San José State University

Math 263: Stochastic Processes

Markov Chains

Dr. Guangliang Chen

This lecture is based on the following textbook sections:

- Sections 4.1, 4.2

Outline of the presentation

- Markov chain concepts
- t-step transition probabilities

HW2: Assigned in Canvas

Recall that a stochastic process is a family of random variables, $\{X_t, t \in T\}$, where X_t measures, at time t, the aspect of a system which is of interest.

The process is called a

- discrete-time process when T is a discrete set such as $\mathbb{Z}, \mathbb{Z}^+, \mathbb{Z}_0^+$, or
- continuous-time process when T is an interval such as $\mathbb{R}^+$ or $\mathbb{R}_0^+$.

In this lecture, we focus on discrete-time, integer-valued processes and will introduce a **probability model** for the collection of random variables X_t.

Basic concepts

Let $\{X_n, n = 0, 1, 2, \ldots\}$ be a stochastic process that takes a countable number of possible values, which is assumed to be $\mathbb{Z}$ for convenience.

If $X_n = i$, then we say that the process is in state i (at time n).

Given the current time and state, $X_n = i$, and the history of the process,

$$X_{n-1} = i_{n-1}, \quad \ldots, \quad X_1 = i_1, \quad X_0 = i_0$$

the process will move to state j at time $n+1$, with **transition probability**

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

Def 0.1. A discrete-time, integer-valued stochastic process $\{X_n\}_{n\geq 0}$ is called a **Markov chain**, or said to have the **Markov property**, if the conditional distribution of any future state X_{n+1}, given the past states $X_0, X_1, \ldots, X_{n-1}$ and the present state X_n, is independent of the past states and depends only on the current state:

$$\begin{aligned} &P(X_{n+1} = j \mid X_n = i,\ X_{n-1} = i_{n-1}, \ldots, X_1 = i_1, X_0 = i_0) \\ =&P(X_{n+1} = j \mid X_n = i) \end{aligned}$$

for all $n \geq 0$ and all $i_0, i_1, \ldots, i_{n-1}, i, j \in S$.

A Markov chain is said to be **time-homogeneous**, or **stationary**, if the transition probabilities from one state to another state are independent of time, i.e.,

$$P(X_{n+1} = j \mid X_n = i) = \underbrace{P(X_1 = j \mid X_0 = i)}_{p_{ij}}$$

for all $n \geq 0$ and all $i, j \in S$.

Unless other specified, Markov chains are assumed to be time-homogeneous.

Given a Markov chain, the transition probabilities form a matrix, $\mathbf{P} = (p_{ij})$, called the **transition matrix**.

For example, when $S = \{0, 1, \ldots, N\}$, the transition matrix has the form

$$\mathbf{P} = \begin{bmatrix} p_{00} & p_{01} & \cdots & p_{0N} \\ p_{10} & p_{11} & \cdots & p_{1N} \\ \vdots & \vdots & \ddots & \vdots \\ p_{N0} & p_{N1} & \cdots & p_{NN} \end{bmatrix}$$

Transition matrices must be

- **nonnegative**:

$$p_{ij} \geq 0 \quad \text{for all } i, j \in S$$

- **row-stochastic** (i.e., row sums are all 1):

$$\sum_j p_{ij} = 1, \quad \text{for all } i \in S$$

The row-stochastic property can be represented in matrix notation

$$\mathbf{P1} = \mathbf{1}, \qquad \text{where} \quad \mathbf{1} = (1, 1, \ldots, 1)^T.$$

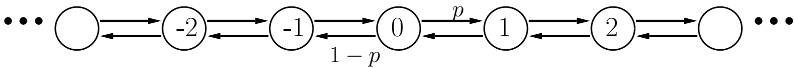
This shows that $\mathbf{1}$ is an eigenvector of $\mathbf{P}$ corresponding to eigenvalue 1.

