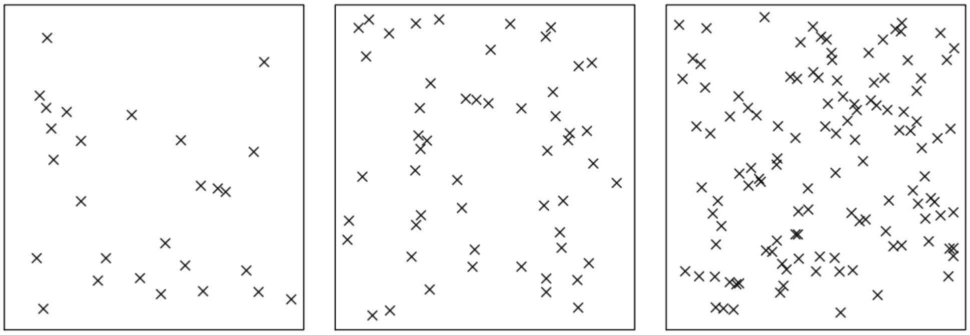


Figure 4: Simulated 2D Poisson process in the square $[0,5]^2$ for $\lambda = 1, 2, 5$.

Solution: In a 3D Poisson process with intensity λ, the number of events in a region of space V is Poisson with mean $\lambda\, volume(V)$. Let R be the distance from you to the nearest star. The key observation is that in order for the event $R > r$ to occur, there must be no stars within a sphere of radius r around you; in fact, these two events are equivalent. Let N_r be the number of events within radius r of you, so $N_r \sim Pois(\frac{4}{3}\pi r^3)$. Then $R > r$ is the same event as $N_r = 0$ so

$$P(R > r) = P(N_r = 0) = \exp(-\lambda\, \frac{4}{3}\pi r^3).$$

This specifies the CDF and hence the distribution of R. The distribution of R is an example of a Weibull distribution, which generalizes the Exponential.

11.10 Non Homogenous Poisson Process

Arrivals may be more or less likely at certain times. This is not captured by the Poisson Process model. To allow this, we can let the rate parameter λ vary over time.

Definition 11.18. *A counting process N_t is a Non Homogenous Poisson Process (NHPP) with intensity function $\lambda(t)$ if*

1. $N_0 = 0$.

2. *For any $t > 0$, N_t has Poisson distribution with mean*

$$\mathbb{E}N_t = \int_0^t \lambda(x)dx.$$

   In general, $N_{t+s} - N_s$ has Poisson distribution with mean $\int_s^{t+s} \lambda(x)dx$.

3. *For any $0 \leq q < r \leq s < t$, counts in disjoint intervals $N_r - N_q$ and $N_t - N_s$ are independent.*

Remark 11.11. *NHPP has independent increments but not stationary increments.*

Example: Let $\lambda(t) = A[1+\cos(\frac{2\pi t}{365})]$. Consider N_t to be a NHPP with intensity $\lambda(t)$.

1. Find $\mathbb{E}[N_{365/4}]$. We have

$$\int_0^{365/4} A[1+\cos(\frac{2\pi t}{365})]dt = \frac{365A}{4} + \frac{365}{2\pi}\sin(\frac{2\pi t}{365})|_0^{365/4} = \frac{365A}{4} + \frac{365A}{2\pi} \sim 149.34$$

2. Find $\mathbb{E}[N_{365/2} - N_{365/4}]$.

   We have

$$\int_0^{365/4} A[1+\cos(\frac{2\pi t}{365})]dt = \frac{365A}{4} + \frac{365}{2\pi}\sin(\frac{2\pi t}{365})|_{365/4}^{365/2} = \frac{365A}{4} - \frac{365A}{2\pi} \sim 33.16$$

Remark 11.12. *Note that $N_{365/4}$ and $N_{365/2} - N_{365/4}$ have different distributions.*

Remark 11.13. *For NHPP, the assumption of uniform arrivals on $[0,t]$ given $N_t = n$ no longer holds. There will be higher probability where λ is larger.*

11.11 A Paradox

Buses arrive at a bus stop according to a Poisson process. The time between buses, on average, is 10 minutes. Lisa gets to the bus stop at time t. How long can she expect to wait for a bus?

Here are two possible answers:

1. By memorylessness, the time until the next bus is exponentially distributed with mean 10 minutes. Lisa will wait, on average, 10 minutes.

2. Lisa arrives at some time between two consecutive buses. The expected time between consecutive buses is 10 minutes. By symmetry, her expected waiting time should be half that, or 5 minutes.

