Markov Chains

Apurva Nakade

2025-02-23

Table of contents

Definitions

In this module, we will learn about Markov Chains. A Markov Chain is a discrete-time stochastic process that satisfies the Markov property. The Markov property states that the future state of the process depends only on the current state and not on the sequence of events that preceded it. In other words, the future is conditionally independent of the past given the present.

Markov Chain: A Markov chain is a sequence of random variables $X_0, X_1, X_2, \ldots$ satisfying the following properties:

1. **State Space**: The random variables X_i take values in a finite set Ω called the state space.
2. **Markov Property**: For all $n \geq 0$ and all states $i_0, i_1, \ldots, i_{n+1} \in \Omega$,

$$P(X_{n+1} = i_{n+1} \mid X_0 = i_0, X_1 = i_1, \ldots, X_n = i_n) = P(X_{n+1} = i_{n+1} \mid X_n = i_n).$$

3. **Time Homogeneity**: The transition probabilities $P(X_{n+1} = j \mid X_n = i)$ do not depend on n.

The transition matrix of a Markov chain is a square matrix whose (i, j)-th entry is the probability of transitioning from state i to state j in one time step.

Transition Matrix: Let $X_1, X_2, X_3, \ldots$ be a Markov chain with state space $\Omega = \{1, 2, \ldots, N\}$. The transition matrix P of the Markov chain is an $N \times N$ matrix whose (i, j)-th entry is given by

$$P(i, j) = P(X_{n+1} = j \mid X_n = i).$$

Throughout this notebook, we'll let $X_0, X_1, X_2, \ldots$ be a Markov chain with state space $\Omega = \{1, 2, \ldots, N\}$ and transition matrix P.

We can represent a Markov chain by a directed graph called a **state diagram**. Each state is represented by a node, and the transition probabilities are represented by directed edges between the nodes. The transition matrix can be derived from the state diagram by assigning the transition probabilities to the corresponding entries of the matrix.

The probability distribution of the Markov chain at time n is a *row vector* π_n whose i^{th} entry is $\mathbb{P}(X_n = i)$ for each $i \in \Omega$.

Theorem 0.1. *Let π_n be the probability distribution of the chain at time n. Then,*

$$\pi_{n+1} = \pi_n P.$$

And hence,

$$\pi_n = \pi_0 P^n.$$

Proof.

$$\begin{aligned}
\pi_{n+1}(j) &= \mathbb{P}(X_{n+1} = j) \\
&= \sum_{i \in S} \mathbb{P}(X_{n+1} = j \mid X_n = i) \mathbb{P}(X_n = i) \\
&= \sum_{i \in S} \pi_n(i) P(i, j) \\
&= \pi_n P(j).
\end{aligned}$$

□

Stationary Distribution

A probability distribution π is called a **stationary distribution** of a Markov chain with transition matrix P if $\pi = \pi P$.

P is guaranteed to have an eigenvalue of 1 because it's row sum is 1 i.e.

$$P\vec{1} = \vec{1},$$

where $\vec{1}$ is the vector of all ones. Since, there is a right eigenvector corresponding to the eigenvalue 1, there will be a left eigenvector as well. The left eigenvector is a stationary distribution of the Markov chain.

It is not hard to see that every eigenvalue of P is less than or equal to 1 in magnitude. Suppose $\vec{v}$ is a left eigenvector of P corresponding to an eigenvalue λ. Let v_I be the largest component of $\vec{v}$ in magnitude. Then, we have

$$\begin{aligned}
\lambda\vec{v} &= \vec{v}P \\
\implies \lambda v_I &= \sum_j v_j P(j, I) \\
&\leq \sum_j |v_j| P(j, I) \\
&\leq \sum_j |v_I| P(j, I) \\
&= |v_I|
\end{aligned}$$

Thus, $|\lambda| \leq 1$. In particular, this means that the Frobenius norm of P is less than or equal to 1 and for all vectors $\vec{v}$, we have

$$\|\vec{v}\|_2 \leq \|\vec{v}P\|_2.$$

We are particularly interested in the case when there is exactly one eigenvector with eigenvalue of magnitude 1 which would then be the stationary distribution of the Markov chain.

