CS265/CME309: Randomized Algorithms and Probabilistic Analysis

Lecture #14: The Fundamental Theorem of Markov Chains and Stationary Distributions

Gregory Valiant[*], updated by Mary Wootters

November 9, 2020

1 The Fundamental Theorem of Markov Chains

We begin with several definitions that will provide some terminology for discussing the long-term behavior of a time-homogeneous Markov chain.

Definition 1. *A (time homogenous) Markov chain, $X_0, X_1, \ldots,$ is* irreducible *if, for all pairs of states, i, j there is a positive probability of eventually getting to state j when starting at state i:*

$$\sum_{t \geq 0} \Pr[X_t = j | X_0 = i] > 0.$$

Given the representation of a Markov chain as a graph, with nodes being the states, and nonzero probability transitions represented as directed edges, the above definition is equivalent to requiring that the graph is *strongly connected.*

For both Markov chains with finite, and infinite state spaces, it is useful to classify states into the set of states that are *transient*—meaning the chain will visit them only a finite number of times—and those that are *recurrent*—meaning that the chain will visit them an unbounded number of times, if the chain is run for sufficiently long:

Definition 2. *Letting*

$$r_i = \sum_{t \geq 1} \Pr[X_t = i \textit{ and for } t' \in \{1, \ldots, t-1\}, X_{t'} \neq i | X_0 = i]$$

denote the probability that a Markov chain returns to state i, given that it starts there, we say that state i is transient *if $r_i < 1$, and* recurrent *if $r_i = 1$.*

In the case that a state i is transient, then the number of times the chain will be in state i, given that it starts in state i, is given by the geometric random variable with parameter $1-r_i$, and hence has expected value $1/(1-r_i)$. For recurrent states, one can also ask what the expected time is between visits to a given state, namely $\mathbf{E}[\min\{t : X_t = i\}|X_0 = i]$. For Markov chains with a finite number of states, a state is recurrent if and only if this expected time is finite. As the following example illustrates, for infinite chains, it is possible that a state is recurrent, but the expected time to return is infinite.

Example 3. *(Gambler's Ruin) Consider the Markov chain $X_0, \ldots$ where $X_0 = 0$, and $X_t = X_{t-1} \pm 1$, with probability $1/2$ of each outcome. First, we claim that state 0 is recurrent. To see this, recall our analysis of the randomized 2-SAT algorithm from last class. An immediate consequence of that analysis showed that, given that $X_t = i$, with probability at least $1/2$ there is some $t' \in \{t, \ldots, t+2n^2\}$ such that $|X_{t'} - i| \geq n/2$. That is, the walk will have wandered at least distance $n/2$ from where we started within those $2n^2$ steps. Hence, whatever the value of X_t, with probability at least $1/4$, we have hit 0 within the next $2X_t^2$ steps. (Here, we have $1/4$ instead of $1/2$ because of symmetry: we are equally as likely to wander to zero from X_t as we are to $2X_t$, so there's at least a $1/4$'th chance of each happening. Hence for every value of X_t, we can name a time $t' = f(t)$ such that with constant probability we will have returned to 0 within the next t' steps, and hence state 0 must be recurrent. (Indeed, the probability that we don't return to 0* ever *is at most $(3/4) \cdot (3/4) \cdot (3/4) \cdot (3/4) \cdot \cdots$, which is equal to zero in the limit.)*

The expected time until we return to 0, however, turns out to be infinite*! We will see a proof of this next week, in the context of analyzing hitting times.*[1]

Intuitively, this means that if we play a fair betting game (we toss a coin and if its heads, you win a dollar, otherwise, you lose a dollar), then if you were to play for infinitely long and the person you were playing with had an infinite amount of money, you are guaranteed to go broke at some point. However, the expected time you will need to play until this happens is infinite! This example is often referred[2] *to as the "Gambler's Ruin," since eventually the gambler will go broke. Another perspective on this is that the gambler is "ruined", not because they expect to go broke eventually, but because they expect to spend so much time gambling.*

The next definition characterizes when a Markov chain has some periodic behavior:

Definition 4. *A Markov chain $X_0, X_1, \ldots$ is* periodic *if there exists a state, i, such that*

$$gcd\left(\{t | \Pr[X_t = i|X_0 = i] > 0\}\right) \neq 1,$$

[1] Intuitively, this is because we expect it to take about X_t^2 steps before we either reach 0 or reach $2X_t$, and we are equally likely to hit 0 first as we are to hit $2X_t$. So, the expected time to hit 0 from, say, 2^i is $\frac{1}{2}2^{2i} + \frac{1}{2} \cdot (\mathbb{E}[\text{time to hit 0 from } 2^{i+1}])$. Repeating this logic, we get

$$\frac{1}{2}2^{2i} + \frac{1}{4}2^{2(i+1)} + \frac{1}{4} \cdot \mathbb{E}[\text{time to hit 0 from } 2^{i+2}\ ]$$

and then

$$\frac{1}{2}2^{2i} + \frac{1}{4}2^{2(i+1)} + \frac{1}{8}2^{2(i+2)} + \cdots$$

and so on. The sum that we end up with (assuming we started with $i = 1$, say) is $\sum_j \frac{2^{2j}}{2^j} = \sum_j 2^j$, which diverges.

