Wednesday, December 5 **Fall 2023**

Section 11: Markov Chains

Srihari Ganesh

Based on section note formatting template by Rachel Li and Ginnie Ma '23

Forms

- Attendance form: http://bit.ly/110attend
- Feedback form: https://bit.ly/SrihariFeedback

Attendance form

Section feedback form

Logistics

- Make sure you check the special end-of-term schedule for December 6-12. Specifically,
  - No office hours from me this week.
  - I will be hosting extra exam OH on Monday, December 11 from 8-10pm in Quincy Dining Hall.
  - Go to the PPPPP tonight from 6-9pm! Good opportunity to practice under testing conditions.
- Always feel free to message me. I'm happy to meet in person through one-on-ones/small-group/extra OH time if you want to ask lots of questions.

1 Summary

1.1 Basic Markov chain definitions and results

Definition 1 (Markov chain)**.** A sequence of random variables $X_0, X_1, X_2, \ldots$ is said to have the **Markov property** if

$$P(X_{j+1} = x_{j+1} | X_0 = x_0, X_1 = x_1, \ldots, X_j = x_j) = P(X_{j+1} = x_{j+1} | X_j = x_j).$$

If we consider to be a series over time, then if we condition on a set of timepoints, we only need to keep the most recent information — the rest is redundant (i.e., conditionally independent) given more recent information. Note that this does NOT mean that X_{j+1} is independent of $X_0, \ldots, X_{j-1}$, but it is conditionally independent of those given X_j.

A sequence with the Markov property is said to be a **Markov chain**.

Definition 2 (States and state space)**.** The support of the X_j is called the **state space**: formally, it is $\bigcup_{i=1}^{\infty} \text{support}(X_i)$. We will often refer to the value of a Markov chain at a certain time as its current **state**. In Stat 110, we will only discuss finite state spaces, usually the integers $\{1, \ldots, M\}$ (M is not random, just a constant).

Definition 3 (Time homogeneity)**.** A Markov chain is said **time-homogeneous** if

$$P(X_{j+1} = x_{j+1} | X_j = x_j) = P(X_1 = x_{j+1} | X_0 = x_j)$$

for all j. That is, the PMF of the next state only depends on the current state, but not the current time. All Markov chains in Stat 110 will be time-homoegeneous.

Definition 4 (Transition matrix)**.** A time-homogeneous Markov chain with M states (using the state space $\{1, \ldots, M\}$) can be represented by the $M \times M$ **transition matrix** Q, where

$$Q_{ij} = P(X_2 = j | X_1 = i)$$

for any $1 \leq i, j \leq M$. Note that the each row of a transition matrix must sum to 1 since $\sum_j P(X_2 = j | X_1 = i)$ is a sum over the conditional PMF of the next state given the current state is i.

Notation 5 (PMF as a vector)**.** For a discrete random variable X with support $\{1, \ldots, M\}$, we can represent the PMF as a vector $\vec{t}$:

$$\vec{t} = \begin{pmatrix} P(X = 1) & \cdots & P(X = M). \end{pmatrix}$$

So we can think of each row of the transition matrix as the PMF of the next state given the current state. Note that when discussing Markov chains in Stat 110 we'll use *row vectors* for the PMFs.

Remark 6 (Initial state)**.** While the transition matrix tells us the PMF of the next state given the current state, we need to start from somewhere — usually the first state, X_0. X_0 can be a constant, or have its own PMF defined completely separately from the chain itself, which can affect the long-term behavior of the chain.

Result 7 (n-step transition probability)**.** The **n-step transition probabilities** $P(X_{n+k} = j | X_k = i)$, are given by Q^n; i.e.,

$$Q^n_{ij} = P(X_{n+k} = j | X_k = i).$$

If the PMF of the initial state X_0 is $\vec{t}$, the unconditional PMF of X_n (the **n-step distribution**) is $\vec{t}Q^n$.