Paradoxically, both answers have some truth to them! On the one hand, the time until the next bus will be shown to have an exponential distribution with mean 10 minutes. But the backwards time to the previous bus is almost exponential as well, with mean close to 10 minutes. Thus, the time when Lisa arrives at the bus stop is a point in an interval whose length is about 20 minutes. And the argument in (ii) essentially holds. By symmetry, her expected waiting time should be half that, or 10 minutes. The surprising result is that the interarrival time of the buses before and after Lisa's arrival is about 20 minutes. And yet the expected interarrival time for buses is 10 minutes!

To explain the paradox, consider the process of bus arrivals. The rate of one arrival per 10 minutes is an average. The time between buses is random, and buses may arrive one right after the other, or there may be a long time between consecutive buses. When Lisa

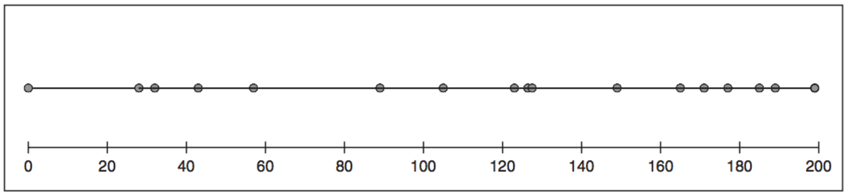


Figure 5: Pick a number from 0 to 200. Is your number in a short or long interval?

gets to the bus stop, she is more likely to get there during a longer interval between buses than a shorter interval.

To illustrate the idea, pick a random number between 1 and 200. Do it now before reading on. Now look at Figure 12.2, which gives arrival times for a Poisson process with parameter $\lambda = \frac{1}{10}$ on $[0, 200]$. Find your number. Is your number in a short interval (length less than 10) or a long interval (length greater than 10)? Most of you will find your number in a long interval.

This example illustrates the phenomenon of length-biased or size-biased sampling. For the bus waiting problem, the expected length of an interarrival time, which contains a fixed time t, is larger, about twice as large, than the average interval length between buses. Here is the calculation.

Fix $t > 0$. The time of the last bus before t is S_{N_t}. The time of the next bus after t is S_{N_t+1}. The expected length of the interval containing t is

$$\mathbb{E}\big(S_{N_t+1} - S_{N_t}\big) = \mathbb{E}S_{N_t+1} - \mathbb{E}S_{N_t}.$$

By memorylessness of the interarrival times, we must have

$$\mathbb{E}S_{N_t+1} = t + \frac{1}{\lambda}.$$

On the other hand, we can write

$$\mathbb{E}S_{N_t} = \mathbb{E}\,\mathbb{E}S_{N_t}|N_t = \mathbb{E}t(1 - \frac{1}{N_t+1}) = t - t\mathbb{E}\frac{1}{N_t+1}$$

where the second equality is because conditional on $N_t = k$, the kth arrival time has the same distribution as the maximum of k i.i.d. uniform random variables distributed on $(0, t)$.

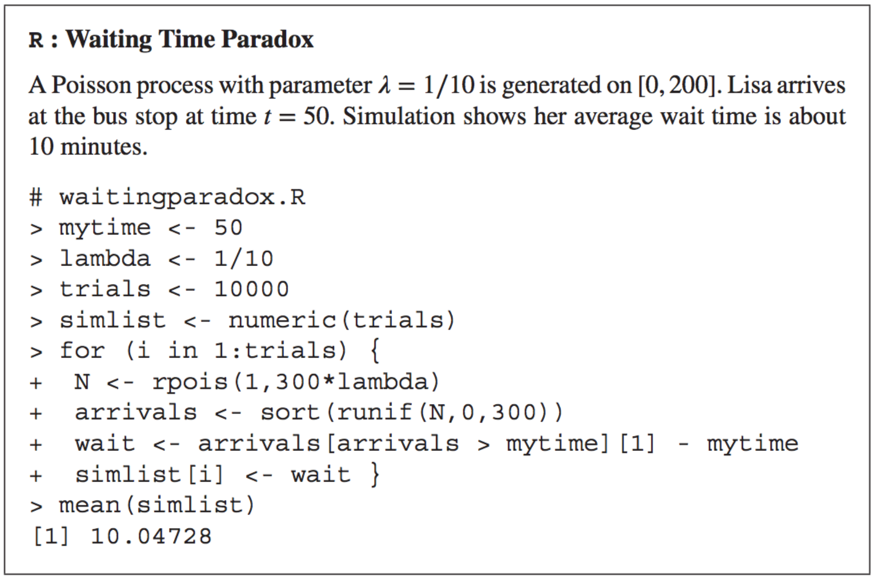
Exercise: Show that the expectation of the maximum of k i.i.d Uniform $(0, t)$ random variables is equal to $tk(k+1)$.