[2] There are also other, different, examples that are often referred to as "Gambler's Ruin." Turns out that there are lots of ways that gambling can lead to ruin...

and is aperiodic *if no such state exists. (Here "gcd" is short-hand for the "greatest common divisor".)*

Example 5. *The "Gambler's Ruin" Markov chain is periodic, because, for example, you can only ever return to state 0 at even time-steps:* $gcd\{t| \Pr[X_t = 0|X_0 = 0] > 0\} = 2.$

Fact 6. *Any irreducible Markov chain that has at least one "self-loop" (ie one state i for which* $\Pr[X_t = i|X_{t-1} = i] > 0$*, is aperiodic.*

Proof. Suppose state i has a self-loop. From any state j, the chain can eventually get to i (by irreducibility), and use the self-loop any number of times, and then return to j (by irreducibility), rendering the greatest common divisor of timesteps at which we could have returned to state j to be 1. □

Given all these definitions, we can now state the Fundamental Theorem of Markov chains, which states that finite, irreducible, aperiodic Markov chains have a unique *stationary distribution.* That is, there is some distribution π so that if X_t is distributed according to π, then so is X_{t+1}. We omit the proof of this theorem (if you are curious, there is a nice overview in Mitzenmacher and Upfal).

Theorem 1 (The Fundamental Theorem of Markov Chains)**.** *Let $X_0, X_1, \ldots$ be a Markov chain over a finite state space, with transition matrix P. Suppose that the chain is irreducible and aperiodic. Then the following hold:*

1. *There exists a unique* ***stationary distribution****, $\pi = (\pi_1, \pi_2, \ldots)$ over the states such that: for any states i and j,*
$$\lim_{t\to\infty} \Pr[X_t = i|X_0 = j] = \pi_i.$$

2. *For each state i, $\pi_i = \frac{1}{\mathbf{E}[\min(t: X_t = i)|X_0 = i]}$, namely π_i is the inverse of the expected return time of state i.*

3. *π is a left eigenvector of matrix P, with eigenvalue 1, namely the vector-matrix product*
$$\pi P = \pi.$$

The crucial aspect of the above theorem is that the probability of being in a given state, i, eventually becomes *independent* of the state we started in, j. This is the crucial property of irreducible and aperiodic Markov chains! It is natural to ask whether analogs of the above theorem hold when the Markov chain in question is either 1) not finite, 2) not irreducible, and 3) not aperiodic.

- (Infinite State Spaces) There is an analog of the theorem that applies to Markov chains with an infinite state space, which you will see on Problem Set 7. This theorem states that, provided the chain is aperiodic and irreducible, either there is a unique stationary distribution such that for all states i, j $\lim_{t\to\infty} \Pr[X_t = i|X_0 = j] = \pi_i > 0$, or for all states $\lim_{t\to\infty} \Pr[X_t = i|X_0 = j] = 0$.

- (Not Irreducible) Given any non-irreducible Markov chain, one can decompose it into (possibly more than one) irreducible components, corresponding to the strongly connected components of the graph. Each of these components will have a stationary distribution (provided the chain is aperiodic), though which of these stationary distributions the chain eventually ends up at, depends on the randomness of the early part of the chain.

- (Not Aperiodic) If a chain is periodic, but is finite and irreducible, there will still be a unique stationary distribution, π, satisfying $\pi P = \pi$, though it is *not* the case that $\lim_{t\to\infty} \Pr[X_t = i|X_0 = j] = \pi_i$, since this probability will depend on whether or not t divides the period, and hence this limit does not exist. This is the only issue: for example, it is still the case that, in the limit, the chain will have been in state i exactly π_i fraction of the time.

The uniqueness of the stationary distribution for irreducible, aperiodic (finite) chains is extremely powerful. It means that, if we are able to guess a distribution π and check that $\pi P = \pi$, then we have proved that π is *the* stationary distribution. One way of thinking about the condition that $\pi P = \pi$, is that this condition simply means that, if we start with π_i probability mass at state i, and evolve the chain by one step, then the amount of probability mass leaving state i (along the outgoing edges in the graph representation of the Markov chain) is *exactly* equal to the probability arriving at state i from its neighbors. The following two examples/propositions illustrate this approach to describing stationary distributions.

