MARKOV CHAIN MONTE CARLO METHOD

MACHINE LEARNING 2 MIDTERM PROJECT

Haolin Ji, Jiachen Jiang, Suwen Wu

Northeastern University

March 2023

1 Introduction

Markov models are mathematical models used to describe systems that evolve over time, and are based on the assumption of memorylessness. Markov chains are a specific type of Markov model, which describe a sequence of events where the probability of each event depends only on the state of the system at the previous event (Grewal). The method simulates complex and high-dimensional distributions from probability distributions based on the law of large numbers, which states that as the sample size increases, the sample mean converges with the population means (Andrieu et al. 270). As the nomenclature suggests, it is derived from two methods, Monte Carlo sampling, and Markov Chain processes. Monte Carlo sampling is utilized to sample from a probability distribution (randomness) to ascertain the quantity of interest. For intricate computations, Monte Carlo methods necessitate simulation on multiple samples. However, the method has limitations as sampling from a large probability distribution is not always facile. In contrast, Markov Chain methods allow sampling from an intractable probability distribution. The Markov Chain process is beneficial in determining the probability of transitioning from one state to another. Markov Chain Monte Carlo (MCMC) employs the Markov Chain to determine a long-run probability distribution in the area of interest. It is a valuable technique for approximating posterior distributions, which may be exceedingly challenging or infeasible to compute analytically. By relying on iterative sampling methods, MCMC can generate representative samples from complex probability distributions and provide insight into a broad range of applications, including statistical inference, machine learning, and optimization.

$$P(X(t+n) = X/X(t)X(t+1)\dots X(t+k)) = X(t+n) = X/X(t)$$

1.1 Background

Markov Chain was first introduced and named after a Russian mathematician, Andrey Markov in early 20th century.

The origins of Monte Carlo algorithms can be traced back to the work of Ulam and von Neumann in the 1940s. Ulam is credited with conceiving the notion of intractable binational computation when he employed it to compute the probability of winning a card game (Robert and Casella 2). In 1947, Ulam and von Neumann introduced accept and reject methods for simulating nonuniform distributions. Von Neumann later adopted this idea and employed it in neutron diffusion techniques.

In order to optimize the limitations of the original Monte Carlo, Nicholas Metropolis, Arianna Rosenbluth, Marshall Rosenbluth, and Edward Teller proposed the Markov chain Monte Carlo in 1950.

The Metropolis algorithm derives its name from Nicolaus Metropolis, who employed it in the second computer named MANIAC in 1952. He utilized the Stanislav-Ulam law to design the hydrogen bomb employing the developed computer facilities. Hastings and Paskin subsequently advanced the Metropolis Algorithm as a statistical method to surmount the dimensionality issues associated with Monte Carlo methods in a paper titled "Equation of State Calculations by Fast Computing Machines" (Robert and Casella 3).

1.2 MCMC Process

MCMC simulations are applied for parameter estimations such as expected values, means, variances, and posterior distribution in Bayesian models. A Metropolis Algorithm is used where a distribution simulation is repeated in small independent steps in a process referred to as a random walk (Hamra et al. 633). Through numerous random walks, distributions similar to the actual distribution are produced. The MCMC process is characterized by several key features. Initially, a random point is selected to serve as the first sample. However, the model may take some time to move away from the selected point, and the resulting estimate might be biased if the targeted distribution is sporadic in that particular region. To mitigate this, an initial part of the Markov Chain is eliminated in the burn-in period. A proposal density is assigned to determine the following sample concerning the subsequent sample. The proposal distribution is determined to investigate the sample space with the most excellent density. An acceptance ratio is employed to assess the suitability

of the proposed subsequent sample. Gibb's sampling, which selects a candidate sample for the probability distribution that represents the density while keeping all other factors constant, can also be utilized.

1.3 Simple Illustration

A tourist intends to travel to multiple cities in a particular country. After visiting each city, he/she makes a decision regarding which city to visit next. The tourist has two options at each city: he/she can either head North or South. His/Her primary objective is to visit all the cities in the country, prioritizing those cities that have more national parks by spending more time in them. However, due to a lack of information, he/she is uncertain about the number of cities in the country or the number of national parks in each city. To obtain this information, he/she relies on her visit to a particular city, where he/she can determine the number of national parks present, and also enlists the help of a local guide to ascertain the number of national parks in the remaining cities.

Solutions:

The tourist can employ a coin-tossing mechanism to make a directional decision, which can be either North or South. His/Her next destination will be the proposed city if it has a larger number of national parks compared to his/her current location. On the other hand, if the proposed city has a lower number of national parks, then the tourist will:

- Each city θ has a population of θ national parks.
- α =ratio of national parks in the proposed city.
- she travels to the proposed city with a probability of α
- (1-α) = spends the day in the current city.
- The population for island is proportional to $\alpha^2 e^{-0.5}$

We can calculate the acceptance probability of α. For example, there are ten cities. City 1 and 5 have no national parks. She can flip a coin to move to city θ+1 or θ-1. The proposed city will be accepted since it has more national parks. On the other hand, θ-1 will be rejected, and she will stay in the current city. However, it can be accepted on the probability that (θ-1)/(θ).