Now we find

$$\mathbb{E}\frac{1}{N_t+1} = \sum_{k=0}^{\infty} \frac{1}{k+1} \exp(-\lambda t)\frac{(\lambda t)^k}{k!} = \frac{\exp(-\lambda t)}{\lambda t} \sum_{k=0}^{\infty} \frac{(\lambda t)^{k+1}}{(k+1)!} = \frac{1 - \exp(-\lambda t)}{\lambda t}$$

Therefore, we obtain

$$\mathbb{E}S_{N_t} = t - \frac{1}{\lambda} + \frac{\exp(-\lambda t)}{\lambda}.$$



R : Waiting Time Paradox

A Poisson process with parameter $\lambda = 1/10$ is generated on $[0, 200]$. Lisa arrives at the bus stop at time $t = 50$. Simulation shows her average wait time is about 10 minutes.

```
# waitingparadox.R
> mytime <- 50
> lambda <- 1/10
> trials <- 10000
> simlist <- numeric(trials)
> for (i in 1:trials) {
+  N <- rpois(1,300*lambda)
+  arrivals <- sort(runif(N,0,300))
+  wait <- arrivals[arrivals > mytime][1] - mytime
+  simlist[i] <- wait }
> mean(simlist)
[1] 10.04728
```

Finally this means that

$$\mathbb{E}(S_{N_{t+1}} - S_{N_t}) = t + \frac{1}{\lambda} - t + \frac{1}{\lambda} - \frac{\exp(-\lambda t)}{\lambda} = \frac{2 - \exp(-\lambda t)}{\lambda} \sim \frac{2}{\lambda}$$

where the last approximation is correct for t not too small.

12 Brownian Motion

12.1 Some History

Brownian motion is one of the most famous and fundamental of stochastic processes. The formulation of this process was inspired by the physical phenomenon of Brownian motion, which is the irregular jiggling sort of movement exhibited by a small particle suspended in a fluid, named after the botanist Robert Brown who observed and studied it in 1827. A physical explanation of Brownian motion was given by Einstein, who analyzed Brownian motion as the cumulative effect of innumerable collisions of the suspended particle with the molecules of the fluid. Einstein's analysis provided historically important support for the atomic theory of matter, which was still a matter of controversy at the time, shortly after 1900. The mathematical theory of Brownian motion was given a firm foundation by Norbert Wiener in 1923; the mathematical model we will study is also known as the Wiener process.

12.2 Introduction

BM is a stochastic process that models random continuous motion. Let us start by writing down some physical assumptions about random continuous motion. Let X_t represent the position of a particle at time t. In this case, t takes on values in $\mathbb{R}_+$ (can be thought of as time) and X_t takes on values in the real line (or the plane or 3D space). This will be an example of a stochastic process with both continuous state space and continuous time.

For simplicity, let us start with $X_0 = 0$. The next assumption is that the motion is completely "random". Consider two times $s < t$. The motion after time s, that is $X_t - X_s$ is independent of s. We will need this assumption for any finite number of times: for any $s_1 < t_1 < s_2 < t_2 \cdots < s_n < t_n$ the random variables $X_{t_1} - X_{s_1}, \ldots, X_{t_n} - X_{s_n}$ are independent. Also the distribution of the random movements should not change with time. Hence we will assume that the distribution of $X_t - X_s$ depends only on $t - s$. For the time being, let us also assume that there is no drift to the process, i.e, $\mathbb{E}X_t = 0$.

The above assumptions are not sufficient to describe random continuous motion. If Y_t is the Poisson process and $X_t = Y_t - t$ then X_t satisfies these assumptions of stationary and independent increments. We will finally assume that the random motion process X_t, viewed as a function of time, is continuous.