Proposition 7. *For any Markov chain where transitions are symmetric (e.g.* $\Pr[X_t = i|X_{t-1} = j] = \Pr[X_t = j|X_{t-1} = i]$*), if the chain is aperiodic and irreducible, then the stationary distribution is the uniform distribution over states.*

Proof. Letting π denote the uniform distribution over states, which assigns probability $1/|S|$ to each of the $|S|$ states, consider the ith entry of πP, which we'll denote by $\pi P(i)$:

$$\pi P(i) = \sum_j \pi_j P_{j,i} = \sum_j \pi_j P_{i,j} = \frac{1}{|S|}\sum_j P_{i,j} = \frac{1}{|S|} = \pi_i,$$

where the second equality used the assumption of symmetry, that $P_{i,j} = P_{j,i}$, the third equality used the fact that $\pi_j = 1/|S|$, and the final equality used the fact that the sum of entries in any row of P is 1, as these entries correspond to the distribution of transitions out of a fixed state. □

Example 8. *Consider the following protocol for shuffling a deck of n cards: choose 2 cards at random, and swap their positions. (If we pick the same card twice, then assume we don't do anything in that step.) Since the transitions are symmetric, (and the chain is aperiodic and irreducible...) by the above proposition, the stationary distribution is the uniform distribution over the $n!$ orderings of the cards, and this is a valid shuffling procedure.*

Proposition 9. *Let $X_0, X_1, \ldots$ represent a random walk on a connected undirected graph, defined by letting X_t be a uniformly random neighboring node of the node corresponding to X_{t-1}. Provided the graph is not bi-partite, then there is a unique stationary distribution that puts probability $\pi(v) = \frac{degree(v)}{2|E|}$ on node v, where $|E|$ denotes the number of edges in the graph.*

The magic of the above proposition is that this stationary distribution depends only on the degrees of the nodes, and not on the structure of the graph! [E.g. suppose we have a social network, and there is a magical stone being passed from friend to friend. The probability you have the stone at some fixed time, t, long in the future, is a function of **only** the number of friends you have, and doesn't depend on who those friends are connected to.

Proof. First, note that π as defined in the problem is actually a distribution, because the $\sum_v degree(v) = 2|E|$. Since the graph is connected, the walk is irreducible. If the graph is not bipartite, there is a cycle of odd length (by definition). There is also a cycle of even length, e.g. take any edge, and cross it then cross back. Hence the greatest-common-divisor of return times can be any even number, and also can include any multiple of this odd length cycle, and hence the gcd is 1. So there is a unique stationary distribution. To see that the claimed distribution is the stationary distribution, consider the amount of probability mass leaving node v at one step of the walk: this is simply $\pi(v)$, since there are no self-loops. The probability entering state v is

$$\sum_{u=Neighbor(v)} \frac{deg(u)}{2|E|}\frac{1}{deg(u)} = \sum_{u=Neighbor(v)} \frac{1}{2|E|} = \frac{deg(v)}{2|E|} = \pi(v),$$

where the second term in the first expression corresponds to the fact that each neighbor, u, will send $1/deg(u)$ of its probability mass to each of its neighbors, including v. Hence $\pi P = \pi$, and so π is the stationary distribution. □

2 Markov Chain Monte Carlo: The Metropolis Algorithm

Given the fundamental theorem of Markov chains, one way of drawing samples from some distribution of interest, would be to 1) construct an irreducible aperiodic Markov chain whose stationary distribution, π is the distribution of interest, and then 2) run the chain for a long time, and return X_t as a samples from something close to the stationary distribution. (Ideally, we would have $t \to \infty$, though then we would never get our sample....)

For many distributions of interest, it is much easier to construct such a Markov chain versus trying to describe the distribution explicitly. (This is especially true when the distribution corresponds to some process whose evolution is easy to model via a Markov chain.)

Example 10. *Suppose that we want to sample a uniformly random proper k-coloring of a graph G. (Here, a "proper k-coloring" is a way to color the vertices so that no two adjacent vertices have the same color). We may want to do this, for example, if we want to estimate the fraction of colorings in which vertex* 1 *and vertex* 10 *have the same color: just draw a bunch of random colorings and ask how many of them color those two vertices the same color.*

It seems a bit tricky to sample a random proper coloring—in fact, it seems tricky even to count the number of proper colorings of an arbitrary graph, and hence tricky even to determine the probability distribution we want to sample from! (This latter problem is #P complete, meaning that it's unlikely that there's an efficient algorithm to do it exactly in general). But we can approximately *sample a proper coloring using a Markov chain.*

Consider the following Markov chain on proper colorings. Suppose that X_t is any proper coloring. We form X_{t+1} as follows:

- *Choose a random vertex and a random color.*
- *If coloring that vertex that color in X_t results in a proper coloring, let X_{t+1} be the new coloring.*
- *Otherwise, $X_{t+1} = X_t$.*

It is not hard to check that this is aperiodic. It turns out that it's irreducible provided that k is big enough (at least the maximum degree plus 2), so by Proposition 7, the stationary distribution is uniform. Hence as $t \to \infty$, the distribution of X_t approaches the uniform distribution.