Fundamental Theorem

We say that a Markov Chain is **irreducible** if for every pair of states $i, j \in \Omega$, there exists an integer n such that $P^n(i,j) > 0$ i.e. it is possible to go from any state to any other state in a finite number of steps.

We say that a state i is **aperiodic** if the greatest common divisor of the set $\{n \geq 1 : P^n(i,i) > 0\}$ is 1. A Markov chain is called **aperiodic** if all its states are aperiodic.

Note that sometimes we add a preliminary requirement that the set $\{n \geq 1 : P^n(i,i) > 0\}$ is non-empty. This condition is called **positive recurrence**. We'll assume this as a part of the definition of aperiodicity.

A Markov chain is called **ergodic** if it is irreducible and aperiodic.

Theorem 0.2. ***Fundamental Theorem of Markov Chains***: *If a Markov chain is ergodic, then it has a unique stationary distribution* Π. *Moreover, in this case, for any initial distribution* π_0*, the distribution of the chain converges to* Π *as* $n \to \infty$ *i.e.*

$$\lim_{n\to\infty} \pi_0 P^n = \Pi,$$

for any initial distribution π_0.

Random Walk on Graphs

Our main example of a Markov chain is the random walk on a graph. Let $G = (V, E)$ be a graph with vertex set V and edge set E. Suppose you want to move from one vertex to another by following the edges of the graph. At each vertex, you choose an edge uniformly at random and move to the adjacent vertex. This process is called a random walk on the graph.

The random walk on G is a Markov chain with state space $\Omega = V$ and transition probabilities given by

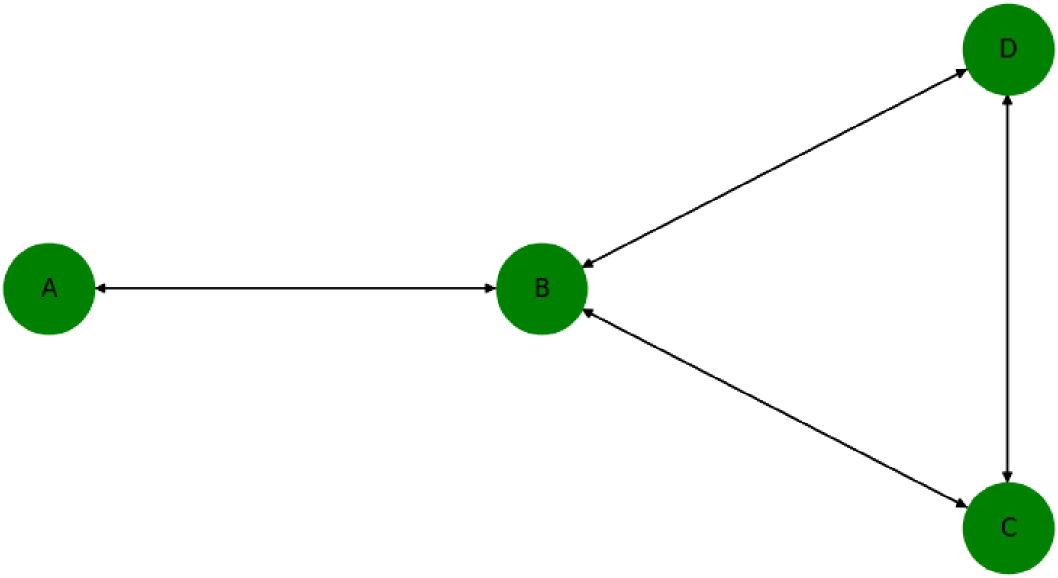
$$P(i,j) = \begin{cases} \frac{1}{\deg(i)} & \text{if } (i,j) \in E, \\ 0 & \text{otherwise,} \end{cases}$$

where $\deg(i)$ is the degree of vertex i i.e. the number of edges incident to i.