Therefore, the acceptable probabilities are:

$$\begin{aligned}
\alpha(1 \longrightarrow 0) &= 0 \\
\alpha(1 \longrightarrow 2) &= 1 \\
\alpha(2 \longrightarrow 1) &= 0.5 \\
\alpha(2 \longrightarrow 3) &= 1 \\
\alpha(3 \longrightarrow 2) &= 0.667 \\
\alpha(3 \longrightarrow 4) &= 1 \\
\alpha(4 \longrightarrow 3) &= 0.75 \\
\alpha(4 \longrightarrow 5) &= 1 \\
\alpha(5 \longrightarrow 4) &= 0.8 \\
\alpha(5 \longrightarrow 6) &= 0
\end{aligned}$$

The starting point is city 3. She decides to move from city three to city 2 or 4. The move to city four is accepted, and she moves to the city. The move to city two is accepted but only with a probability of 0.667; otherwise, she stays in city 3.

1.4 Application

MCMC suits posterior distributions-summarizing uncertain factors (Hamra et al. 633). For example, it is suitable where the data is incomplete in a survey. The missing data can be simulated using MCMC procedures in such a situation. The ability of MCMC analysis to compute complex data allows it to be applied in hierarchical modeling. For example, it can show the relationship between alcohol and tobacco consumption in assessing the risk of developing esophageal cancer. Given the different data sets, such as wine, beer, and digestifs under the alcohol category, a researcher can combine factors in a single term or choose a subset (Hamra et al. 633).

1.5 Limitations

MCMC algorithm is majorly applied in the analysis of multi-dimensional problems. However, its samples are correlated in 1d distribution, resulting in a less accurate estimation (Hamra et al. 633). It can be corrected by applying other algorithms, such as adaptive rejection sampling. MCMC algorithm is slow in a sample with a high-dimensional space (Hamra et al. 633). The Hamilton-Monte Carlo algorithm alleviates this challenge by producing more significant steps between samples, leading to a lesser correlation.

2 Law of Large Numbers

Let $Y_1, Y_2 \cdots$ be any collection of IID r.v .'s (independent and identically distributed random variables) with common mean u and variance σ^2. The sample mean of the first n variables is

$$\overline{Y_n} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)$$

The **LLN** (Law of Large Numbers) says that Y_n converges to the true mean μ as $n \to \infty$.

Note that $\overline{Y_n} = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n)$ is the long-run average value of a sequence of measurements. For each possible value y_k define:

$$n(k) = \#i : Y_i = y_k$$

So then:

$$\frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \sum_k y_k \frac{n(k)}{n}$$

The **LLN** is equivalent to the statement that the relative frequency of occurrence of each value converges to its probability:

$$\frac{n(k)}{n} \to P(Y = y_k) \text{ as } n \to \infty$$

And therefore

$$\frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \sum_k y_k \frac{n(k)}{n} \to \sum_k y_k P(Y = y_k) = E(Y)$$

Real issue is what 'convergence' means. Given any $\varepsilon > 0$, $\delta > 0$, there is $N < \infty$ such that for all $n \leq N$, Y n will with probability at least $1 - \delta$ lie inside the interval $\mu \pm \varepsilon$. Or more succinctly, for every $\varepsilon > 0$,

$$\lim_{n \to \infty} P\left(|\frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) - x| > \varepsilon \right) = 0$$

3 Markov's Inequality

Many useful results in probability are proved using inequalities. The relevant one for the **LLN** is called Markov's inequality and is easily stated. For any random variable X and for any numbers $a > 0$ and $k > 0$,

$$P(|X| \geq a) \leq \frac{1}{a^k} E[|X|^k]$$

We can easily prove it:

$$\begin{aligned} E[|X|^k] &= \sum_i |x_i|^k P(X = x_i) \\ &\geq \sum_{i:x_i \geq a} |x_i|^k P(X = x_i) \\ &\geq a^k \sum_{i:x_i \geq a} P(X = x_i) \\ &= a^k P(|X| \geq a) \end{aligned}$$

An important special case of Markov's inequality is called Chebyshev's inequality: take $X = E(Y)$ and $k = 2$ to get:

$$P(|Y - E(Y)| \geq a) \leq \frac{1}{a^2} Var(Y)$$

The LLN follows easily from this. Take

$$X = \frac{1}{n}(Y_1 + Y_2 + \cdots + Y_n) = \overline{Y_n}$$

Then $E(X) = \mu$ and $Var(X) = \frac{\sigma^2}{n}$, hence

$$P(|Y - E(Y)| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2} \to 0$$

4 Central Limit Theorem

Let Y_1, Y_2...be IID with finite mean $E(Y_i) = \mu$ and finite variance $VAR(Y_i) = \sigma^2$. Define

$$Z_n = \frac{\sqrt{n}(\overline{Y_n} - \mu)}{\sigma} = \frac{Y_1 + Y_2 + \cdots + Y_n - n\mu}{\sigma\sqrt{n}}$$

Then for all $a < b$, as $n \to \infty$,

$$P(a < Z_n < b) \to P(a < Z_n \leq b = \int_a^b \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx$$

The integrand on the right side is the pdf of the standard normal. So another way to state the **CLT** is

$$z_n \to Z \sim N(0,1) \text{ (Convergence in disitribution)}$$

Even more informally, we can say that for n large,

$$\frac{1}{n}\sum_{i=1}^{n} Y_i = \mu + \frac{\sigma}{\sqrt{n}} Z + \cdots$$

Or even

$$\sum_{i=1}^{n} Y_i = n\mu + \sigma\sqrt{n} Z + \cdots$$

where $\cdots$ goes to zero faster than the leading order terms as as $n \to \infty$. This is the most useful way for us to think about the meaning of the **CLT**.

