Notes on Finite-State Discrete-Time Markov Chains

Rose Enos

Abstract

Part of the Department of Computer Science's intelligent systems series, COMPSCI 177 introduces probability theory and several topics central to machine learning to computer science students who are not expected to have experience in a proof-based course [1]. Markov chains are defined in COMPSCI 177 with reference only to basic probability theory and linear algebra, and are motivated by historical examples from the field of computer science. A Markov chain is a series of states that evolve from an initial state according to a static transition function. We explore the properties of Markov chains and the PageRank algorithm.

0. Introduction

Part of the Department of Computer Science's intelligent systems series, COMPSCI 177 introduces probability theory and several topics central to machine learning to computer science students who are not expected to have experience in a proof-based course [1]. Among these is the finite-state discrete-time Markov chain, with applications to search systems and large language models. Markov chains are defined in COMPSCI 177 with reference only to basic probability theory and linear algebra, and are motivated by historical examples from the field of computer science.

A Markov chain is a series of states that evolve from an initial state according to a static transition function. We explore the properties of Markov chains and methods for calculating those properties. We also describe the PageRank algorithm, which relies on the theory of Markov chains to sort web search results.

For brevity, and because COMPSCI 177 is not proof based, we do not provide proofs in these notes. However, the proofs in the course text, Harchol-Balter [2], are accessible to students who are at the course level and are mostly presented informally as discussions.

1. Properties of Discrete-Time Markov Chains

A stochastic process is a sequence of random variables.

Definition 1.1 (Markov chain)**.** A discrete-time Markov chain is a stochastic process satisfying the Markovian property

$$\mathrm{P}(X_{n+1} = j \mid X_n = i, X_{n-1} = i_{n-1}, \ldots, X_0 = i_0) = \mathrm{P}(X_{n+1} = j \mid X_n = i)$$

and the stationary property

$$\mathrm{P}(X_{n+1} = j \mid X_n = i) = P_{ij}$$

The transition probability matrix P has P_{ij} representing the probability of transitioning from i to j in one step.

Definition 1.2 (Limiting probability)**.** The limiting probability, or long-run ensemble average fraction of time, of j

$$\pi_j = \lim_{n\to\infty} (P^n)_{ij}$$

is independent of i.

A distribution $\vec{\pi}'$ is stationary if

$$\vec{\pi}' \cdot P = \vec{\pi}'^T$$

Since the nth probability of a discrete distribution is determined by the other probabilities, solving for the stationary distribution requires solving the stationary equations

$$\begin{cases} \vec{\pi}' \cdot P = \vec{\pi}'^T \\ \sum_{i=0}^{M-1} \pi'_i = 1 \end{cases}$$

Theorem 1.3 (Uniqueness of the stationary distribution)**.** *If it exists, the limiting distribution is the unique stationary distribution*

$$\vec{\pi} = (\pi_0, \ldots, \pi_{M-1})$$

A Markov chain is stationary, or steady state, if the initial state is the chosen according to a stationary distribution. Every chain has at least one stationary distribution.

The stationary equations may be computationally expensive to solve. The balance equations equate the input and output rate of each state:

$$\begin{cases} \sum_{j\neq i} \pi_i P_{ij} = \sum_{j\neq i} \pi_j P_{ji} \\ \sum_i \pi_i = 1 \end{cases}$$

Solving the balance equations gives a stationary distribution. The time-reversibility equations equate the pairwise transition rates:

$$\begin{cases} \pi_i P_{ij} = \pi_j P_{ji} \\ \sum_i \pi_i = 1 \end{cases}$$

Solving the time-reversibility equations gives a stationary distribution. Then the chain is time reversible.

2. Ergodicity of Finite-State Markov Chains

Definition 2.1 (Period)**.** The period of j is

$$\gcd\{n \in \mathbb{N} : (P^n)_{jj} > 0\}$$

j is aperiodic if its period is 1. A chain is aperiodic if all states are aperiodic. The Euclidean number property states that, if $\gcd\{i_1, \ldots, i_k\} = 1$, then every $n \geq n_0$ is representable as an $\mathbb{N}_0$-linear combination of $i_1, \ldots, i_k$.

j is accessible from i if $(P^n)_{ij} > 0$. i and j communicate if they are mutually accessible. j is recurrent if it communicates with all accessible states, and is otherwise transient. j is absorbing if $P_{ii} = 1$. A recurrent class is a set of pairwise communicating states.