Definition 8 (Irreducible chain)**.** A chain is **irreducible** if it is possible to (eventually) reach any state j from any other state i. That is, there exists some n such that $Q_{ij}^n > 0$.

Definition 9 (Recurrent state)**.** A state i is **recurrent** if, when we start at state i, the probability of returning to state i is 1. In irreducible chains, all states are recurrent.

Definition 10 (Transient state)**.** Conversely, a state i is **transient** if there is a nonzero probability of never returning to state i. If we call this probability p, then the number of returns to state i (before never returning) is distribued $\text{Geom}(p)$.

Definition 11 (Periodicity)**.** On the right-hand chain of Figure 1, we can see that to return to state 1, we must take 3 steps. The **period** of a state is the greatest common denominator of the possible amounts of steps one must take to return to the state. A state is **aperiodic** if its period is 1; a chain is **aperiodic** if all *states* are aperiodic.

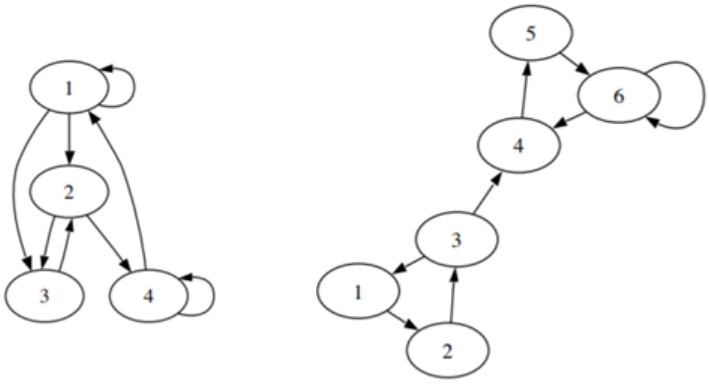


Figure 1: Blitzstein & Hwang, Figure 11.2

Definition 12 (Reversibility)**.** Let $\vec{s}$ represent a valid PMF as a row vector. A Markov chain with transition matrix Q is said to be **reversible** if for all i, j, if

$$s_i Q_{ij} = s_j Q_{ji}.$$

Remark 13 (Graphical representation of Markov chains)**.** Figure 1 shows two graphical representations of Markov chains. The **nodes** (circles with numbers in them) are the possible states in the chain, while the **edges** (arrows) show which transitions between states are possible. Edges can **weights** (numbers) next to them, either representing the probability of the denoted transition or a number proportional to it. If the edges have no weights, then you should assume that every outgoing edge from a certain state has equal probability. For example, on the left-hand graph in Figure 1, state 2 has two outgoing edges, $P(X_{j+1} = 3|X_j = 2) = P(X_{j+1} = 4|X_j = 2) = \frac{1}{2}$. If an edge is not drawn then that transition probability is 0; i.e., $P(X_{j+1} = 1|X_j = 1) = 0$. If edges do not have arrows (or have arrows at both ends), then you can assume that both transition directions are possible.

Result 14 (Random walk on an undirected network)**.** We can represent some Markov chains as **undirected graphs** with uniform edge weights; in our graphical representation of Markov chains, that means edges won't have weights drawn and will not have directional arrows. This means that

1. Every positive term of the i-th row of Q is equal (for all i): valid examples of rows for a 4-state chain include $(0, 1, 0, 0)$, $(\frac{1}{3}, 0, \frac{1}{3}, \frac{1}{3})$, $(\frac{1}{4}, \frac{1}{4}, \frac{1}{4}, \frac{1}{4})$.

2. For all i, j, if $Q_{ij} > 0$, then $Q_{ji} > 0$.

We define the **degree sequence** as the number of edges attached to each node; the i-th entry is the number of positive terms in the i-th row of Q. Note that self-edges are included in this count. These chains are reversible.

Result 15 (Birth-death chain)**.** A **birth-death** chain is a Markov chain where each step moves at most one position to the left or right: e.g., if the current state is 5, then the next state must be 4, 5, or 6. Any birth-death chain is reversible.