4.1 Central Limit Theorem: sketch of proof

The main idea of the proof is to consider the moment generating function of Z_n, that is

$$M_n(t) = E[e^{tZ_n}], t \in R$$

and show that for every number t it converges to the moment generating function of the standard normal, which is

$$M_n(t) = E[e^{tZ_n}] = e^{t^2/2}$$

(this result for the standard normal follows from a simple calculation using Gaussian integrals).

Once this convergence has been shown for every t, it follows that all the moments of Z_n converge to the corresponding moments of Z, and this gives the result. The proof of convergence follows a calculation which we sketch below. First, for each $i = 1, 2, ...n$ we define:

$$X_i = \frac{Y_i - \mu}{\sigma}$$

It follows that $E(X_i) = 0$ and $VAR(X_i) = 1$, and

$$Z_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} X_i$$

Also, since the $\{X_i\}$ are independent we can compute

$$M_n(t) = E[e^{tn^{-1/2}\sum_i X_i}] = E(e^{tn^{-1/2}X_i})^n$$

Now we just observe that for large n, using the Taylor series for the exponential

$$E[e^{tn^{-1/2}X}] = 1 + E[tn^{-1/2}X] + \frac{1}{2}E[t^2n^{-1}X^2] + \cdots = 1 + 0 + \frac{t^2}{2n} + \cdots$$

The remainder terms go to zero faster than n^{-1} as $n \to \infty$, so we can ignore these to compute

$$M_n(t) = (1 + \frac{t^2}{2n})^n \to e^{\frac{t^2}{2}} = M(t)$$

5 Monte Carlo method

5.1 Introduction

Monte Carlo method is also called computer stochastic simulation method. Its theoretical basis is actually Law of large number, which proves that in the case of a large sample, the frequency of events is equal to the probability. MC method is a calculation of some deterministic problems based on the statistical results of a large number of events. The core of this method is to transform the problem into a probability problem, and generate a series of random numbers by computer simulation, and then carry out statistical work on the random numbers.

We can also understand this method in this way: through the repeated execution of the algorithm (at the cost of increasing the execution time of the algorithm), the probability of error can be reduced to a negligible level (the more calculation, the more accurate the result)

5.2 Definition

When the problem we need to solve is the probability of the occurrence of a certain random event, or the expected value of a certain random variable, we can estimate the probability of this random event with the frequency of the occurrence of this event through some "experimental" method, or obtain some numerical characteristics of this random variable. We take the result of calculation as the solution of the problem. This is a statistical method. In short, the frequency or probability of the occurrence for many random samples is used as the solution of the problem. Monte Carlo method is generally divided into three steps, including the process of constructing random probability, sampling from the construction of random probability distribution, and solving the estimator.

5.3 Properties of Monte Carlo

(1) We can obtain approximately accurate results through random sampling

(2) The more samples we take, the closer our calculation results will be to the real value.

5.4 Advantages:

(1) It is a complete valuation method, which can deal with nonlinear, large fluctuation and thick tail problems;

(2) We can use the computer to repeatedly generate simulation data, and the calculation results are more reliable and accurate;

(3) We can use the historical data information of risk factor changes to improve and modify the stochastic simulation model, and the simulation of future changes of risk factors is closer to reality.

5.5 Disadvantages:

(1) We need to construct a random model in advance, and there is a risk of model and parameter estimation.

(2) The convergence speed is slow, the computational efficiency is low, and a large number of times of repeated simulation are required.

(3) In order to reduce the time cost and improve the calculation efficiency, it will increase the sample variance and reduce the reliability and accuracy of the calculation results

6 Generation and Theoretical Analysis of Markov Chains

In the process of our study, we learned that after Markov random process was proposed by Markov, the concept of Markov chain also continued to emerge, and the state transition matrix and initial state probability distribution became an important part of the Markov chain. Because Markov chain has its own unique properties, the construction of Markov chain model should follow these rules. Our team will analyze the definition, basic properties, state classification, stationarity and ergodicity of Markov chain in detail below.

6.1 Markov Process

Markov process has become an important part of the branch theory of stochastic process in modern probability theory and mathematical statistics. When the state of a system or process at a specific time t1 which is known, the conditional probability distribution of the state of the system or process at time t>t1 has nothing to do with the time and the previous state, but only related to time t1. In other words, when the "present" state of the system or process is known, the "future" and "past" are independent of each other, and its "future" state does not depend on the "past" state. This property is usually called Markov property or no aftereffect.

The Markov property is described by the probability distribution function, which can be expressed in the following form:

Let the state space of a random process $\{X(t), t \in T\}$ be S. If for any moment of time t, here $t_1 < t_2 < ... < t_n$, $n \geq 3$, $t_i \in T$, in the case of $X(t_i) = x_i$, $x_i \in S$, $i = 1, 2, ..., n-1$, if the conditional probability distribution function of $X(t_n)$ is exactly be equal to the conditional probability distribution function of $X(t_n)$ as $X(t_n) = X_{n-1}$, it is:

$$P\{X(t_n) < X_n, X(t_1) = x_1, X(t_2) = x_2, ..., X(t_{n-1}) = x_{n-1}\} = P\{X(t_n) < X_n, X(t_{n-1}) = x_{n-1}\}$$

$$X_n \in R$$

In this case, we call the random process X, which has Markov property or has no aftereffect, and further call this process a Markov process.