Example 0.1. Consider a graph G with 4 vertices as shown below. The transition matrix of the random walk on G is given by

$$P = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1/3 & 0 & 1/3 & 1/3 \\ 0 & 1/2 & 0 & 1/2 \\ 0 & 1/2 & 1/2 & 0 \end{pmatrix}.$$

Suppose we start at vertex A. This means that the initial distribution is $\pi_0 = [1, 0, 0, 0]$. The n^{th} distribution π_n can be obtained by multiplying π_0 with the transition matrix P^n, as shown below.



```
Transition matrix of the Markov chain:
[[0.         1.         0.         0.        ]
 [0.33333333 0.         0.33333333 0.33333333]
 [0.         0.5        0.         0.5       ]
 [0.         0.5        0.5        0.        ]]
```

```
Example of evolution of a Markov chain:
State at time 0 : [1, 0, 0, 0]
State at time 1 : [0. 1. 0. 0.]
State at time 2 : [0.333 0.    0.333 0.333]
State at time 3 : [0.    0.667 0.167 0.167]
State at time 4 : [0.222 0.167 0.306 0.306]
State at time 5 : [0.056 0.528 0.208 0.208]
State at time 6 : [0.176 0.264 0.28  0.28 ]
State at time 7 : [0.088 0.456 0.228 0.228]
State at time 8 : [0.152 0.316 0.266 0.266]
State at time 9 : [0.105 0.418 0.238 0.238]
State at time 10 : [0.139 0.344 0.259 0.259]
```

In HW, you'll prove the following theorem about ergodicity of random walks on graphs.

Theorem 0.3. ***Theorem**: A random walk on a graph is*

1. *Irreducible if and only if the graph is connected.*
2. *Aperiodic if and only if the graph is not bipartite.*

If these conditions hold, then the stationary distribution of the random walk is given by

$$\Pi(i) = \frac{deg(i)}{2|E|},$$

where $deg(i)$ is the degree of vertex i and $|E|$ is the number of edges in the graph.

Mixing Time

Assume that the Markov chain is ergodic, and hence has a stationary distribution Π. We know that any initial distribution π_0 converges to Π as $n \to \infty$. The mixing time measures the rate of this convergence.

The **total variation distance** between two probability distributions μ and ν on a finite state space Ω is defined as

$$\begin{aligned} \|\mu - \nu\|_{\text{TV}} &= \frac{1}{2} \sum_{i \in \Omega} |\mu(i) - \nu(i)| \\ &= \frac{1}{2} \|\mu - \nu\|_{L^1}. \end{aligned}$$

One can show that

$$\|\mu - \nu\|_{\text{TV}} = \sup\{|\mu(A) - \nu(A)| : A \subseteq \Omega\},$$

i.e. $\|\mu - \nu\|_{\text{TV}}$ is the maximum difference in the probability of any event under the two distributions.

We use the total variation distance to measure the distance between the distribution of the Markov chain at time n and the stationary distribution. The **mixing time** of the Markov chain is defined as the smallest N such that for the stationary distribution Π and any initial distributions π_0, we have