Def 0.2 (Random walk). A Markov chain $\{X_n, n \geq 0\}$ with state space $S = \mathbb{Z}$ is called a **random walk** if it has the following transition probabilities

$$p_{i,i+1} = p = 1 - p_{i,i-1}, \quad \text{for all } i, j \in S$$

that is,

$$p_{ij} = \begin{cases} p, & j = i+1 \\ 1-p & j = i-1 \\ 0, & \text{otherwise} \end{cases}, \quad \text{for all } i \in S$$

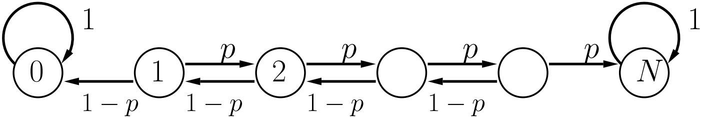


It is called a symmetric random walk if $p = \frac{1}{2}$.

Example 0.1 (Gambler's Ruin)**.** This can be modeled as a Markov chain with state space $S = \{0, 1, 2, \ldots, N\}$ and transition probabilities

$$P_{i,i+1} = p = 1 - P_{i,i-1}, \quad 1 \le i \le N-1$$

$$P_{00} = 1 = P_{NN} \quad \text{(absorbing states)}$$



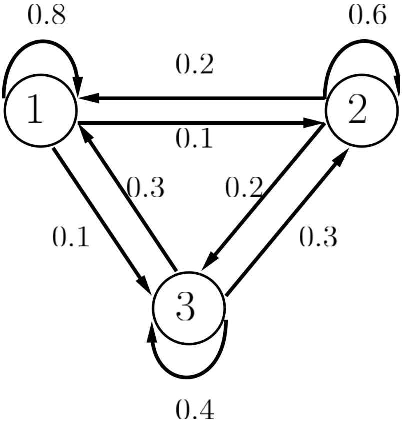
It is a random walk on a finite state space and with two absorbing barriers.

The transition matrix of such a process has the following form (when $N = 5$):

$$\mathbf{P} = \begin{bmatrix} 1 & & & & & \\ 1-p & 0 & p & & & \\ & 1-p & 0 & p & & \\ & & 1-p & 0 & p & \\ & & & 1-p & 0 & p \\ & & & & & 1 \end{bmatrix}$$

Example 0.2 (Social mobility). Let X_n be a family's social class: 1 (lower), 2 (middle), 3 (upper) in the nth generation. We can model this process as a Markov chain with certain kind of transition probabilities such as

$$\mathbf{P} = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.2 & 0.6 & 0.2 \\ 0.3 & 0.3 & 0.4 \end{pmatrix}$$



Application to clustering

Markov chains can be used for data clustering, which is an unsupervised learning task in machine learning. Its informal formulation is the following.

Problem 0.3. Given a set of objects and a similarity measure, partition the data set into k disjoint subsets (i.e., clusters) such that

- objects in the same cluster are similar to each other;
- objects in different clusters are generally not similar.

We often represent such information via an undirected, weighted graph, called similarity graph:

Accordingly, **clustering is equivalent to graph partitioning.**

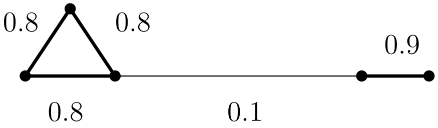
Remark. An undirected, weighted graph $\mathscr{G} = (V, E, \mathbf{W})$ is a mathematical object that has the following components:

- vertex set $V = \{v_1, \ldots, v_n\}$
- edge set $E = \{e_{ij}\}$
- weight matrix $\mathbf{W} = (w_{ij})$

Note that an edge exists between two vertices i, j if and only if $w_{ij} > 0$.

Remark. A similarity graph is uniquely defined by a given weight matrix.

$$\mathbf{W} = \begin{pmatrix} & 0.8 & 0.8 & & \\ 0.8 & & 0.8 & & \\ 0.8 & 0.8 & & 0.1 & \\ & & 0.1 & & 0.9 \\ & & & 0.9 & \end{pmatrix}$$



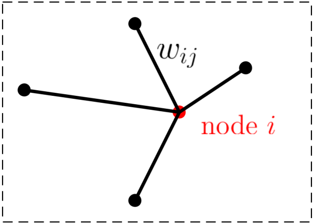
Given an undirected, weighted graph $\mathscr{G} = (V, E, \mathbf{W})$, define