1.2 Stationary distributions and how to find them

Definition 16 (Stationary distribution)**.** Let $\vec{s} \in \mathbb{R}^M$ be a row vector that is valid PMF for a Markov chain with M states. Then $\vec{s}$ is **stationary** if

$$\vec{s}Q = \vec{s}.$$

Result 17 (Irreducible chain/expected return time)**.** Irreducible Markov chains have unique stationary distributions; every state has positive probability in this distribution. If this stationary distribution is $\vec{s}$, the expected time to return to state i when starting at state i is $\frac{1}{s_i}$.

Result 18 (Columns sum to 1)**.** If each column of the transition matrix Q sums to 1, then the stationary distribution is uniform.

Result 19 (Irreducible and aperiodic convergence)**.** A chain is **irreducible and aperiodic** if and only if there exists some power of the transition matrix Q^m with all entries positive. Such chains **converge** to their unique stationary distribution.

Result 20 (Reversible chain)**.** If a Markov chain with transition matrix Q is reversible with respect to $\vec{s}$, then $\vec{s}$ is stationary. The converse is not necessarily true: reversible implies stationary, but stationary does not imply reversible.

Result 21 (Random walk on an undirected network)**.** If a Markov chain can be represented as an undirected graph with uniform edge weights, then the stationary distribution is proportional to the degree sequence. It is also shown in Blizstein & Hwang (Exercise 11.20) that for an undirected *weighted* graph, that the stationary distribution is proportional to the sum of edge weights connected to each node, $s_i \propto \sum_j w_{ij}$, where w_{ij} are the edge weights.

Result 22 (Birth-death chain)**.** You can find the stationary distribution of a birth-death chain by doing the algebra.

1. Solve for s_2 in terms of s_1 using the reversibility condition.
2. Solve for s_3 in terms of s_2 using reversibility. Then plug the previous step to get s_3 in terms of s_1.
3. Continue this way until you have all s_i in terms of s_1.
4. Since the stationary distribution must be a valid PMF, $\sum_i s_i = 1$. Plug in all values in terms of s_1 and solve for s_1, which thus gives you the full stationary distribution.

Remark 23. If you are familiar with linear algebra, $\vec{s}$ is a left-eigenvector of Q with eigenvalue 1. In Stat 110, you will never have to solve such a matrix equation explicitly. Instead, if asked to find the stationary distribution of a complicated-looking Markov chain, try to use one of the results above

(e.g., show the Markov chain matches one of the special cases above for which we are given the result).

1.3 Metropolis-Hastings

One of the most important uses of Markov chains is for sampling complicated distributions. This general class of sampling methods is called **Markov Chain Monte Carlo (MCMC)**, of which we'll introduce a fundamental example:

Algorithm 24 (Metropolis-Hastings)**.** Suppose we want to draw samples from a distribution with PMF (in vector form) $\vec{s} = (s_1, \ldots, s_M)$ over support $\{1, \ldots, M\}$ where all $s_i > 0$. Suppose $P = (p_{ij})$ is an $M \times M$ transition matrix over the state space $\{1, \ldots, M\}$. We construct a Markov chain $X_0, X_1, \ldots$ in the following way:

1. Start at X_0 (use any procedure to set the initial state, as long as it is in the support $\{1, \ldots, M\}$).
2. At iteration n, suppose we have $X_n = i$. Propose a new state by sampling $j \sim \text{Cat}(p_{i1}, \ldots, p_{iM})$, where Cat represents a **categorical** distribution over states $\{1, \ldots, M\}$ where state j have probability p_{ij} of being selected.
3. Compute **acceptance probability**
$$a_{ij} = \min\left(\frac{s_j p_{ij}}{s_i p_{ij}}, 1\right).$$
4. Flip a coin that lands Heads with probability a_{ij}.
5. If the coin lands Heads, accept the proposal by setting $X_{n+1} = j$. If the coin lands Tails, reject the proposal by setting $X_{n+1} = i$.
6. Repeat steps 2–5 for each iteration.

