Markov Chains on Countable State Space

1 Markov Chains Introduction

1. Consider a discrete time Markov chain \( \{X_i, i = 1, 2, \ldots \} \) that takes values on a countable (finite or infinite) set \( S = \{x_1, x_2, \ldots \} \), where \( x_i \) is the \( i \)-th state, the state of the stochastic process at time \( i \).

Examples of countable \( S \) are:

- \( \mathbb{Z}^d \), \( d \)-dimensional integers lattice
- set of all partitions of \( \{1, 2, \ldots, n\} \)
- set of all configurations of magnetic spins (-1 or 1) of a \( n \times n \) lattice grid (Ising model)
- set of possible moves of a king on a chess board
- set of possible route of visitation of \( n \) cities (traveling salesman problem)
- set of all length \( d \) vectors of 0’s and 1’s
- any finite set

It can be viewed as a stochastic process \( X_i \) indexed over time \( i = 1, 2, \ldots \). It is called a Markov chain if it satisfies the Markov property: the next state \( X_{i+1} \) depends only on the current state \( X_i \) (memoryless property), i.e.

\[
P(X_{i+1} = x_j | X_i = x_i, \ldots, X_0 = x_0) = P(X_{i+1} = x_j | X_i = x_i).
\]

The probability \( P_{ij} = P(X_{i+1} = x_j | X_i = x_i) \) is called the transition matrix. Any row of a transition matrix will sum to one, i.e., \( \sum_j P_{ij} = 1 \) for all \( i \). The initial distribution for \( X_0 \) determine the distribution for any \( n \)-th state.

2. Random walk. You play a coin tossing game. You win $1 if a head appears and lose $1 if a tail appears. Let \( X_i \) be your total gain after \( i \) tosses. Assume the coin is unbiased, then the transition matrix is given by

\[
P(X_{i+1} = x + 1 | X_i = x) = \frac{1}{2},
\]

\[
P(X_{i+1} = x - 1 | X_i = x) = \frac{1}{2}.
\]

We can model the coin tossing a simple random walk \( X_{i+1} = X_i + \epsilon_i \), with \( P(\epsilon_i = 1) = \frac{1}{2} \) and \( P(\epsilon_i = -1) = \frac{1}{2} \). The expected pay off does not change over time, i.e. \( E X_{i+1} = E X_i \).

In this case, \( X_{i+1} = \sum_{j=0}^{n} \epsilon_j + X_0 \). Then \( P(X_{n+1} = x | X_0 = 0) = P(\sum_{j=0}^{n} \epsilon_j = x) \). Please compute this probability.
3. Chapman-Kolmogorov equation. Let $P^n_{xy}$ be the transition kernel from $x$ to $y$ in $n$-steps, i.e. $P^n_{xy} = P(X_{i+n} = y|X_i = x)$. Then we have

$$P^m_{xy} = \sum_z P^m_{xz} P^n_{zy}.$$ 

The the probability of reaching $y$ from $x$ in $n$-steps is the sum of all probabilities going from $x$ to $y$ through an intermediate point $z$.

Let $P^n = (P^n_{ij})$ be a matrix.

Then in terms of matrix, Chapman-Kolmogorov equation can be trivially written as $P^{m+n} = P^m P^n$.

4. Example. A rat became insane and moves back and forth between position 1 and 2. Let $X_i$ be the position of the rat at the $i$-th move. Suppose that the transition probability is given by

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

On a finite state space, a state $i$ is called recurrent if the Markov chain returns to $i$ with probability 1 in a finite number of steps otherwise the state is transient. The recurrent of a state is equivalent to a guarantee of a sure return. Is the state 1 and 2 recurrent?

$$\sum_{i=2}^{\infty} P(X_i = j|X_0 = j, X_1 \neq j, \ldots, X_{i-1} \neq j) = \frac{1}{2} + \frac{1}{2^2} + \cdots = 1$$

for $j = 1, 2$. So the state 1 and 2 are recurrent.