- the **degree** of a single vertex $v_i \in V$:

$$d_i = \sum_{j \in V} w_{ij}$$

- and also the **degree matrix**:

$$\mathbf{D} = \operatorname{diag}(d_1, \ldots, d_n) \in \mathbb{R}^{n \times n}$$
$$= \operatorname{diag}(\mathbf{W1}).$$



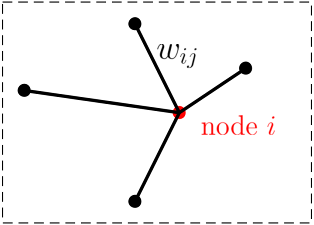
Note that d_i measures the connectivity of node i in the graph: The larger the degree, the more strongly connected the node.

For example, the degree matrix associated with the previous graph is

$$\mathbf{W} = \begin{pmatrix} & 0.8 & 0.8 & & \\ 0.8 & & 0.8 & & \\ 0.8 & 0.8 & & 0.1 & \\ & & 0.1 & & 0.9 \\ & & & 0.9 & \end{pmatrix} \longrightarrow \mathbf{D} = \begin{pmatrix} 1.6 & & & & \\ & 1.6 & & & \\ & & 1.7 & & \\ & & & 1.0 & \\ & & & & 0.9 \end{pmatrix}$$

Now, suppose that a person initially stands on some vertex of the graph (say $X_0 = i$) and moves from vertex to vertex along the edges randomly according to the following transition probabilities:

$$p_{ij} = \frac{w_{ij}}{d_i}, \quad \text{for all (connected) nodes } j \in V.$$



Remark. Let $\mathbf{P} = (p_{ij})$. Then

- $\mathbf{P} = \mathbf{D}^{-1}\mathbf{W}$,
- $\mathbf{P}$ is nonnegative ($\mathbf{P} \geq 0$),
- $\mathbf{P}$ is row-stochastic ($\mathbf{P}\mathbf{1} = \mathbf{1}$).

Let X_n be the location of the person in the graph after n steps.

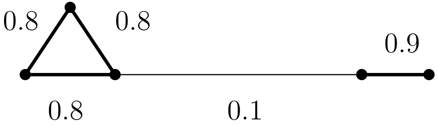
Then $\{X_n : n = 0, 1, 2, \ldots\}$ is a Markov chain with

- state space $S = V$, and
- transition matrix $\mathbf{P} = (p_{ij})$.

Under this model, **clusters are subsets of states where one spends a long time in each of them and seldom jumps between them**.

In the toy example, the state space of the Markov chain is $S = \{1,2,3,4,5\}$ and the transition matrix is

$$\mathbf{W} = \begin{pmatrix} & 0.8 & 0.8 & & \\ 0.8 & & 0.8 & & \\ 0.8 & 0.8 & & 0.1 & \\ & & 0.1 & & 0.9 \\ & & & 0.9 & \end{pmatrix} \longrightarrow \mathbf{P} = \begin{pmatrix} & \frac{1}{2} & \frac{1}{2} & & \\ \frac{1}{2} & & \frac{1}{2} & & \\ \frac{8}{17} & \frac{8}{17} & & \frac{1}{17} & \\ & & \frac{1}{10} & & \frac{9}{10} \\ & & & 1 & \end{pmatrix}$$



t-step transition probabilities

Consider a Markov chain with state space S. For any $i, j \in S$ and $t \geq 0$, the t-step transition probability from i to j is defined as

$$p_{ij}^{(t)} = P(X_t = j \mid X_0 = i).$$

Define also the t-step transition matrix

$$\mathbf{P}^{(t)} = \left(p_{ij}^{(t)}\right).$$

Clearly, $\mathbf{P}^{(1)} = \mathbf{P}$ (one-step transition matrix). What about $t \geq 2$?