Definition 2.2 (Irreducibility)**.** A chain is irreducible if it has only one recurrent class.[1]

The proof of 2.3 relies on the following two results: First, iff $M_n = \max P^n \hat{e}$, $m_n = \min P^n \hat{e}$, and $s = \min P$, then

$$M_n - m_n \leq (1 - 2s)(M_{n-1} - m_{n-1})$$

Second, a chain is aperiodic and irreducible if and only if $\dim P^n \hat{e} = \dim P^{n-1} \hat{e}$ for all $n \geq n_0$.

Theorem 2.3 (Ergodicity)**.** *The limiting distribution exists if and only if the chain is aperiodic and irreducible. Then the chain is ergodic.*

3. More on Aperiodicity and Irreducibility

A statement is true with probability 1 if it is true on almost every sample path. The strong law of large numbers states that, if $X_1, \ldots, X_n$ are independent and identically distributed, then, with probability 1,

$$\lim_{n \to \infty} \frac{\sum_{i=1}^{n} X_i}{n} = \mathrm{E}(X)$$

A renewal process is a process such that the times between events are independent and identically distributed. The renewal theorem states that, with probability 1,

$$\lim_{t \to \infty} \frac{N(t)}{t} = \frac{1}{\mathrm{E}(X)}$$

In a recurrent class, the expected number of steps m_{ij} between i and j is finite, and the unique stationary distribution has

$$\pi'_j = \frac{1}{m_{jj}}$$

The long-run time-average fraction of time in state j on a random walk is

$$p_j = \lim_{t \to \infty} \frac{N_j(t)}{t}$$

With probability 1,

$$p_j = \frac{1}{m_{jj}}$$

If the chain is aperiodic, then, with probability 1,

$$p_j = \pi'_j$$

[1] Irreducibility is defined differently by different authors. For instance, our definition is from the course lectures [3], since it allows for the "if and only if" statement in 2.3. Harchol-Balter [2] defines irreducibility to mean that the entire chain is a recurrent class; ours requires only that there be exactly one recurrent class in the chain.

4. Entropy Rates of Markov Chains

The entropy rate of a stochastic process is

$$\bar{H} = \lim_{n\to\infty} \frac{1}{n} H(X_1, \ldots, X_n)$$

If a stationary distribution exists, then the entropy rate exists as

$$\bar{H} = \lim_{n\to\infty} H(X_n \mid X_{n-1}, \ldots, X_1)$$

For a Markov chain, from the Markovian property,

$$\bar{H} = \lim_{n\to\infty} H(X_n \mid X_{n-1}) = \sum_{i=1}^{m} h_i \pi_i$$

where

$$h_i = H(X_n \mid X_{n-1} = i)$$

The entropy rate represents the average number of bits per time step to compress a sequence sampled from the chain or to simulate the chain [3].

5. Applications: PageRank and Language Models

Google's first web search algorithm, PageRank, ranks webpages by their stationary probability based on forward hyperlink navigation. To make the chain irreducible, we scale the link probabilities down and use the excess to link to every other page. Since a web graph is sparse, estimating the limiting distribution by matrix multiplication is more efficient than solving for the exact distribution. Thus, modern search engines apply these and other techniques to rank pages, rather than relying on pure Markov chains. Nonetheless, Markov chains provide a framework within which to understand the basics of web search.

We can consider a Markov chain representing the sequence of natural language units (e.g. characters or words) [3]. Under our model, the chain gives the most likely sequence of words, starting from some initial word. We can extend our model to consider more than one previous state in its calculations, yielding a word sequence that may make more sense to a human reader. As we extend the model to consider more words, we get more coherent sequences—sentences!—but vastly higher computational cost. Modern techniques for large language models use other machine learning techniques to simulate the results of a Markov chain while avoiding the computational cost. Still, Markov chains again provide the basis for intuitive understanding.

References

[1] UCI General Catalogue. Compsci 177, 2025.

[2] Mor Harchol-Balter. *Introduction to Probability for Computing.* Cambridge University Press, 2024.

[3] Erik B. Sudderth. Compsci 177. Course lectures at the University of California, Irvine.