It turns out that the above assumptions uniquely define the stochastic process at least up to a scaling constant. Suppose the process X_t satisfies all these assumptions. What is the distribution of X_t? Let us consider $t = 1$. For any $n \geq 1$, we can write

$$X_1 = [X_{1/n} - X_0] + [X_{2/n} - X_{1/n}] + \ldots [X_1 - X_{1-1/n}].$$

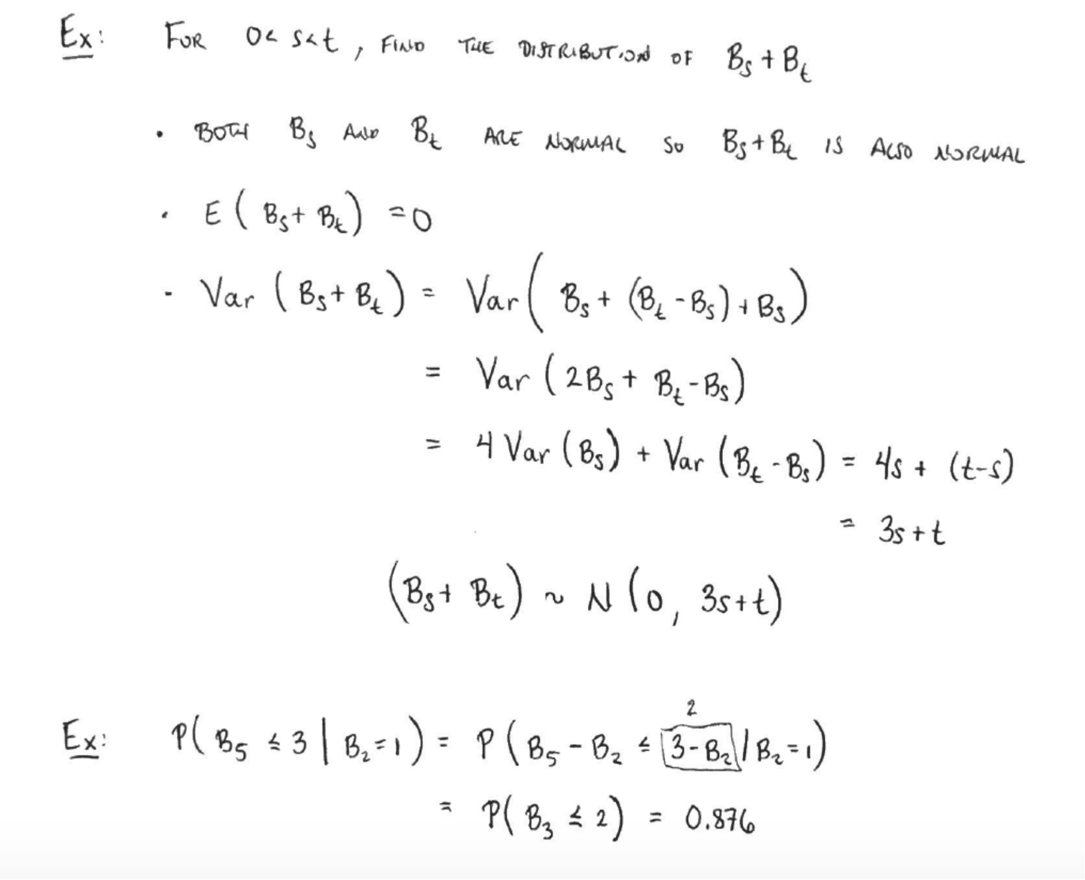
In words, X_1 can be written as the sum of n i.i.d random variables. Moreover, if n is large, each of the random variables is small. To be more precise, if we let

$$M_n = \max\{|X_{1/n} - X_0|, |X_{2/n} - X_{1/n}|, \ldots, |X_1 - X_{1-1/n}|\}$$

then as $n \to \infty$ the random variable $M_n \to 0$ because of continuity (why? uniform continuity anyone?). It is a theorem of probability theory (versions of Central Limit Theorem) that the only distribution that can be written as a sum of n i.i.d random variables such that the maximum of the random variables goes to 0 is a normal distribution. We can therefore conclude that the distribution of X_1 is normal.

Definition 12.1. *A Brownian Motion (BM) or a Weiner process with variance parameter σ^2 is a stochastic process X_t taking values in real numbers satisfying*

1. $X_0 = 0$.
2. *For any $s_1 < t_1 < s_2 < t_2 \cdots < s_n < t_n$ the random variables $X_{t_1} - X_{s_1}, \ldots, X_{t_n} - X_{s_n}$ are independent. This is the independent increments property.*
3. *For any $s < t$, the random variable $X_t - X_s$ has a normal distribution with mean 0 and variance $(t - s)\sigma^2$. This is the stationary normal increments property.*
4. *The paths are continuous, i.e, the function $t \to X_t$ is a continuous function of t.*