6.2 The Definition of a Markov chain

Suppose a random process $\{X(t), t \in T\}, T = \{0, 1, 2, \dots\}$ with a state space $S = \{0, 1, 2, \dots\}$. If for any positive integer m, n, p and any non-negative integer $j_m > j_{m-1} > \cdots > j_2 > j_1$ $(n > j_m)$ and $i_{n+p}, i_n, i_{j_m}, \dots, i_{j_2}, i_{j_1}$, there is:

$$P\{X(n+p) < i_{n+p}, X(n) = i_n, X(j_m) = i_{j_m}, \dots, X(j_2) = i_{j_2}, X(j_1) = i_{j_1}\} = P\{X(n+p) = i_{n+p}, X(n) = i_n\}$$

, then we can call X_T as a Markov chain.

6.2.1 One-step transition probability

We call $P_{ij} = P(X_{t+1} = j, X_t = i), i, j \in I$ as the transition probability of the Markov chain X(t). And $P_{ij} = (P_{ij})_{i,j \in I}$ is called a one-step transition probability matrix. For example, there are three states in total, represented by 1, 2, and 3, then the one-step transition probability matrix is:

$$\mathbf{P} = \begin{bmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{bmatrix}$$

6.2.2 K-step Transition Probability

$p_{ij}(0) = (X_{t+1} = j|X_t = i)$is equal to 1 when i=j, otherwise it is equal to 0. It means that there can only be one state at a time, and there cannot be multiple states at one time;

$p_{ij}(k) = (X_{t+k} = j|X_t = i)$ is the k-step transition probability of $\{X_t\}$. Matrix $P^k = p_{ij}(k)$ is the k-step transition probability matrix of $\{X_t\}$. We can write the k-step transition matrix P of the Markov chain as:

$$\mathbf{P}^k = \begin{bmatrix} p_{11}^k(n) & p_{12}^k(n) & \dots & p_{1m}^k(n) \\ p_{21}^k(n) & p_{22}^k(n) & \dots & p_{2m}^k(n) \\ \vdots & \vdots & \ddots & \vdots \\ p_{m1}^k(n) & p_{m2}^k(n) & \dots & p_{mm}^k(n) \end{bmatrix}$$

$P^{(0)}$ is an identity matrix. When $P^{(1)} = P$, it is a one-step transition probability matrix. According to the definition we can write the one-step transition matrix P of the Markov chain:

$$\mathbf{P} = \begin{bmatrix} p_{11} & p_{12} & \dots & p_{1n} \\ p_{21} & p_{22} & \dots & p_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ p_{n1} & p_{n2} & \dots & p_{nn} \end{bmatrix}$$

6.2.3 Basic Properties of Markov chains

For the transition matrix of a Markov chain, its elements satisfy the following two basic properties.

(1) The elements in the matrix are non-negative numbers, that is $p_{ij}^{(h)}(n) \geq 0, i, j = 1, 2, ...n$

(2) The sum of the elements in each row in the matrix is 1, it means: $\sum p_{ij}^{(h)}(n) = 1$

(3) Kolmogorov-Chepman Equation (C-K equation)

For any $m, n \geq 0$ there is :$p_{ij}(h + l) = \sum_{k \in I} P_{ik}^h P_{kj}^l$, which is $P^{(h+l)} = P^{h+l}$. We can understand that the (h+l)-step transfer matrix is equal to the $(h + l)th$ power of the one-step transfer matrix.

Here is our group's proof of the $C - K$ equation

For $\forall h, l \in Z^*$, there are:

$$
\begin{aligned}
P_{ij}^{(h+l)}(n) &= p\{X(n+h+l)=j|X(n)=i\} = p\{X(n+h+l)=j, \bigcup_{r\in S}\{X(n+h)=r\}|X(n)=i\} \\
&= \sum_{r\in S} p\{X(n+h+l)=j|X(n+h)=r|X(n)=i\} \\
&= \sum_{r\in S} \frac{p\{X(n+h+l)=j|X(n+h)=rX(n)=i\}}{p\{X_{(n)}=i\}} \\
&= \sum_{r\in S} \frac{p\{X(n+h+l)=j|X(n+h)=rX(n)=i\}}{p\{X_{(n+h)}=r, X_{(n)}=i\}} * \frac{p\{X(n+h)=rX(n)=i\}}{p\{X_{(n)}=i\}} \\
&= \sum_{r\in S} p\{X(n+h+l)=j|X(n+h)=r, X(n)=i\} * p\{X(n+h)=r|X(n)=i\} \\
&= \sum_{r\in S} p_{ir}^{h}(n) p_{rj}^{l}(n+h), S=\{1,2,\cdots n\}
\end{aligned}
$$

We can express the above $C-K$ equation in matrix form as: $p^{h+l}(n) = p^h(n)p^l(n+h)$.

Assuming l=1, according to $C-K$ equation we can get: $p_{ij}^{(h+1))}(n) = \sum p_{ir}^{(h)}(n)p_{rj}(n+h)$.

We recurse the formula to get $p_{ij}^{(h+1)}(n) = \sum p_{rj1}(n)p_{j1j2}(n+1)...p_{jhj}(n+h)$, which is $p^{(h)} = p^{(1)}p^{(h-1)}$. Therefore, we can conclude that: $p^{(h)} = (P)^h$. This means that the Markov chain multi-step transition matrix can be directly obtained from the one-step transition matrix.

Through our study and analysis, we understand that for a homogeneous Markov chain, its finite-dimensional probability distribution can be completely predicted and determined by the initial probability distribution and one-step transition probability matrix.

