1 Markov Chain Monte Carlo

Early in the course, we used a Monte Carlo simulation to estimate π. That was pretty cool, who knew π is really what people always told you it was and they weren't lying all along. We used the fact that there is a standard Python command that simulates sampling a point from a uniform distribution. But what if I asked you to sample a point according to some complicated density or distribution? Chances are you'd have no idea how to go about it, and not for nothing - finding an algorithm that simulates a given distribution is out of reach for contemporary computers - even just finding a normalizing factor would often take a long time even on a very powerful machine. But we can approximate a distribution with a Markov chain.

Coming up with a Markov chain that gives you a specific distribution is, again, out of our reach. If you have a probability vector, finding a suitable transition matrix is quite a problem. However, if we have another transition matrix, with some other stationary distribution we may be able to modify it to suit our needs. Here's the intuitive idea.

Suppose you're at the bus terminal at Boston South Station and the departures board is broken, so you don't now what buses are departing in the future. You do however know how many buses a day depart to each destination so you know if they're common or rare. You have a set of possible destinations with varying preference. When a bus comes, you have to decide if you're gonna get on the bus or stay where you are. This is a very rough outline of how the Metropolis-Hastings algorithm works.

Definition 1 (Metropolis-Hastings Algorithm) Suppose we have some target stationary distribution $s = (s_1, \dots, s_M)$ on states $\{1,\ 2, \dots,\ M\}$, and $s_i > 0$ for all i (otherwise, delete that state.) Let P be a transition matrix on the same states (it may have a different stationary distribution.) Starting in some state X_0, at each stage do:

Suppose that $X_n = i$.

- Draw a proposed next state j according to the probability distribution in ith row of P.
- Accept this as X_{n+1} with probability:

$$a_{ij} = \min(1, \frac{s_j P_{ji}}{s_i P_{ij}}),$$

otherwise $X_{n+1} = i$.

The probability distribution of this MC will approach s. However, how fast it approaches s if we run the chain for long enough will depend on the choice of initial Markov chain, i.e. on P.

Example 1 (Smaug's Lair) Bilbo Baggins managed to get into Smaug's lair and found m treasures. Each treasure is worth some amount g_i of gold and weights w_i pounds. Bilbo can carry at most W pounds. How should he pick which treasures to steal?

This is called the *knapsack problem*, and it's really hard to optimize. Even powerful computers would time out if you asked for calculation on many items, and Smaug's lair is sure to have many, many treasures. So instead of finding the best possible combination, let's settle for some solution that gives us high value. Consider the following model. Each subset of the treasures is a binary vector $x = (x_1, \dots, x_M)$, with 1 if the treasure is included and 0 if it isn't. Let X be the set of vectors such that the total wights is at most W. Suppose we start from some $x \in X$ and run the following MC: at each stage pick a coordinate of the vector uniformly at random, and flip it as long as the result of the flipping is in X (i.e. Bilbo can carry it), stick to x otherwise.

What are the transition probabilities between $x, y \in X$? Well, if they differ by more than one coordinate, the transition probabilities are 0. If they only differ by one coordinate, we go from one to the other if that coordinate is picked, so transition probabilities from x to y and from y to x are both $1/M$. So this transition matrix is symmetric! And is the transition matrix is symmetric, the stationary distribution is uniform over X.

We just found a perfectly functional MC, and we know its transition probabilities and stationary distribution. How do we modify it to help us get items with high value? We use the Metropolis-Hastings algorithm.