EX: FOR $0 < s < t$, FIND THE DISTRIBUTION OF $B_s + B_t$

- BOTH B_s AND B_t ARE NORMAL SO $B_s + B_t$ IS ALSO NORMAL
- $E(B_s + B_t) = 0$
- $Var(B_s + B_t) = Var(B_s + (B_t - B_s) + B_s)$

$$= Var(2B_s + B_t - B_s)$$

$$= 4\,Var(B_s) + Var(B_t - B_s) = 4s + (t-s)$$

$$= 3s + t$$

$$(B_s + B_t) \sim N(0, 3s+t)$$

EX: $P(B_5 \leq 3 \mid B_2 = 1) = P(B_5 - B_2 \leq \overset{2}{\boxed{3 - B_2}} \mid B_2 = 1)$

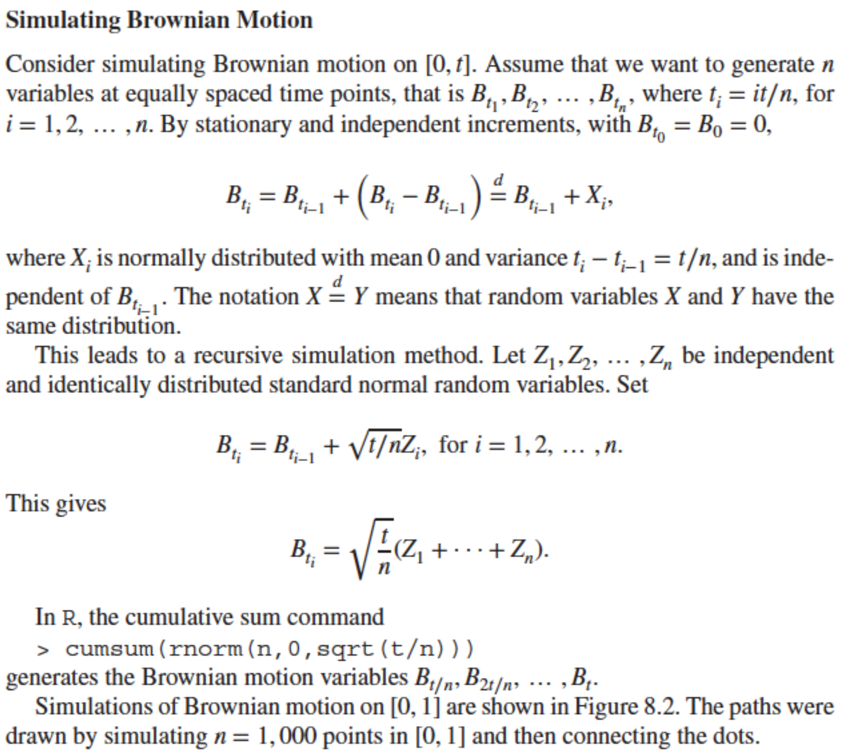
$$= P(B_3 \leq 2) = 0.876$$

While it is standard to include the fact that the increments are normally distributed in the definition, it is actually true that normality can be deduced from the physical assumptions.

Proposition 12.2. *If a stochastic process X has continuous paths and stationary, independent increments, then X is a Brownian motion.*

Thus, the assumptions of path continuity and stationary, independent increments is enough to give the normality of the increments for free.

Remark 12.1. *Standard Brownian Motion (SBM) is a BM with $\sigma^2 = 1$. We can also speak of a BM starting from x; this is a process satisfying conditions 2 to 4 in the above definition along with the initial condition $X_0 = x$. If X_t is a SBM then the process $Y_t = X_t + x$ is a BM starting at x. We can also speak of Brownian Motion with drift μ. If X_t is a SBM and $Y_t = X_t + \mu t$ then Y_t is a BM with drift μ. We can refer to a (μ, σ^2) BM as a Brownian motion where the mean and variance increases at rate μ and σ^2 per second, respectively. This situation here is analogous to that with normal distributions, where $Z \sim N(0,1)$ is called a standard normal random variable, and general normal random variables are obtained by multiplying a standard normal random variable by something and adding something.*



Simulating Brownian Motion

Consider simulating Brownian motion on $[0,t]$. Assume that we want to generate n variables at equally spaced time points, that is $B_{t_1}, B_{t_2}, \dots, B_{t_n}$, where $t_i = it/n$, for $i = 1, 2, \dots, n$. By stationary and independent increments, with $B_{t_0} = B_0 = 0$,

$$B_{t_i} = B_{t_{i-1}} + \left(B_{t_i} - B_{t_{i-1}}\right) \overset{d}{=} B_{t_{i-1}} + X_i,$$

where X_i is normally distributed with mean 0 and variance $t_i - t_{i-1} = t/n$, and is independent of $B_{t_{i-1}}$. The notation $X \overset{d}{=} Y$ means that random variables X and Y have the same distribution.