5. Let us find $P(X_n = 1|X_0 = 2)$. Note that this is $n$ step transition probability denoted by $P^n_{12}$. It is computed by the Chapman-Kolmogorov equation. For this we need to compute $P^n$. In general there are theorems to do this. We are glad to do this computationally for now and find out that

$$P = \begin{bmatrix} 0.5 & 0.5 \\ 1 & 0 \end{bmatrix},$$

$$P^2 = \begin{bmatrix} 0.75000 & 0.25000 \\ 0.50000 & 0.50000 \end{bmatrix},$$

$$P^5 = \begin{bmatrix} 0.65625 & 0.34375 \\ 0.68750 & 0.31250 \end{bmatrix},$$

$$P^{100} = \begin{bmatrix} 0.66667 & 0.33333 \\ 0.66667 & 0.33333 \end{bmatrix}.$$ 

So we can see that the transition probabilities converge: $\lim_{n \to \infty} P^n_{ij} = \pi_j$. $\pi_1 = \frac{2}{3}$ and $\pi_2 = \frac{1}{3}$.
6. Suppose we have a probability distribution \( \pi \) defined on a state space \( S = \{x_1, x_2, \ldots\} \) such that \( \pi_j = P(X = x_j) \). We have \( \sum_{j \in S} \pi_j = 1 \). Then \( \pi \) is a stationary distribution for the Markov chain with transition probability \( P \) if \( \pi' = \pi P \).

An interesting property on \( \pi_j \) is that it is the expected number of states a Markov chain should take to return to the original state \( j \).

In our crazy rat example, the rat will return to position 2 in average 3 steps if it was at the position 2 initially.

7. An increased stability for a Markov chain can be achieved if we let \( X_0 \sim \pi \), i.e. \( P(X_0 = j) = \pi_j \). We have \( X_1 \sim \pi'P = \pi' \). Similarly \( X_i \sim \pi \).

Note that \( \pi \) is the eigenvector corresponding to the eigenvalue 1 for the matrix \( P \).

Actually you don’t need to have \( X_0 \sim \pi \) to have a stable chain. If \( X_0 \sim \mu \) for any probability distribution, \( \mu'P^n \rightarrow \pi' \) as \( n \rightarrow \infty \).

8. In general, the Markov chain is completely specified in terms of the distribution of the initial state \( X_0 \sim \mu \) and the transition probability matrix \( P \).

9. This is the basis of MCMC. Given probability distribution \( \pi \), we want to estimate some functions of \( \pi \), e.g., \( \mathbb{E}_\pi g(X) \). The idea is to construct a Markov chain \( X_i \) and let the chain run a long time (burn in stage) so that it will converge to its stationary distribution \( \pi \). We then sample from an dependent sample from this stationary distribution and use it to estimate the integral function.

The difference between MCMC and a simple Monte-Carlo method is that MCMC sample are dependent and and simple Monte Carlo sample are independent.

2 Markov Chains Concepts

We want to know:

- Under what conditions will the chain will converge?
- What is the stationary distribution it converges to?
- We will talk about finite state space first, then extend it to infinite countable space

2.1 Markov Chains on Finite S

One way to check for convergence and stationarity is given by the following theorem:

**Peron-Frobenius Theorem** If there exists a positive integer \( n \) such that \( P^m_{ij} > 0, \forall i, j \) (i.e., all the elements of \( P^n \) are strictly positive), then the chain will converge to the stationary distribution \( \pi \).
Although the condition required in Peron-Frobenious theorem is easy to state, sometimes it is difficult to establish for a given transition matrix $P$. Therefore, we seek another way to characterize the convergence of a Markov chain with conditions that may be easier to check.

For finite state space, the transition matrix needs to satisfy only two properties for the Markov chain to converge. They are: **irreducibility** and **aperiodicity**.

**Irreducibility** implies that it is possible to visit from any state to any state in a finite number of steps. In other words, the chain is able to visit the entire $S$ from any starting point $X_0$.

**Aperiodicity** implies that the Markov chain does not cycle around in the states with a finite period. The problem with periodicity is limiting the chain to return to certain states at some constant time intervals.

Some definitions:

1. Two states are said to communicate if it is possible to go from one to another in a finite number of steps. In other words, $x_i$ and $x_j$ are said to communicate if there exists a positive integer such that $P^n_{ij} > 0$.