Suppose that the value of collection x is $V(x)$ and we are shooting at a probability distribution that is proportional to $V(x)$, i.e:

$$s_x = \alpha V(x) \text{ for some normalizing factor } \alpha.$$

Then the Metropolis-Hastings algorithm can be applied by, at any state x, drawing y according to the previous method and setting the next step to y with probability:

$$a_{xy} = \min(1, \frac{\alpha V(y) P_{yx}}{\alpha V(x) P_{xy}}) = \min(1, \frac{V(y)}{V(x)}),$$

and stay at x otherwise. Notice that we used the fact that $P_{xy} = P_{yx} = 1/M$. In other words, if $V(y) \geq V(x)$ we always go to y, but there is also some probability we might go to y even if the value is smaller. This way the sampling favors collections with high value. We can make it even more skewed towards high value, by aiming at probability $e^{\beta V(x)}$ for some normalizing factor β, then we'd have:

$$a_{xy} = \min(1, \frac{e^{\beta V(y)}}{e^{\beta V(x)}}) = \min(1, e^{\beta(V(y)-V(x))}).$$

2 Solutions to practice problems

I. Which of the following matrices are transition matrices for regular Markov chains?

Recall that a MC is *regular* if there is a natural number n such that one can get from any state to any other state in exactly n steps. Equivalently, if some power of the transition matrix has all entries positive.

a. This matrix has all entries positive, so the chain is clearly regular.
b. P^2 has all entries positive. Alternatively, it's easy to check that the following routes are all possible in two steps:
   A to A: stay at A twice
   A to B: stay at A, go to B
   B to A: go to A, stay at A
   B to B: go to A, go to B
c. No, B is an absorbing state so there is no way to get anywhere else from B.
d. No, you will alwasy be at one state on even steps and the other at odd dteps.
e. Yes, you can clearly get anywhere from anywhere in at most 2 steps, and if you have steps to spare just wait at that particular state.

II. A cat and a mouse are in a two-room apartment. At each time step, the cat will stay in the same room with probability 0.2 and go to the other room with probability 0.8. If the mouse is in room 1, it will stay with probability 0.7 and go to room 2 with probability 0.3. If the mouse is in room 2, it will stay with probability 0.4 and go to room 1 otherwise.

– Find stationary distributions of the cat and mouse Markov chains.

  The two Markov chains are:

$$\text{cat } = \begin{bmatrix} 0.2 & 0.8 \\ 0.8 & 0.2 \end{bmatrix} \quad \text{mouse} = \begin{bmatrix} 0.7 & 0.3 \\ 0.6 & 0.4 \end{bmatrix}$$

  And the stationary distributions can be derived from a system of linear equations. For the cat:
  $a + b = 1$
  $0.2a + 0.8b = a$
  $0.8b = 0.8a$
  $a = b = 1/2$
  And for the mouse:
  $a + b = 1$
  $0.7a + 0.6b = a$
  $0.6b = 0.3a$
  $2b = a = 2/3$
  And so the stationary distributions are:

$$s_{\text{cat}} = (1/2, 1/2), \ s_{\text{mouse}} = (2/3, 1/3).$$

– Set up a joint Markov chain with four different states (how?). What is the expected time until they are in the same room?

  Let (x, y) be the state where x is the room the cat is in and y is the room the mouse is in. So the possible states are $(1,1), (1,2), (2,1), (2,2)$. Set up $(1,1)$ and $(2,2)$ as absorbing states. The transition matrix becomes:

$$P = \begin{array}{c} (1,2) \\ (2,1) \\ (1,1) \\ (2,2) \end{array} \begin{bmatrix} 2/25 & 12/25 & 3/25 & 8/25 \\ 12/50 & 7/50 & 28/50 & 3/50 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix}$$

And we're only interested in the upper left part of this matrix:

$$Q = \begin{bmatrix} 2/25 & 12/25 \\ 12/50 & 7/50 \end{bmatrix}$$

The fundamental matrix is:

$$N = (I - Q)^{-1} = \begin{bmatrix} 23/25 & -12/25 \\ -12/50 & 43/50 \end{bmatrix}^{-1} = \frac{250}{169} \begin{bmatrix} 43/50 & 12/25 \\ 12/50 & 23/25 \end{bmatrix} = \frac{5}{169} \begin{bmatrix} 43 & 24 \\ 12 & 46 \end{bmatrix}$$