(4) Stationary Distribution

The probability distribution of the Markov chain at time t is called the stationary distribution of time t:

$$
\pi(t) = \begin{bmatrix} \pi_1(t) \\ \pi_2(t) \\ \pi_3(t) \\ \cdots \end{bmatrix}
$$

Where $\pi_i(t) = P(X_t = i), i = 1, 2, ...$ represents the probability of the $i-th$ state at time t. For example, states 1, 2, and 3 in the above example represent that a person is eating or sleeping or playing games at noon every day. We can arbitrarily assume an initial state, that is, at time t, there is 50% probability of eating and 30% probability of sleeping, 20% chance to play games.

Assume that the chain is irreducible and persistent (no transient states in the chain). Let $\{\pi_i\}$ $(i \in R)$ be a probability distribution on R. We say that π_i is stationary if

$$
\sum_i \pi_i P_{ij} = \pi_j, \text{ for all states j} \in \text{R}
$$

So we suppose that $P(X_0 = j) = \pi_j$ for all $j \in R$, where π is the stationary distribution. Then $P(X_0 = j) = \pi_j$ for all $j \in R$, and all $n \geq 1$.

Here we can prove it:

$$\begin{aligned} P(X_1 = j) &= \sum_i P(X_1 = j|X_0 = i)P(X_0 = j) \\ &= \sum_i \pi_i P_{ij} \\ &= \pi_j \end{aligned}$$

where we used the stationarity property in the last line.

Now repeat the argument for X_2 by conditioning on X_1, and so on. This result explains why it is called a stationary distribution: once the chain enters this distribution it must remain there.

6.3 Properties of Markov chains

6.3.1 irreducible and regular

Given a Markov chain $X = \{X_0, X_1, ..., X_t, ...\}$, for any state $i, j \in S$, if there exists a time t that satisfies: $P(X_t = i|X_0 = j) > 0$, that is, the probability of starting from state j at time 0 and arriving at state i at time t is greater than 0, then this Markov chain is irreducible, otherwise the Markov chains is reducible.

An irreducible Markov chain, starting from any state, can reach any state after a sufficiently long time.

Here is some definitions for states in Markov chain.

Let P be the transition matrix of a Markov chain.

(1) The Markov chain is irreducible if for all states i, j there is an integer $n(i, j)$ such that $P_{ij}(n(i, j)) > 0$.

(2) The Markov chain is regular if there is an integer n such that $P^n > 0$.

6.3.2 Finite and irreducible

Now suppose the chain is irreducible but not regular. Then we get a similar but weaker result.

Let P be the transition matrix of an irreducible Markov chain. Then there is a unique strictly positive probability vector w such that

$$w^T P = w^T$$

Furthermore

$$\frac{1}{(n+1)}(I + P + P^2 + \cdots + P^n) \to ew^T, \text{as } n \to \infty$$

This Theorem allows the following interpretation: for an irreducible chain, w_j is the long-run fraction of time the chain spends in state j.

Here we need to prove finite state irreducible

Define:

$$Q = \frac{1}{2}I + \frac{1}{2}P$$

Then Q is a transition matrix. Also:

$$2^n Q^n = \sum_{k=0}^{n} \binom{n}{k} P^k$$

Because the chain is irreducible, for all pairs of states i, j there is an integer $n(i,j)$ such that $(P^{n(i,j)})_{ij} > 0$. Let $n = \max n(i,j)$, then for all i,j we have

$$2^n (Q^n)_{ij} = \sum_{k=0}^{n} \binom{n}{k} (P^k)_{ij} \geq \binom{n}{n(i,j)} (P^{n(i,j)})_{ij} > 0$$

And hence Q is regular. Let w be the unique stationary vector for Q then

$$w^T Q = w^T \leftrightarrow w^T P = w^T$$

which shows existence and uniqueness for P.

Let $W = w^T e$ then a calculation shows that for all n

$$(I + P + P^2 + \cdots + P^{n-1})(I - P + W) = I - P^n + nW$$

Note that $I - P + W$ is invertible: indeed if $y^T(I - P + W) = 0$ then

$$y^T - y^T P + (y^T e)w = 0$$

Multiply by e on the right and use $Pe = e$ to deduce

$$y^T e - y^T Pe + (y^T e)(w^T e) = (y^T e)(w^T e) = 0$$

Since $w^T e = 1 > 0$ it follows that $y^T - y^T P = 0$. By uniqueness this means that y is a multiple of w, but then $y^T e = 0$ means that y=0. Therefore $I - P + W$ is invertible, and so

$$I + P + P^2 + \cdots + P^{n-1} = (I - P^n + nW)(I - P + W)^{-1}$$

Now $WP = W = W^2$ hence

$$W(I - P + W) = W \rightarrow W = W(I - P + W)^{-1}$$

Therefore:

$$I + P + P^2 + \cdots + P^{n-1} = W + \frac{1}{n}(I - P^n)(I - P + W)^{-1}$$

It remains to show that the norm of the matrix $(I - P^n)(I - P + W)^{-1}$ is bounded as $n \rightarrow \infty$, or equivalently that $(I - P^n)$ is uniformly bounded. This follows from the bound

$$||p^n z|| \leq \sum_{ij} (P^n)_{ij} |z_j| = \sum_j |z_j|$$

Therefore $\frac{1}{n}(I - P^n)(I - P + W)^{-1} \rightarrow 0$ and the result follows.