Theorem 0.1 (**Chapman-Kolmogorov Equations**). For any $n, m \in \mathbb{Z}_0^+$,

$$p_{ij}^{(n+m)} = \sum_{k\in S} p_{ik}^{(n)} p_{kj}^{(m)}, \quad i, j \in S.$$

This implies that

$$\mathbf{P}^{(n+m)} = \mathbf{P}^{(n)} \mathbf{P}^{(m)}.$$

Proof. By the law of total probability,

$$\begin{aligned} p_{ij}^{(n+m)} &= P(X_{n+m} = j \mid X_0 = i) \\ &= \sum_{k\in S} P(X_{n+m} = j \mid X_n = k, \cancel{X_0 = i}) P(X_n = k \mid X_0 = i) \\ &= \sum_{k\in S} p_{kj}^{(m)} p_{ik}^{(n)}. \end{aligned}$$

By mathematical induction, we can obtain the following result.

Corollary 0.2. $\mathbf{P}^{(t)} = \mathbf{P}^t$ for any integer $t \geq 1$.

Remark. $\mathbf{P}^{(t)}$ is also nonnegative and row-stochastic:

$$\mathbf{P}^{(t)}\mathbf{1} = \mathbf{P}^t\mathbf{1} = \underbrace{\mathbf{P}\cdots\mathbf{P}\cdot(\mathbf{P}}_{t \text{ copies}}\cdot\mathbf{1}) = \underbrace{\mathbf{P}\cdots\mathbf{P}}_{t-1 \text{ copies}} \cdot \mathbf{1} = \mathbf{1}$$

Example 0.4 (Social mobility, cont'd)**.**

$$\mathbf{P}^2 = \begin{pmatrix} .69 & .17 & .14 \\ .34 & .44 & .22 \\ .42 & .33 & .25 \end{pmatrix}, \quad \mathbf{P}^3 = \begin{pmatrix} .628 & .213 & .159 \\ .426 & .364 & .210 \\ .477 & .315 & .208 \end{pmatrix}$$

How to interpret them?

Example 0.5 (modified from Example 4.10, page 198)**.** An urn always contains 2 balls. Ball colors are red and blue. At each stage a ball is randomly chosen and then replaced by a new ball, which with probability 0.8 is the same color as the ball it replaces, and with probability 0.2 is the opposite color. If initially both balls are red, find the probability that the third ball selected is red.

Solution. Let X_n be the number of red balls in the urn after n steps (and $X_0 = 2$). Clearly, $\{X_n : n = 0, 1, 2, \ldots\}$ is a Markov chain with state space $S = \{0, 1, 2\}$ and transition matrix

$$\mathbf{P} = \begin{pmatrix} 0.8 & 0.2 & 0 \\ 0.1 & 0.8 & 0.1 \\ 0 & 0.2 & 0.8 \end{pmatrix} \quad \longrightarrow \quad \mathbf{P}^2 = \begin{pmatrix} 0.66 & 0.32 & 0.02 \\ 0.16 & 0.68 & 0.16 \\ 0.02 & 0.32 & 0.66 \end{pmatrix}$$

Let R_3 denote the event that the third selected ball is red. Then

$$\begin{aligned} P(R_3 \mid X_0 = 2) &= \sum_{i=0}^{2} P(R_3 \mid X_2 = i, \cancel{X_0 = 2}) P(X_2 = i \mid X_0 = 2) \\ &= \sum_{i=0}^{2} \frac{i}{2} \cdot p_{2,i}^{(2)} = 0 \cdot 0.02 + 0.5 \cdot 0.32 + 1 \cdot 0.66 = 0.82. \end{aligned}$$

Example 0.6 (Example 4.11, page 199)**.** Suppose that balls are successively distributed among 8 urns, with each ball being equally likely to be put in any of these urns. What is the probability that there will be exactly 3 nonempty urns after 9 balls have been distributed?