2.1 The Metropolis Algorithm

We got lucky in the example above: we happened to guess a Markov chain that just so happened to have the stationary distribution we were after. But what if we don't happen to get so lucky?

The Metropolis Algorithm is one generic way of constructing a chain to have a desired stationary distribution, π. To apply this approach, we need the following ingredients:

- A connected graph G whose nodes are the set S corresponding to the support of the distribution π. Ideally, the degree of G should be pretty small, and given $i \in S$, it should be efficient to compute the neighbors of i in G.

  [In Example 10, S would be the set of all proper colorings, and one proper coloring would be connected to another if it only differs on one vertex.]

- For two states, i, j, in the above graph, we need to be able to calculate the ratio $\pi(i)/\pi(j)$, where π is the distribution we care about.

  [In Example 10, we have $\pi(i)/\pi(j) = 1$ for all i, j, since we are looking for the uniform distribution]

  Notice that if we only know π up to some scaling factor—e.g., if π is uniform over some unknown number of things—then we can compute $\pi(i)/\pi(j)$.]

Given the above setup, consider the following Markov chain over states S: Let d be any constant larger than the maximum degree of the graph we are given. Then define the transition matrix $P_{i,j}$ as follows:

$$P_{i,j} = \begin{cases} 0 & \text{if } i, j \text{ not neighbors} \\ \frac{1}{d}\min(1, \frac{\pi(j)}{\pi(i)}) & \text{if } i \neq j \text{ and they are neighbors} \\ 1 - \sum_{\ell \neq i} P_{i,\ell} & \text{if } i = j. \end{cases} \tag{1}$$

Theorem 2. *The Markov chain constructed above is irreducible, aperiodic, and has stationary distribution π.*

Proof. The irreducibility is from the connectedness of the graph, and the aperiodicity is because d is larger than the maximum degree, and hence there are self-loops that have positive probability. We now analyze the probability leaving, and entering at each node, given the distribution π, to prove that $\pi P = \pi$. In the following expression, $N(i)$ denote the set of neighboring states in the graph of state i, and $i \notin N(i)$. The probability leaving node i is

$$\pi(i) \sum_{j \in N(i)} \frac{1}{d} \min(1, \frac{\pi(j)}{\pi(i)}) = \left(\sum_{j \in N(i): \pi(i) \geq \pi(j)} \frac{\pi(j)}{d} \right) + \left(\pi(i) \sum_{j \in N(i): \pi(i) < \pi(j)} \frac{1}{d} \right).$$

The total probability mass arriving at state i is

$$\sum_{j \in N(i)} \frac{\pi(j)}{d} \min(1, \frac{\pi(i)}{\pi(j)}) = \left(\sum_{j \in N(i): \pi(i) \geq \pi(j)} \frac{\pi(j)}{d} \right) + \left(\sum_{j \in N(i): \pi(i) < \pi(j)} \frac{\pi(i)}{d} \right).$$

Hence the probability entering and leaving each node is equal, so $\pi P = \pi$, and hence π is the (unique) stationary distribution, by the fundamental theorem of Markov chains. □

If we this theorem to the set-up in Example 10, wth d equal to $|V| \cdot k$ (the number of vertices in G times the number of colors), we will recover the same Markov chain from the example.[3]

To see the power of the Metropolis algorithm, consider trying to adapt Example 10 to the distribution π where a coloring with $k - j$ colors is j times more likely than a coloring with exactly k colors. That is, for a proper coloring C with exactly $k - j$ colors, $\pi(C) = j/Z$, where

$$Z = \sum_{0 \leq j \leq k} j \cdot [\text{number of proper colorings with exactly } k - j \text{ colors }]$$

is a normalizing constant.

In this case, it seems quite hard to compute $\pi(C)$ for any given coloring C, and it's not a priori obvious [to me...] how to set up a Markov chain with π as a stationary distribution. But $\pi(C)/\pi(C')$ is pretty easy to compute, and so we can use the Metropolis algorithm.

Of course, in order for the Metropolis approach to be meaningful in any concrete sense, we not only need that, in the limit as $t \to \infty$, X_t is drawn from π, but we need that for some (ideally small) finite value of t, the distribution of X_t is close to π. In the next class, we will define the notion of *mixing time*, which is the amount of time it takes before X_t becomes close to π, for an appropriate definition of "close", and will discuss several techniques for bounding the mixing time. One of these techniques is "coupling", which is an incredibly elegant approach that often requires a bit of creativity to apply.

[3] We can choose $d = |V| \cdot k$ since the degree of this graph is strictly less than $|V| \cdot k$.