6.3.3 Aperiodic: the return time common divisor is 1

Given a Markov chain $X = \{X_0, X_1, ..., X_t, ...\}$, for any state $i \in S$, if starting from state i at time 0 and returning to the state at time t, the greatest common divisor of all time lengths $t : P(X_t = i | X_0 = j) > 0$ is 1, that is, period $d = 1$. Then this Markov chain is aperiodic, otherwise it is periodic.

6.4 State Classification of Markov chains

In the actual problem research process, we often have to study the probability that a system or process will reach another state after going through a lot of state transitions from one state. Therefore, we need to find the value of $\lim_{+\infty} p_{ij}^{(n)}$ to determine whether this limit value will exist after the system has undergone a long state transition. If so, whether this state is related to the start state. After the study of our group, the limit value is related to the initial state and the final state. The following are the specific properties:

6.4.1 First Arrival Time:

$T_{ij} = minn : n \geq 1, X_0 = i, X_n = j$ represents the time when starting from state i and arriving at state j for the first time. If starting from state i can never reach state j, then $T_{ij} = \infty$

6.4.2 Probability of First Arrival:

$$f_{ij}^{(n)} = P(X_n = j, X_m \neq j, m = 1, 2, ..., n-1 | X_0 = i) \tag{1}$$

$f_{ij}^{(n)}$ is the probability of starting from state i and arriving at state j for the first time after n steps.

$f_{ij}^{(\infty)}$ is the probability that starting from state i will never reach state j.

Starting from state i, the probability of reaching state j for the first time within finite steps:

$$f_{ij} = \sum_{n=1}^{+\infty} f_{ij}^{(n)} = P(T_{ij} < +\infty) \tag{2}$$

State i is a return state: $f_{ii} = 1$. It must be able to return to state i within a limited number of steps;

State i is an abnormal return state: $f_{ii} < 1$. It cannot go back to state i in a finite number of steps.

Average return time: $u_i = \sum_{n=1}^{+\infty} n f_{ii}^{(n)}$

In addition, we also learned some other attachment definitions of Markov chains:

Positive recurrent state: if $u_i < +\infty$, then the normal return state i is positive recurrent state.

(Leave this state with a limited number of steps and it will be able to come back)

Null recurrent state: if $u_i = +\infty$, then the normal return state i is null recurrent state.(This state can come back, but it takes infinite steps)

Ergodic state: a Markov chain is said to be ergodic if there exists a positive integer such that for all pairs of states in the Markov chain, if it is started at time 0 in state then for all , the probability of being in state at time is greater than .

6.5 Time reversible Markov chains

Consider an ergodic chain $\{...X_{n-1}, X_n, ...\}$ with transition probabilities P_{ij} and stationary distribution π_j. We have

$$P_{ij} = P(X_n = j | X_{n-1} = 1)$$

Now consider the reversed chain, where we run the sequence backwards:$\{...X_{n-1}, X_n, ...\}$. The transition matrix is

$$\begin{aligned} q_{ij} &= P(X_n = j | X_{n-1} = i \\ &= \frac{P(X_n = j, X_n = i)}{P(X_n = i)} \\ &= P(X_n = j | X_n = i) \frac{P(X_{n-1} = j)}{P(X_n = i)} \\ &= P_{ij} \frac{P(X_{n-1} = j)}{P(X_n = i)} \end{aligned}$$

Assume that the original chain is in its stationary distribution so that $P(X_n = i) = \pi_i$ for all i, then this is

$$q_{ij} = P_{ij}\frac{\pi_j}{\pi_i}$$

So we can get a definition that the Markov chain is reversible if $q_{ij} = P_{ij}$ for all $i, j \in S$.

And there is a lemma.

Consider an irreducible Markov chain with transition probabilities P_{ij}. Suppose there is a distribution $w_i > 0$ such that for all $i, j \in S$.

$$w_i P_{ij} = w_j P_{ji}$$

Then the chain is time reversible and w_j is the stationary distribution.

The quantity $w_i P_{ij}$ has another interpretation: it is the rate of jumps of the chain from state i to state j. More precisely, it is the long-run average rate at which the chain makes the transition between these states:

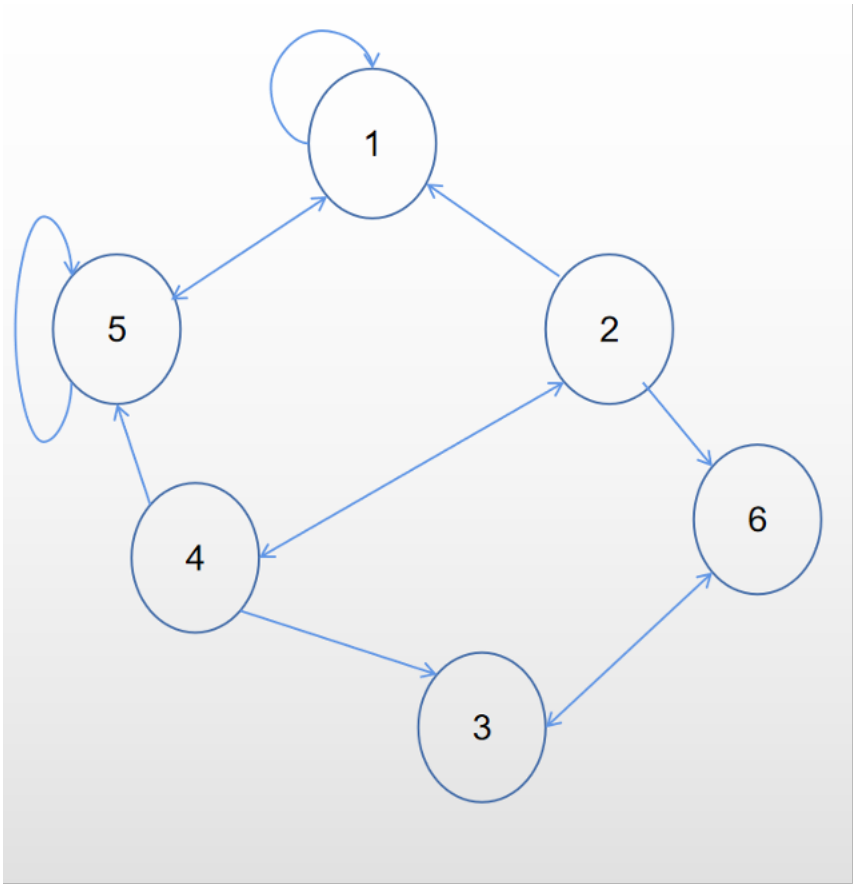
$$\lim_{n\to\infty} P(X_n = i, X_{n+1} = j) = \pi_i P_{ij}$$