$$\|\pi_0 P^n - \Pi\|_{\text{TV}} \leq \frac{1}{4},$$

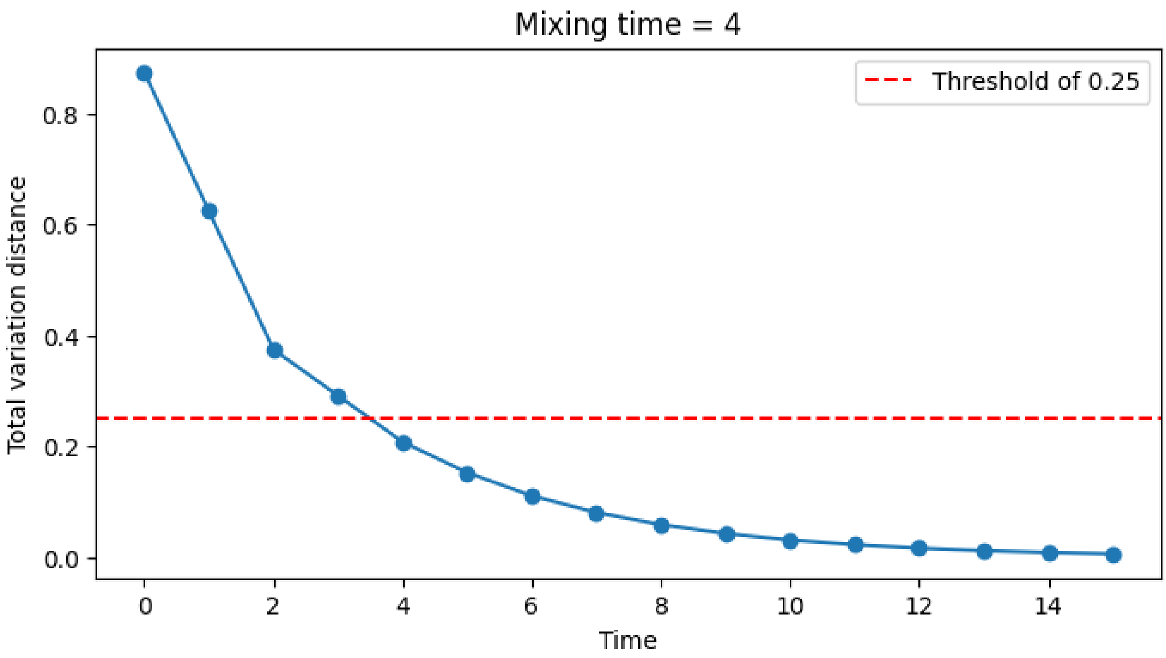
for all $n \geq N$. The constant $\frac{1}{4}$ is arbitrary and can be replaced by any other constant in $(0, 1)$. This will only change the value of the mixing time by a constant factor and not the order of magnitude.

When using Markov chains for sampling, we want the mixing time to be as small as possible. This ensures that the distribution of the chain is close to the stationary distribution after a small number of steps. We think of the time before the chain mixes as a transient phase - the chain has not yet reached equilibrium. This is a burn-in period where we discard the samples. The bigger the mixing time, the more the number of wasted samples in the burn-in phase.

Example. Consider Example 0.1 again. If we start at the vertex A, by step 4 we have already reached the distribution $\pi_4 = [0.222, 0.167, 0.306, 0.306]$. The stationary distribution is $\Pi = [1/6, 4/6, 2/6, 2/6]$. The total variation distance between π_4 and Π is $\|\pi_4 - \Pi\|_{\text{TV}} = 0.21$. This is less than $1/4$, and hence the mixing time is 4.

This only computes the mixing time for the initial distribution $\pi_0 = [1, 0, 0, 0]$. In general, we need to compute the mixing time for all possible initial distributions. The mixing time is the maximum of these mixing times.

In practice, we either provide a theoretical bound on the mixing time or “visually” inspect the convergence of the chain to the stationary distribution. Running a simulation to compute the mixing time is computationally expensive and not commonly done.



Connection to Spectral Theory

Continuing the example from above, the matrix $I_4 - P$ is called the **normalized Laplacian matrix** of the graph G.

$$\mathcal{L} = I_4 - P = \begin{pmatrix} 1 & -1 & 0 & 0 \\ -1/3 & 1 & -1/3 & -1/3 \\ 0 & -1/2 & 1 & -1/2 \\ 0 & -1/2 & -1/2 & 0 \end{pmatrix}.$$

One can show that 0 is an eigenvalue of $\mathcal{L}$ and all eigenvalues are non-negative. The second smallest eigenvalue of $\mathcal{L}$ is called the **spectral gap** of the graph (which could be 0).

Spectral graph theory, in particular Cheeger inequalities, prove that there is an inverse relationship between the spectral gap of the graph and the mixing time of the random walk on the graph. The smaller the spectral gap, the larger the mixing time. You'll explore this connection in the homework.