Solution. Let X_n be the number of occupied urns after n steps. Clearly, $\{X_n : n = 1, 2, \ldots\}$ is a Markov chain with state space $S = \{1, 2, \ldots, 8\}$ and transition probabilities

$$p_{i,i} = \frac{i}{8}, \quad i = 1, \ldots, 8$$
$$p_{i,i+1} = 1 - \frac{i}{8}, \quad i = 1, \ldots, 7$$

The resulting transition matrix is

$$\mathbf{P} = \begin{pmatrix} 1/8 & 7/8 & & & & & & \\ & 2/8 & 6/8 & & & & & \\ & & 3/8 & 5/8 & & & & \\ & & & 4/8 & 4/8 & & & \\ & & & & 5/8 & 3/8 & & \\ & & & & & 6/8 & 2/8 & \\ & & & & & & 7/8 & 1/8 \\ & & & & & & & 1 \end{pmatrix}$$

The desired probability is

$$p_{13}^{(8)} = .00756 \text{ (by software).} \quad \square$$

Marginal distribution of X_n

To compute the marginal distributions of X_n, we need to be given the initial distribution of the chain:

$$\boldsymbol{\alpha} = (\alpha_i)_{i \in S}, \quad \alpha_i = P(X_0 = i)$$

where $\boldsymbol{\alpha}$ is a row vector.

Remark. If the initial state of a chain is fixed (say i), then $\boldsymbol{\alpha} = \mathbf{e}_i$.

Theorem 0.3. The marginal distribution of X_n is given by $\boldsymbol{\alpha}\mathbf{P}^n$ (over S).

Proof. For any j,

$$P(X_n = j) = \sum_i P(X_n = j \mid X_0 = i) P(X_0 = i) = \sum_i p_{ij}^{(n)} \alpha_i.$$

which is just the matrix product of $\boldsymbol{\alpha}$ and the jth column of $\mathbf{P}^n$. □

Remark. If $\boldsymbol{\alpha} = \mathbf{e}_i$ for some i (i.e., the chain always starts from state i), then the marginal distribution of X_n is given by the ith row of $\mathbf{P}^n$.

Example 0.7 (Social mobility, cont'd). Suppose the initial distribution of the chain is $\boldsymbol{\alpha} = (\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$. Then the distribution of X_2 (social status after two generations) is

$$\boldsymbol{\alpha}\mathbf{P}^2 = \left(\frac{1}{3}, \frac{1}{3}, \frac{1}{3}\right) \begin{pmatrix} .69 & .17 & .14 \\ .34 & .44 & .22 \\ .42 & .33 & .25 \end{pmatrix} = (0.4833, 0.3133, 0.2033).$$

Example 0.8 (2 balls, change color, cont'd). If the initial distribution of the number of red balls is $\boldsymbol{\alpha} = (\frac{1}{4}, \frac{1}{2}, \frac{1}{4})$, what is the probability that the third ball selected is red?

Solution. The marginal distribution of X_2 is

$$\boldsymbol{\alpha}\mathbf{P}^2 = \left(\frac{1}{4}, \frac{1}{2}, \frac{1}{4}\right) \begin{pmatrix} 0.66 & 0.32 & 0.02 \\ 0.16 & 0.68 & 0.16 \\ 0.02 & 0.32 & 0.66 \end{pmatrix} = (0.25, 0.5, 0.25).$$

It follows that

$$P(R_3) = \sum_{i=0}^{2} P(R_3 \mid X_2 = i) P(X_2 = i) = 0 \cdot 0.25 + 0.5 \cdot 0.5 + 1 \cdot 0.25 = 0.5.$$

Example 0.9 (Example 4.11, page 199). In a sequence of independent flips of a fair coin, let N denote the number of flips until there is a run of three consecutive heads, for example,

$$T \quad H \quad H \quad T \quad T \quad H \quad T \quad H \quad H \quad H$$

Find $P(N \leq 8)$ and $P(N = 8)$.

Solution. Let X_n be the number of consecutive heads at the end of the sequence from flipping a fair coin n times (and suppose the game has not ended earlier).

After the game has ended, we will just let $X_n = 3$ for all n.

Then $\{X_n, n = 0, 1, 2, \ldots\}$ is a Markov chain with state space $S = \{0, 1, 2, 3\}$, where state i means that we are currently on a run of i consecutive heads (and if $i = 3$, the experiment would just end).

The transition matrix is

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

From this, we obtain (by software) that

$$P(N \le 8) = P(X_8 = 3) = p_{03}^{(8)} = \frac{107}{256}$$

$$P(N = 8) = p_{02}^{(7)} \cdot p_{23} = \frac{13}{256}$$