And this often helps us to figure out if a chain is reversible or not.

7 Class Example

7.1 Example 1

Determine classification for the states in the Markov Chain



$\{1, 5\}$,$\{3, 6\}$ is permanent.
$\{2, 4\}$ is transient.

7.2 Example 2

Calculate the stationary distribution for Markov Chain

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

$$\begin{cases} w_1 = 3/4w_1 + 1/4w_3 \\ w_2 = 1/4w_1 + 2/3w_2 + 1/3w_3 \\ w_3 = 1/3w_2 + 1/2w_3 \end{cases}$$

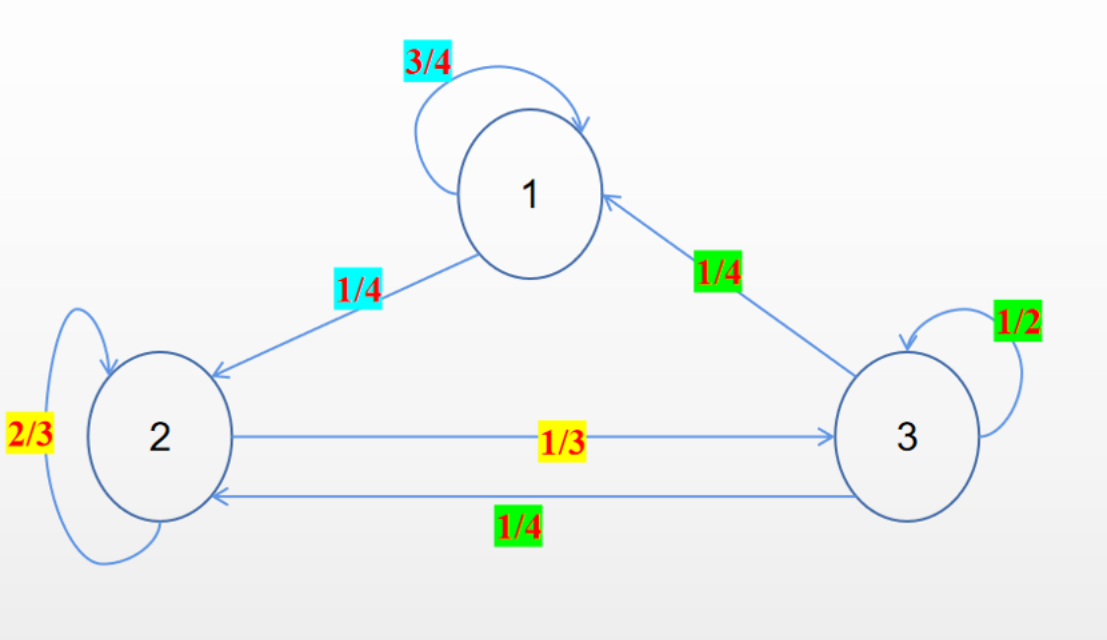
and $w_1 + w_2 + w_3 = 1$

$$\begin{cases} w_1 = 2/7 \\ w_2 = 3/7 \\ w_3 = 2/7 \end{cases}$$

The expected number of return to state1 is 7/2

The expected number of return to state2 is 7/3

The expected number of return to state3 is 7/2



7.3 Example 3

Determine whether the Markov chain is reversible.

$$\begin{Bmatrix} 0 & 0 & 0.8 & 0.2 \\ 0 & 0 & 0.5 & 0.5 \\ 0.4 & 0 & 0.6 & 0 \\ 0.2 & 0.4 & 0.4 & 0 \end{Bmatrix}$$