This leads to a recursive simulation method. Let $Z_1, Z_2, \dots, Z_n$ be independent and identically distributed standard normal random variables. Set

$$B_{t_i} = B_{t_{i-1}} + \sqrt{t/n} Z_i, \text{ for } i = 1, 2, \dots, n.$$

This gives

$$B_{t_i} = \sqrt{\frac{t}{n}}(Z_1 + \cdots + Z_n).$$

In R, the cumulative sum command

```
> cumsum(rnorm(n,0,sqrt(t/n)))
```

generates the Brownian motion variables $B_{t/n}, B_{2t/n}, \dots, B_t$.

Simulations of Brownian motion on [0, 1] are shown in Figure 8.2. The paths were drawn by simulating $n = 1,000$ points in [0, 1] and then connecting the dots.

12.3 What is Brownian Motion Really?

What is a Brownian Motion? We know that it is a stochastic process satisfying certain properties. What this means is that Brownian Motion (like other stochastic processes such as the Poisson Process) is really the name of a probability distribution. What is this a probability distribution over? Let's consider Standard Brownian Motion on the interval $[0,1]$. Then consider the sample space of all continuous functions $f : [0,1] \to \mathbb{R}$ such that $f(0) = 0$. Let's call this space of functions C_0. The Standard Brownian Motion on the interval $[0,1]$ is **a probability distribution which is supported over the space** C_0. One simulation or realization of Brownian Motion on $[0,1]$ gives me a random C_0 function.

12.4 Brownian Motion as a limit of Random Walk

Continuous-time, continuous-state Brownian motion is intimately related to discrete-time, discrete-state random walk. Brownian motion can be constructed from simple symmetric random walk by suitably scaling the values of the walk while simultaneously speeding up the steps of the walk.

Let $X_1, X_2, \dots$ be an i.i.d. sequence with each X_i taking values ± 1 with probability $1/2$ each. Set $S_0 = 0$ and for any integer $t > 0$, let $S_t = X_1 + \cdots + X_t$. Then, $S_0, S_1, S_2, \dots$ is a simple symmetric random walk with $\mathbb{E}(S_t) = 0$ and $Var(S_t) = t$ for $t = 0, 1, \dots$. As a sum of i.i.d. random variables, for large t, S_t is approximately normally distributed by the

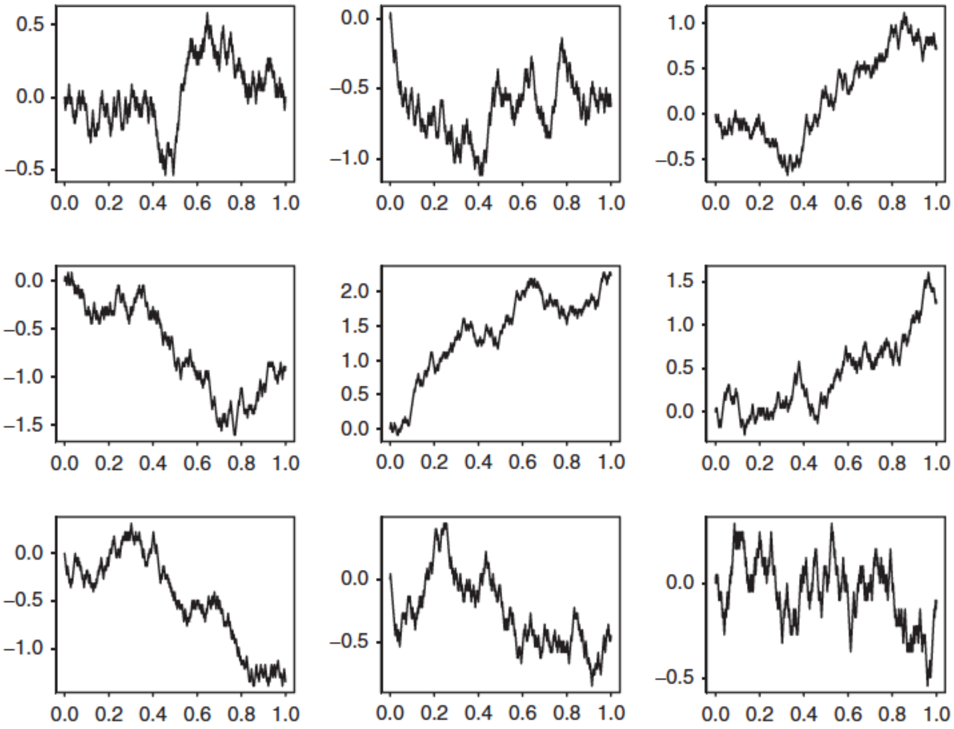


Figure 8.2 Sample paths of Brownian motion on [0, 1].