The big result is that the Markov chain $X_0, X_1, \ldots$, is **reversible** with **stationary distribution** $\vec{s}$.

1.3.1 Big-picture remarks (not necessary for Stat 110)

Metropolis-Hastings can be used to sample from complicated distributions over massive state spaces as long as we have

1. An easy-to-sample distribution over the state space.
2. Possibly unnormalized probabilities $\vec{t} = (t_1, \ldots, t_M)$

   This is allowed since $\vec{s}$ is only used as a ratio (in step 2) as s_j/s_i. So if you have $\vec{t} = c\vec{s}$, then $t_j/t_i = cs_j/cs_i = s_j/s_i$.

   The distinction between $\vec{s}$ and $\vec{t}$ is that $\sum_i s_i = 1$ (normalized) but $\sum_i t_i$ can be anything (unnormalized). Yes, there do exist many distributions for which we can easily get unnormalized probabilities, but not normalized probabilities (for example, many energy-based Boltzmann distributions that you might hear about in statistical thermodynamics).

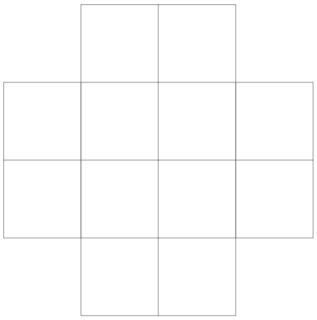
MCMC methods like Metropolis-Hastings still have many problems and are an active area of research: for example,

- They can be slow to converge to the stationary distribution and it may be difficult to tell how long to wait.

- Samples between adjacent steps can be highly dependent (especially with a bad proposal distribution), so you might have to throw out intermediate samples if you want i.i.d. samples from the desired distribution $\vec{s}$.

2 Practice Problems

1. **Chess** (*Extension of Ben Banavige and Jeremy Welborn, 2016*).
   A chess piece is moving randomly on the modified chess board below. In one move, the piece can move right, left, up or down (unless on an edge), with equal probabilities for each legal move.

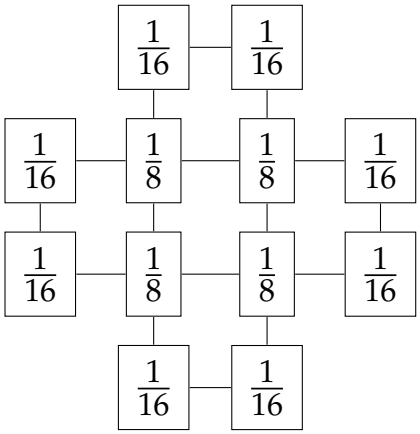


   (a) How can we think of this board as a Markov chain? Specifically, what are the states and the transition probabilities?

   (b) Is this chain irreducible? Is it aperiodic?

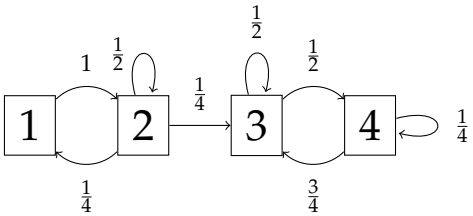
(c) At some step far in the future, what is probability of the piece being at each square in the grid?

(d) Consider a distribution on the chess board given by the PMF below. Show that the Markov chain given by the piece's position is reversible with respect to this distribution.



(e) Suppose your piece starts at the first square in the second row. What is the expected time for the piece to return to that square?

2. Consider the Markov chain below.



(a) Suppose the Markov chain starts at state 2. What is the expected number of times that the chain is at state 2 (including the initial state)?

(b) Suppose the Markov chain starts at state 3. What is the expected number of returns to state 3?

(c) What is the stationary distribution of this chain?