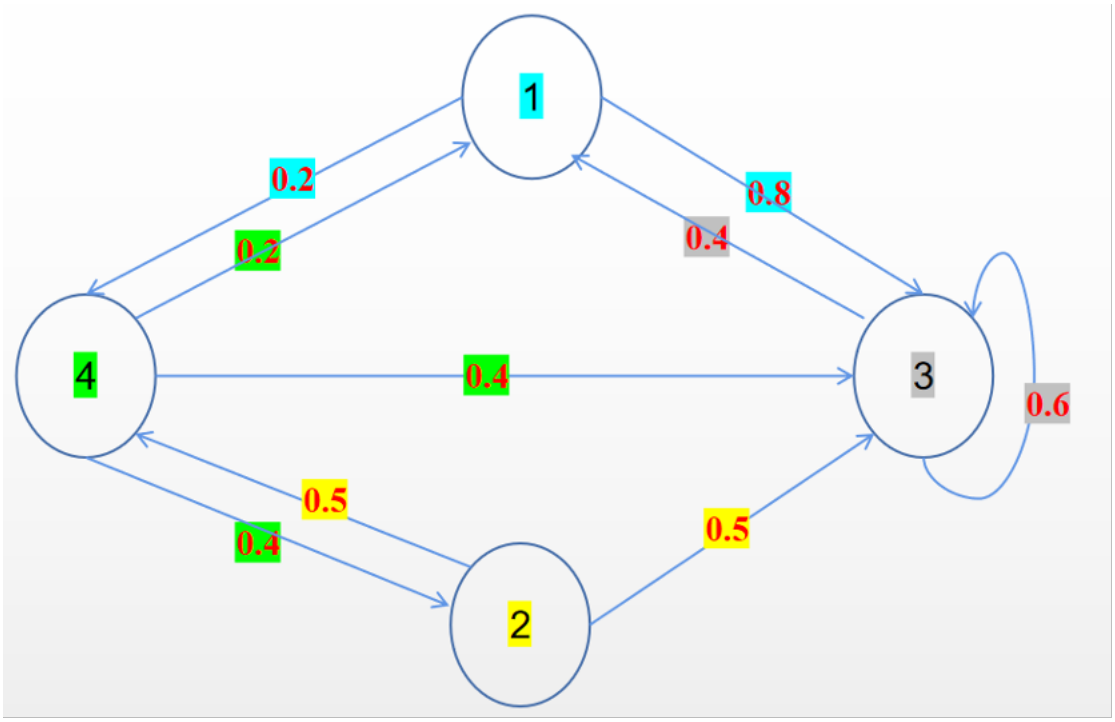
$$\begin{cases} w_1 = 0.4w_3 + 0.2w_4 \\ w_2 = 0.4w_4 \\ w_3 = 0.8w_1 + 0.5w_2 + 0.6w_3 + 0.4w_4 \\ w_4 = 0.2w_1 + 0.5w_2 \end{cases}$$

$w_3 = 9.5w_4$

or $w_3 = 2w_4$

There is no $w_1 + w_2 + w_3 + w_4 = 1$

The chain is not reversible.



7.4 Example 4

A knight moves randomly on a standard 8*8 chessboard. At each step it chooses at random one of the possible legal moves available. Given that the knight starts in its original position of the chessboard, how can we find the expected number of steps until its first return to its initial position.

It can move to every position on the chessboard. Each value on the table represents the number of places it can reach.

So we can know the total number of steps is $(2 + 3 + 4 + 4 + ... + 2) = 336$

Then we can get stationary distribution:

So time to return to its original state is its reciprocal 122.

2	3	4	4	4	4	3	2
3	4	6	6	6	6	4	3
4	6	8	8	8	8	6	4
4	6	8	8	8	8	6	4
4	6	8	8	8	8	6	4
4	6	8	8	8	8	6	4
3	4	6	6	6	6	4	3
2	3	4	4	4	4	3	2

1/166	1/122	1/84	1/84	1/84	1/84	1/122	1/166
1/122	1/84	1/56	1/56	1/56	1/56	1/84	1/122
1/84	1/56	1/42	1/42	1/42	1/42	1/56	1/84
1/84	1/56	1/42	1/42	1/42	1/42	1/56	1/84
1/84	1/56	1/42	1/42	1/42	1/42	1/56	1/84
1/84	1/56	1/42	1/42	1/42	1/42	1/56	1/84
1/84	1/84	1/56	1/56	1/56	1/56	1/84	1/84
1/166	1/122	1/84	1/84	1/84	1/84	1/122	1/166

8 References

[1] Andrieu, Christophe, et al. "Particle Markov Chain Monte Carlo Methods." Journal of the Royal Statistical Society Series B-statistical Methodology, vol. 72, no. 3, Wiley-Blackwell, June 2010, pp. 269–342. https://doi.org/10.1111/j.1467-9868.2009.00736.x.

[2] Grewal, Jasleen K., et al. "Markov models—Markov Chains." Nature Methods, vol. 16, no. 8, 2019, pp. 663–64, https://doi.org/10.1038/s41592-019-0476-x.

[3] Hamra, Ghassan B., et al. "Markov Chain Monte Carlo: An Introduction for Epidemiologists." International Journal of Epidemiology, vol. 42, no. 2, Oxford UP, Apr. 2013, pp. 627–34. https://doi.org/10.1093/ije/dyt043.

[4] Robert, Christian P., and George Casella. "A Short History of Markov Chain Monte Carlo: Subjective Recollections From Incomplete Data." Statistical Science, vol. 26, no. 1, Institute of Mathematical Statistics, Feb. 2011, https://doi.org/10.1214/10-sts351.

[5] 颜荣芳, 股票市场预测的随机过程[J].模型西北师范大学学报(自然科学版), 2003(3):44-46

[6] 韦丁源, 股市大盘指数的马尔科夫链预测法[J].广西广播电视大学学报, 2008(9):66-69

[7] 基于数据挖掘的股票价格预测研究[D].吉林,长春大学,2005,6.

[8] 台文志, 利用马尔可夫链模型预测股票市场的近期走势[J].西南民族大学学报(自然科学版), 2008(3):477-481.