R: Simulating Brownian Motion

```
# bm.R
> n <- 1000
> t <- 1
> bm <- c(0, cumsum(rnorm(n,0,sqrt(t/n))))
> steps <- seq(0,t,length=n+1)
> plot(steps,bm,type="l")
```

More generally, to simulate $B_{t_1}, B_{t_2}, \ldots, B_{t_n}$, for time points $t_1 < t_2 < \cdots < t_n$, set

$$B_{t_i} = B_{t_{i-1}} + \sqrt{t_i - t_{i-1}} Z_i, \text{ for } i = 1, 2, \ldots, n,$$

with $t_0 = 0$.

central limit theorem.

It is clear that the random walk has stationary and independent increments. To obtain a continuous time process with continuous sample paths, we can connect the values by linear interpolation. Also, to model continuous time random motion it is reasonable to speed up the random walk. So let's fix the time interval $[0, 1]$. Let's take n random walk steps of size $\pm\delta$ at time gaps of $\frac{1}{n}$. This defines a piecewise linear continuous function or rather a distribution over continuous functions C_0. Now the question is what should we take δ to be? We should have the variance of the process at time 1 be 1. Therefore, we should take $\delta = 1/\sqrt{n}$.

Now we can imagine letting $n \to \infty$. For each n, we get a distribution over the space C_0. This sequence of distributions converges to a limiting distribution which is precisely the Brownian Motion. What is the meaning of a sequence of distributions on C_0 converging to another distribution? This is a topic for an advanced class.

12.5 Gaussian Process

Here is a very useful alternative characterization of standard Brownian motion. While describing this characterization we will also introduce two important definitions. First, W is a Gaussian process, which means that for all numbers n and times $t_1, \dots, t_n$ the random vector $(W(t_1), ..., W(t_n))$ has a joint normal distribution. An equivalent characterization of the Gaussianity of W is that the sum

$$a_1 W(t_1) + \cdots + a_n W(t_n)$$

is normally distributed for all all $t_1, \dots, t_n$ and all real numbers $a_1, \dots, a_n$.

Being a Gaussian process having mean 0, the joint distribution of all finite collections of random variables $W(t_1), ..., W(t_n)$ are determined by the covariance function

$$r(s, t) = Cov(W_s, W_t).$$

For standard Brownian motion, $Cov(W_s, W_t) = \min\{s, t\}$. To see this, suppose that $s \leq t$, and observe that

$$Cov(W_s, W_t) = Cov(W_s, W_s + W_t - W_s) = Var(W_s) = s$$

where we have used the independent increments property to say that $Cov(W_s, W_t - W_s) = 0$.

It is easy to see (Exercise!) that a process W is Gaussian with mean 0 and covariance function $r(s, t) = \min\{s, t\}$ if and only if properties 2 and 3 of Definition 12.1 hold for W. This shows the following fact.

Lemma 12.3. *A Gaussian process having continuous paths, mean* 0*, and covariance function* $r(s, t) = \min\{s, t\}$ *is a standard Brownian motion.*

This characterization of Brownian motion can be a convenient and powerful tool and can be used to show that other transformed processes are also BM.

12.6 Transformations and Properties

Lemma 12.4. *Let $(B_t)_{t\geq 0}$ be a standard Brownian motion. Then, each of the following transformations is a standard Brownian motion.*

1. *Rescaling. For any $a > 0$,*

$$a^{-1/2}B_{at}.$$

2. *Time Inversion.*

   The process $(X_t)_{t\geq 0}$ defined by $X_0 = 0$ and $X_t = tB_{1/t}$ for $t > 0$.

Proof. Let us prove the time inversion fact. To start we ask: is X_t a Gaussian process? Given $n, t_1, \ldots, t_n$, and $a_1, \ldots, a_n$ we have

$$a_1X(t_1) + \cdots + a_nX(t_n) = a_1t_1B(1/t_1) + \cdots + a_nt_nB(1/t_n)$$

which, being a linear combination of B evaluated at various times, has a normal distribution. Thus, the fact that B is a Gaussian process implies that X is also. Next, observe that the path continuity of X is also a simple consequence of the path continuity of B: if $t \to B(t)$ is continuous, then so is $t \to tB(1/t)$. (Well, this proves that with probability one $X(t)$ is continuous for all positive t. For $t = 0$, if you believe that $\lim_{s\to\infty} B(s)/s = 0$ with probability one, which is eminently believable by the SLLN, then making the substitution $s = 1/t$ gives $\lim_{t\to 0} tB(1/t) = 0$ with probability 1, so that X is also continuous at $t = 0$. Lets leave it at this for now.) The fact that $X(t)$ has mean 0 is trivial. Finally, to check the covariance function of X, let $s \leq t$ and observe that

$$Cov(X(s), X(t)) = Cov(sB(1/s), tB(1/t)) = stCov(B(1/s), B(1/t)) = st\min\{1/s, 1/t\} = st\frac{1}{t} = s.$$

Thus, X is a SBM.