2. A Markov chain is said to be irreducible if all states communicate with each other for the corresponding transition matrix.

   For the above example, the Markov chain resulting from the first transition matrix will be irreducible while the chain resulting from the second matrix will be reducible into two clusters: one including states $x_1$ and $x_2$, and the other including the states $x_3$ and $x_4$.

3. For a state $x_i$ and a given transition matrix $P_{ij}$, define the period of $x_i$ as $d(x_i) = GCD\{n : P^n_{i,i} > 0\}$, where GCD implies the greatest common divisor.

4. If two states communicate, then it can proved that their periods are the same.

   Thus, in an irreducible Markov chain where all the states communicate, they all have the same period. This is called the period of an irreducible Markov chain.

5. An irreducible Markov chain is called aperiodic if its period is one.

6. (**Theorem**) For a irreducible and aperiodic Markov chain on a finite state space, it can be shown that the chain will converge to a stationary distribution.

   If we let the chain run for long time, then the chain can be view as a sample from its (unique) stationary probability distribution.

   These two conditions also imply: that all the entries of $P^n_{ij}$ are positive for some positive integer $n$. The requirement needed for the **Peron-Frobenius Theorem**.
2.2 Markov Chains on Infinite but countable \( S \)

1. In case of infinite but countable state space, the Markov chain convergence requires an additional concept — positive recurrence — to ensure that the chain has a unique stationary probability.

2. The state \( x_i \) is recurrent iff \( P \) (the chain starting from \( x_i \) returns to \( x_i \) infinitely often) \( = 1 \). A state is said to be transient if it is not recurrent.

3. It can be shown that a state is recurrent iff
\[
\sum_{n=0}^{\infty} P^n_{ii} = \infty
\]

4. Another way to look at recurrent/transient: A state \( i \) is said to be transient if, given that we start in state \( i \), there is a non-zero probability that we will never return back to \( i \). Formally, let the random variable \( T_i \) be the next return time to state \( i \)
\[
T_i = \min\{n : X_n = i | X_0 = i\}
\]

Then, state \( i \) is transient if \( T_i \) is not finite with some probability, i.e., \( P(T_i < \infty) < 1 \).
If a state \( i \) is not transient, i.e., \( P(T_i < \infty) = 1 \) (it has finite hitting time with probability 1), then it is said to be recurrent.

Although the hitting time is finite, it need not have a finite average. Let \( M_i \) be the expected (average) return time, state \( i \) is said to be positive recurrent if \( M_i \) is finite; otherwise, state \( i \) is null recurrent.

5. If two states \( x_i \) and \( x_j \) communicate, and \( x_i \) is (positive) recurrent, then \( x_j \) is also (positive) recurrent.

6. (Theorem) For a irreducible, aperiodic, and positive recurrent Markov chain on a countably infinite state space, it can be shown that the chain will converge to a stationary distribution. This chain is said to be ergodic.

3 Markov Chains Examples

1. Let \( P \)
\[
P = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} & 0 \\
\frac{1}{4} & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
0 & \frac{1}{3} & \frac{1}{3} & \frac{1}{3}
\end{bmatrix}
\]

The eigenvector corresponding to the eigenvalue 1 for the matrix \( P \) is the stationary distribution
2. Consider a Markov chain with the following transition probability matrix

\[ P = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 \\
0.5 & 0 & 0.5 & 0 & 0 \\
0 & 0.5 & 0 & 0.5 & 0 \\
0 & 0 & 0.5 & 0 & 0.5 \\
0 & 0 & 0 & 1 & 0
\end{bmatrix} \]

Calculate: \( P^n \)

if \( n \) is even:

\[ P^n = \begin{bmatrix}
0.25 & 0 & 0.50 & 0 & 0.25 \\
0 & 0.50 & 0 & 0.50 & 0 \\
0.25 & 0 & 0.50 & 0 & 0.25 \\
0 & 0.50 & 0 & 0.50 & 0 \\
0.25 & 0 & 0.50 & 0 & 0.25
\end{bmatrix} \]

and if \( n \) is odd:

\[ P^n = \begin{bmatrix}
0 & 0.50 & 0 & 0.50 & 0 \\
0.25 & 0 & 0.50 & 0 & 0.25 \\
0 & 0.50 & 0 & 0.50 & 0 \\
0.25 & 0 & 0.50 & 0 & 0.25 \\
0 & 0.50 & 0 & 0.50 & 0
\end{bmatrix} \]

