CS265, Fall 2022

Class 14 Agenda: Markov Chains II

1 Announcements

- No HW for a while! But HW7 should be out soon.

2 Questions/Lecture Recap

Any questions or reflections from the quiz or minilectures? (Definitions about Markov chains; stationary distributions; fundamental theorem of Markov chains; Markov Chain Monte Carlo)

3 Gibbs Sampling

In this group work, we'll explore a special case of MCMC, called "Gibbs Sampling" which arises in many settings in machine learning and language modeling.

Suppose that π is a joint distribution on X and Y. Suppose that it is hard to sample from π, but relatively easy to sample from $\pi(X|Y=y)$ or $\pi(Y|X=x)$ for any x, y in the support of X and Y respectively.

Consider the following way to set up a Markov chain $(X_0, Y_0), (X_1, Y_1), \ldots$:

- Suppose $(X_t, Y_t) = (x, y)$.
- Draw $x' \sim \pi(X|Y=y)$.
- Draw $y' \sim \pi(Y|X=x')$.
- Set $(X_{t+1}, Y_{t+1}) = (x', y')$.

That is, we first condition on $Y = y$ and draw a new value x' for X, and then we condition on that value x' for X and draw a new value y' for Y.

Group Work

1. With the setup above, show that π is a stationary distribution for this Markov chain.

   Hint: Recall that you want to show that for all x, y,

$$\pi(x,y) = \sum_{x',y'} \pi(x',y') \Pr[(x',y') \to (x,y)]$$

(Why?)

2. Does the Fundamental Theorem of Markov Chains automatically apply in this setting? If not, what additional assumptions do you need to make?
3. This procedure is called "Gibbs Sampling." If it's easy to sample from the marginal distributions, but difficult to sample from π itself, explain why the previous two parts (assuming your assumptions in the previous part are met) give us an algorithm to approximately sample from π. (Don't worry about how efficient the algorithm is for now...)
4. What happens if you apply (an appropriately multivariate) form of Gibbs sampling to the problem of sampling a random proper coloring of a graph? Do you get the same algorithm as we saw in the minilecture, or a different algorithm?

   Note: To do this question you'll have to think about how to extend what we did above to more than two variables.
5. (This one is a bit more open-ended...) Suppose your goal is to create a language model that allows you to sample a uniformly random 7-word sentence from the distribution of naturally occurring 7-word sentences. How could you use Gibbs sampling to do this, and what would the challenges be?

   Note: Same note as the previous question.
6. Has anyone in your group encountered MCMC before? In what context? If not, what else can you think of that Gibbs sampling or MCMC more generally might be useful for?

Group Work: Solutions

1. We want to show that $\pi = \pi \cdot P$, which is the same as showing

$$\pi(x, y) = \sum_{x', y'} \pi(x', y') \Pr[(x', y') \to (x, y)]$$

   for all x, y. (This just follows from the definition of matrix multiplication). To see

this, note that

$$\begin{aligned}\sum_{x',y'} \pi(x',y') \Pr[(x',y') \to (x,y)] &= \sum_{x',y'} \pi(x',y')\pi(x|y')\pi(y|x) \\ &= \pi(y|x) \sum_{y'} \pi(x|y') \sum_{x'} \pi(x',y') \\ &= \pi(y|x) \sum_{y'} \pi(x|y')\pi(y') \\ &= \pi(y|x) \sum_{y'} \pi(x,y') \\ &= \pi(y|x)\pi(x) \\ &= \pi(x,y),\end{aligned}$$

as desired.

2. We need to check the following things:
   - Aperiodic: check! There's a self-loop for any x, y with $\pi(x,y) > 0$. Indeed, the probability that we stay at x, y is $\pi(x|y) \cdot \pi(y|x) > 0$.
   - Irreducible? Not necessarily! One way to envision what the requirement is is the following. Consider a bipartite graph, with x's on the left and y's on the right. There is an edge between x and y if $\pi(x,y) > 0$. Then we can view our states as edges of this graph. Two states are connected to each other (that is, we can get from one to the other in one step of our markov chain) if the corresponding edges are incident. We need our underlying state graph to be connected, which is the same as saying that we need this bipartite graph to be connected.
     Another way to say this is that if Π is a matrix with rows indexed by x's and columns indexed by y's, and $\Pi_{x,y} = \pi(x,y)$, then if we look at the non-zero patter of Π, it should correspond to the adjacency matrix of a connected bipartite graph.
   - Finite? Sure, as long as π has finite support.
3. The fundamental theorem of markov chains says that, eventually, $(X_t, Y_t) \to \pi$. Since we can efficiently sample from the marginals $\pi(X|Y)$ and $\pi(Y|X)$, we can efficiently step this Markov chain along. Of course, we don't yet know how large we need to take t to get close to π.... (that's next time!!)
4. It's not quite the same. The algorithm looks like:
   - Start with an arbitrary coloring.
   - While True:
     - For each vertex v:

- Uncolor v.
- Choose a uniformly random color for v among all of the colors that are legitimate, and color v that color.

5. (Some discussion in class. Note that this question gets a bit away from the theory, and in particular you aren't responsible for this sort of thing for HW or the exam).