The rescaling fact can be shown similarly. □

Brownian motion is continually restarting in a probabilistic sense. The next proposition is one way of formulating this idea mathematically.

Proposition 12.5. *Suppose that W is a standard Brownian motion, and let $c > 0$. Define $X(t) = W(c + t) - W(c)$. Then $\{X(t) : t \geq 0\}$ is a standard Brownian motion that is independent of $\{W(t) : 0 \leq t \leq c\}$.*

The proof is left as an exercise.

The Proposition says that, at each time c, the Brownian motion forgets its past and continues to wiggle on just as if it were a new, independent Brownian motion. That is, suppose that we know that $W(c) = w$, say. Look at the graph of the path of W; we are assuming the graph passes through the point (c, w). Now imagine drawing a new set of coordinate axes, translating the origin to the point (c,w). So the path now goes through the new origin. The above proposition says that if we look at the path past time c, relative to the new coordinate axes, we see the path of a new standard Brownian motion, independent of what happened before time c. Brownian motion is a Markov process: given the current state, future behavior does not depend on past behavior.

12.7 Nowhere Differentiability

The rescaling fact that $Y(t) = B(at)a^{-1/2}$ says that Brownian Motion looks the same if you zoom in on any small interval, say of length 10^{-12}; then after resizing by a factor 10^6 we would see is the realization of a SBM. This shows the *fractal* structure of Brownian Motion.

Another rather mind boggling property of BM is that with probability 1, a sample path of Brownian motion does not have a derivative at any time! Its easy to imagine functions like $f(t) = |t|$, that fail to be differentiable at isolated points. But try to imagine a function that everywhere fails to be differentiable, so that there is not even one time point at which the function has a well-defined slope. Such functions are not easy to imagine. In fact, before around the middle of the 19th century mathematicians generally believed that such functions did not exist, that is, they believed that every continuous function must be differentiable somewhere. Thus, it came as quite a shock around 1870 when Karl Weierstrass produced an example of a nowhere-differentiable function. Some in the mathematical establishment reacted negatively to this work, as if it represented an undesirable preoccupation with ugly, monstrous functions. It is interesting to reflect on the observation that, in a sense, the same sort of thing happened in mathematics much earlier in a different context with which we are all familiar. Pythagorus discovered that $\sqrt{2}$, which he knew to be a perfectly legitimate number, being the length of the hypotenuse of a right triangle having legs of length one is irrational. Such numbers were initially viewed with great distrust and embarrassment. They were to be shunned; notice how even the name irrational still carries a negative connotation. Apparently some Pythagoreans even tried to hide their regretable discovery. Anyway, now we know that in a sense almost all numbers are of this undesirable type, in the sense that the natural measures that we like to put on the real numbers (like Lebesgue measure (ordinary length)) place all of their mass on the set of irrational numbers and no mass on the set of rational numbers. Thus, the proof of existence of irrational numbers by producing an example of a particular irrational number was dwarfed by the realization that if one chooses a real number at random under the most natural probability measures, the result will be an irrational number with probability 1. The same sort of turnabout has occurred in connection with these horrible nowhere differentiable functions. Weierstrass constructed a particular function and showed that it was nowhere differentiable. The strange nature of this discovery was transformed in the same sense by Brownian motion, which puts probability 0 on nice functions and probability 1 on nowhere differentiable functions.

The nondifferentiability of the Brownian paths actually should not be very surprising, by the assumption of independent increments. Indeed, for each t and each $\delta > 0$, the increment $B(t+h) - B(t)$ is independent of the increment $B(t) - B(t-h)$, so that it would just be the wildest stroke of luck if the increments on both sides of t matched up well enough for W to be differentiable at t! Actually, if we just look at a one sided derivative, even that does not exist. In a tiny interval of length h around, the BM travels distance $B(t+h) - B(h) = \sqrt{h}Z$ which is of the order $\sqrt{h}$. Therefore, the right derivative at t would not be well defined.