The Markov chain is irreducible because there exists a \( n \) such that \( P^n_{ij} > 0 \) for all \( i, j \)

Periodic with period \( d = 2 \)

It means that starting in state 1, for example, \( P^n_{1,1} > 0 \) at times \( n = 2, 4, 6, 8, \ldots \) and the greatest common divisor of the values that \( n \) can take is 2.

The unique stationary distribution is

\[ \pi = \lim_{n \to \infty} \frac{1}{2}(\pi^{(n)} + \pi^{(n+1)}) \]

\[ = (0.125, 0.25, 0.25, 0.25, 0.125) \]

we can verify that the eigenvalues of \( P \) are \(-1, 0, 1, -1/\sqrt{2}, 1/\sqrt{2}\)

and there are \( d = 2 \) eigenvalues with absolute value 1
3. Let
\[ P = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0.5 & 0.5 & 0 & 0 & 0 \\
0 & 0.5 & 0 & 0.5 & 0 \\
0 & 0 & 0.5 & 0.5 & 0 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix} \]

this system has 5 different stationary distributions for large \( n \):

Calculate \( P^n \)

\[
\lim_{n \to \infty} P^n = \begin{bmatrix}
1 & 0 & 0 & 0 & 0 \\
0.75 & 0 & 0 & 0 & 0.25 \\
0.50 & 0 & 0 & 0 & 0.50 \\
0.25 & 0 & 0 & 0 & 0.75 \\
0 & 0 & 0 & 0 & 1
\end{bmatrix}
\]

The rows of \( P^n \) represent the five stationary distributions, and each of these satisfy \( \pi' = \pi P \)

4. Let
\[ P = \begin{bmatrix}
\frac{1}{2} & \frac{1}{2} & 0 & 0 & 0 \\
\frac{1}{6} & \frac{1}{6} & 0 & 0 & 0 \\
0 & 0 & \frac{1}{3} & \frac{1}{3} & 0 \\
0 & 0 & \frac{1}{4} & \frac{1}{4} & \frac{1}{4} \\
0 & 0 & 0 & \frac{1}{6} & \frac{1}{6}
\end{bmatrix} \]

\[
\lim_{n \to \infty} P^n = \begin{bmatrix}
0.25 & 0.75 & 0 & 0 & 0 \\
0.25 & 0.75 & 0 & 0 & 0 \\
0 & 0 & 0.182 & 0.364 & 0.455 \\
0 & 0 & 0.182 & 0.364 & 0.455 \\
0 & 0 & 0.182 & 0.364 & 0.455
\end{bmatrix}
\]

The chain splits into two sub-chain, each converges to a stationary distribution and one can not move from state spaces \{0, 1\} to states \{2, 3, 4\}. The resulting stationary distribution depends on the starting point \( X_0 \).

5. Ehrenfest Model
The general model will be easily extended. For now, consider two urns that, between them, contain four balls. At each step, one of the four balls is chosen at random and
moved from the urn that it is in into the other urn. We choose, as states, the number of balls in the first urn.

The transition matrix is

\[
P = \begin{bmatrix}
0 & 1 & 0 & 0 & 0 \\
\frac{1}{4} & 0 & \frac{3}{4} & 0 & 0 \\
0 & \frac{1}{2} & 0 & \frac{1}{2} & 0 \\
0 & 0 & \frac{3}{4} & 0 & \frac{1}{4} \\
0 & 0 & 0 & 1 & 0
\end{bmatrix}
\]

The stationary distribution is \( \pi = (.0625, .2500, .3750, .2500, .0625) \).

Meaning: we can interpret \( \pi_i \) as the proportion of times the process is in each of the states in the long run. For example, the proportion of times in state 0 is .0625 and the proportion of times in state 1 is .375.