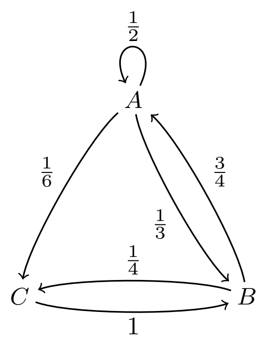
And the times to absorption:

$$t = \begin{pmatrix} 5 \times 67/169 \\ 5 \times 58/169 \end{pmatrix}$$

III. Consider the Markov chain with transition Matrix:

$$P = \begin{bmatrix} 1/2 & 1/3 & 1/6 \\ 3/4 & 0 & 1/4 \\ 0 & 1 & 0 \end{bmatrix}$$

a. Show that this is a regular Markov chain.



It's possible to get from any state to any other state in exactly three steps:

$A \to A \to A \to A$
$A \to A \to A \to B$
$A \to A \to A \to C$
$B \to A \to A \to A$
$B \to A \to A \to B$
$B \to A \to A \to C$
$C \to B \to A \to A$
$C \to B \to A \to B$
$C \to B \to A \to C$

b. The process is started in state A, find the probability that it is in state C after two steps.

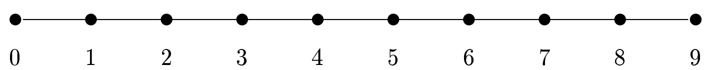
$$P_{11}P_{13} + P_{12}P_{23} + P_{13}P_{33} = \frac{1}{2}\frac{1}{6} + \frac{1}{3}\frac{1}{4} + \frac{1}{6}0 = \frac{2}{12} = \frac{1}{6}$$

c. Find the stationary distribution.

$a + b + c = 1$
$2a + 3b = 4a$
$3b = 2a$
$a + 3c = 3b$
$a + 3c = 2a$
$3c = a$
$c + 2c + 3c = 1$
$c = 1/6$

$$s = (1/2, 1/3, 1/6)$$

IV. Is a simple random walk on the path pictured below an ergodic Markov chain? Is it regular?



It's ergodic but not regular, this is another bipartite graph (same reason as 1d.)

V. Toss a fair die repeatedly. Let S_n denote the sum of the outcomes after n tosses. Let P_n be the proportion of the first n values S_n that are divisible by 7. It converges to a limit. Find this limit, by setting this process up as a 7-state Markov chain.

The states in the chain correspond to possible remainders of the sum. If S_n is divisible by 7, then S_{n+1} can be anything but divisible by 7, and so on:

$$P = \begin{bmatrix} 0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 0 & 1/6 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 0 & 1/6 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 & 1/6 \\ 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 1/6 & 0 \end{bmatrix}$$

Since this matrix is symmetric (even more so, every state is equivalent... but we don't need a condition as strong as that.) the stationary distribution is $(1/7,\ 1/7,\ 1/7,\ 1/7,\ 1/7,\ 1/7,\ 1/7)$. So the sum is divisible by 7 about 1/7th of the time.

VI. Prove that in an r-state ergodic chain it's possible to go from any state to any other state in at most $r-1$ steps.

By pidgeonhole principle, if the sequence $X_0, X_1, \ldots$ has at least r states, and it doesn't include the destination state, some state will appear at least twice. Then you can reduce the sequence by deleting what came before the last appearance of that state.

VII. Consider a Markov chain with the following transition matrix, for some $a,\ b$:

$$P = \begin{bmatrix} 1-a & a \\ b & 1-b \end{bmatrix}$$

a. Under what conditions is P absorbing?

   If either a or b is 0, or both.

b. Under what conditions is P ergodic but not regular?

   If $a = b = 1$. Otherwise, suppose that $a \neq 1$, and $a, b \neq 0$ Then can go $A \to A \to A$, $A \to A \to B$, $B \to A \to A$, $B \to A \to B$, and the chain is regular. The same argument works for $b \neq 1$.

c. Under what conditions is P regular?

   $a, b \neq 0$, and either $a \neq 1$ or $b \neq 1$.