Note that these numbers are the binomial distribution 1/16, 4/16, 6/16, 4/16, 1/16. We could have guessed this answer as follows: If we consider a particular ball, it simply moves randomly back and forth between the two urns. This suggests that the equilibrium state should be just as if we randomly distributed the four balls in the two urns. If we did this, the probability that there would be exactly \( j \) balls in one urn would be given by the binomial distribution \( \text{Binomial}(n, p) \) with \( n = 4 \) and \( p = .5 \).

Let us consider the Ehrenfest model for gas diffusion for the general case of \( 2n \) balls. Every second, one of the \( 2n \) balls is chosen at random and moved from the urn it was in to the other urn.

What is the transition matrix?

What is the stationary distribution \( \pi \)?

What is the mean recurrence time for state \( i \)?

6. **Stepping Stone Model**

This model is often used as an illustration in the study of genetics. In this model we have an \( n \times n \) array of squares, and each square is initially any one of \( k \) different colors. For each step, a square is chosen at random. This square then chooses one of its eight neighbors at random and assumes the color of that neighbor. To avoid boundary problems, we assume that if a square \( s \) is on the left-hand boundary, say, but not at a corner, it is adjacent to the square \( t \) on the right-hand boundary in the same row as \( s \), and \( s \) is also adjacent to the squares just above and below \( t \). A similar assumption is made about squares on the upper and lower boundaries. (These adjacencies are much easier to understand if one imagines making the array into a cylinder by gluing the top and bottom edge together, and then making the cylinder into a doughnut by gluing the two circular boundaries together.) With these adjacencies, each square in the array is adjacent to exactly eight other squares.
A state in this Markov chain is a configuration of the $k$ colors on each square of the $n \times n$ array. Thus, the Markov chain has a total of $k^{n^2}$ possible states.

What is the long term behavior of this chain?

What are the two absorbing states? What is the probability that eventually the chain will stop at one absorbing state?

What is the mean number of steps for the chain to stop eventually?

7. **Random Walk Model on Integers** Consider the random walk on the real line of integers

$$P(X_{i+1} = x + 1|X_i = x) = p$$
$$P(X_{i+1} = x - 1|X_i = x) = 1 - p$$

with $X_0 = 0$?

It can be shown that the chain is **null recurrent** when $p = .5$ (symmetric random walk), and **transient** when $p \neq .5$.

8. **Symmetric Random Walk on Integers on $k$-dimension**

For $k = 1$, the mean distance between start and finish points in $n$ steps is of the order of $\sqrt{n}$.

Every integer point will be visited infinitely often, but the expected return time is infinite (**null recurrent**).

This property has many names: the level-crossing problem, the recurrence problem, or the gambler’s ruin problem. The interpretation of the last name is as follows: if you are a gambler with a finite amount of money playing a fair game against a bank with an infinite amount of money, you will surely lose.

For $k = 2$, it can be shown that a symmetric random walk in two dimension is also **null recurrent**.

For $k \geq 3$, the symmetric random walk is **transient**! Too many rooms to "get lost" forever!

The trajectory of a random walk is the collection of integer points the chain visited, regardless of visit orders at the point. In one dimension, the trajectory is simply all points between the minimum and the maximum visits. In higher dimensions the set has interesting geometric properties. In fact, one gets a discrete fractal, that is a set which exhibits stochastic self-similarity on large scales, but on small scales one can observe "jugginess" resulting from the grid on which the walk is performed. (cf. A famous fractal is the Mandelbrot Set)

9. **Random Walk Model on Non-negative Integers** Consider the random walk on the real line of non-negative integers. With $X_0 = 0$, let the transition matrix of the chain be

$$P(X_{i+1} = x + 1|X_i = x) = p$$
\[ P(X_{i+1} = x - 1 | X_i = x) = 1 - p \]

\[ \forall x > 0. \text{ When } x = 0, \]

\[ P(X_{i+1} = 1 | X_i = 0) = p \]
\[ P(X_{i+1} = 0 | X_i = 0) = 1 - p \]

It can be shown that the chain is

- **positive recurrent** when \( p < .5 \)
- **null recurrent** when \( p = .5 \)
- **transient** when \( p > .5 \)


4 Reversible Markov chain

Consider a Markov chain with state space $S$ that converges to a stationary distribution $\pi$.

Let $x \in S$ denote the current state of the system.
Let $y \in S$ denote the current state at the next step.
Let $p(x, y)$ denote the probability of a transition from $x$ to $y$
Thus $p(y, x)$ is the probability of a transition from $y$ to $x$.
A Markov chain is said to be reversible if it satisfies the detail balance condition:

$$\pi(x)p(x, y) = \pi(y)p(y, x)$$

Note:

$$\pi(x)p(x, y) = P(X_n = x)P(X_{n+1} = y|X_n = x) = P(X_n = x, X_{n+1} = y), \forall x, y \in S. \quad (1)$$

So the reversibility condition can also be written as

$$P(X_n = x, X_{n+1} = y) = P(X_n = y, X_{n+1} = x), \forall x, y \in S. \quad (2)$$

As seen later, the reversibility condition will be important in deriving the appropriate transition kernels function for the Metropolis algorithm or other Markov chain.

5 Historical Remarks

Markov chains were introduced by Andrei Andreevich Markov (1856-1922) and were named in his honor. He was a talented undergraduate who received a gold medal for his undergraduate thesis at St. Petersburg University. Besides being an active research mathematician and teacher, he was also active in politics and participated in the liberal movement in Russia at the beginning of the twentieth century.

In 1913, Markov organized a celebration of the 200th anniversary of Bernoulli’s discovery of the Law of Large Numbers. Markov was led to develop Markov chains as a natural extension of sequences of independent random variables. In his first paper, in 1906, he proved that for a Markov chain with positive transition probabilities and numerical states the average of the outcomes converges to the expected value of the limiting distribution. In a later paper he proved the central limit theorem for such chains. Writing about Markov, A. P. Youschkevitch remarks:

"Markov arrived at his chains starting from the internal needs of probability theory, and he never wrote about their applications to physical science. For him the only real examples of the chains were literary texts, where the two states denoted the vowels and consonants."

In a paper written in 1913, Markov chose a sequence of 20,000 letters from Pushkin’s Eugene Onegin to see if this sequence can be approximately considered a simple chain. He obtained the Markov chain with transition matrix
The limiting distribution for this chain is (.432, .568), indicating that we should expect about 43.2 percent vowels and 56.8 percent consonants in the novel, which was borne out by the actual count.

Claude Shannon considered an interesting extension of this idea in his book *The Mathematical Theory of Communication*, in which he developed the information theoretic concept of entropy. Shannon considers a series of Markov chain approximations to English prose. He does this first by chains in which the states are letters and then by chains in which the states are words. For example, for the case of words he presents first a simulation where the words are chosen independently but with appropriate frequencies.

*REPRESENTING AND SPEEDILY IS AN GOOD APT OR COME CAN DIFFERENT NATURAL HERE HE THE A IN CAME THE TO OF TO EXPERT GRAY COME TO FURNISHES THE LINE MESSAGE HAD BE THESE.*

He then notes the increased resemblance to ordinary English text when the words are chosen as a Markov chain, in which case he obtains

*THE HEAD AND IN FRONTAL ATTACK ON AN ENGLISH WRITER THAT THE CHARACTER OF THIS POINT IS THEREFORE ANOTHER METHOD FOR THE LETTERS THAT THE TIME OF WHO EVER TOLD THE PROBLEM FOR AN UNEXPECTED.*

A simulation like the last one is carried out by opening a book and choosing the first word, say it is the. Then the book is read until the word the appears again and the word after this is chosen as the second word, which turned out to be head. The book is then read until the word head appears again and the next word, and, is chosen, and so on.

Other early examples of the use of Markov chains occurred in Galton’s study of the problem of survival of family names in 1889 and in the Markov chain introduced by P. and T. Ehrenfest in 1907 for diffusion. Poincare in 1912 discussed card shuffling in terms of an ergodic Markov chain defined on a permutation group. Brownian motion, a continuous time version of random walk, was introduced in 1900-1901 by L. Bachelier in his study of the stock market, and in 1905-1907 in the works of A. Einstein and M. Smoluchowsky in their study of physical processes.