COMP 480/580 — Probabilistic Algorithms and Data Structure Mar 28, 2019

Lecture 19

Lecturer: Anshumali Shrivastava *Scribe By: Hamza Nauman*

1 Motivation: Sampling from State Space

In this lecture, we continue our discussion of Monte Carlo Markov Chain (MCMC) methods. To motivate this lecture, we introduce the problem of sampling from a very large state space. Given a state space, Ω, we wish to sample x from Ω such that the probability of getting any $x \in \Omega$ is proportional to the weight associated with x, $w(x)$. In other words,

$$P[\text{sampling } x \in \Omega] = \frac{w(x)}{\sum_{y \in \Omega} w(y)}$$

We will suppose that Ω is so large that $\sum_{y \in \Omega} w(y)$ is not easy to compute.

1.1 Example: Random Permutation Problem

An example illustrating this concept is the random permutations problem. Essentially, given N numbers, we wish to generate a random permutation of these numbers. Here, Ω is the set of all permutations. Initially, we will suppose that $w(x) = 1$ for each permutation, x (i.e. each permutation is equally likely). A concrete example following this model is the act of shuffling a deck of N cards. We represent each permutation of cards as a number and we want to randomly sample from the numbers $[1, N!]$. We can solve the random permutation problem using the algorithm below.

```
Algorithm 1: Random Permutation Problem
  Input: N numbers
  Result: A random permutation of N numbers
  Start with the order initialized to 1, 2, ... N;
  for repeat for 1...k do
      select i, j ∈ [1...$N$] randomly;
      swap the ith element and jth element;
  end
  return The current order of elements;
```

This algorithm swaps different integers within the list k times. As we increase the number k - in particular, as $k \to \infty$ - the returned permutation becomes truly random. This algorithm works for this specific problem, but only because all of our weights $w(x)$ were equal to 1. In the general case when $w(x) \neq 1$, we require Markov chains.

2 Markov Chains

A Markov chain is a stochastic process that consists of a sequence of states, $\{X_0, X_1, ..., X_t, ...\}$. Each state, X_i is from the set of all states, Ω. The interesting thing about Markov chains is that they are memory-less. This means that the next state is determined only by the current state, and no other state in the past matters. Thus, the probability of reaching state y at time $t+1$, when we were at state x at time t can be written as:

$$Pr[X_{t+1} = y | X_t = x, X_{t-1} = x_{t-1}, ..., X_0 = x_0] = Pr[X_{t+1} = y | X_t = x] = P(y|x)$$

For the rest of our discussion of Markov chains, we will use $P(x, y)$ to denote the conditional probability $P(y|x)$. The following facts, which follow direction from the basic rules of probability, will be important in our analysis: $P(x, y) \geq 0$ for all $x, y \in \Omega$, and $\sum_{y \in \Omega} p(x, y) = 1$, for any $x \in \Omega$.

2.1 Markov Chains as Graphs

Based on the definition of a Markov chain, we can visualize the chain as a directed graph. Recall from graph theory that we can represent a graph g as a set of nodes V and edges E: $g = (V, E)$. The set of nodes is the set of states, so $V = \Omega$. The set of edges (x, y) is determined by the state transition probabilities of the Markov chain. A directed edge between two nodes represents a non-zero probability of transitioning from state x to state y. This edge has a weight of $P(x, y)$.

Due the laws of probability, our graph has some structural properties. For instance, the weights of all outgoing edges from any node must add up to 1. It should be noted that a node can also have an edge to itself. We can also see that representing the Markov chain as a graph captures the idea that Markov chains are memoryless, since it does not matter what path we took to arrive at a particular node when traversing the graph. To traverse the graph, all that matters are the relative probabilities of all outgoing edges from that node.

2.2 Markov Chains as Matrices

Recall that all graphs can be represented as an adjacency matrix P, where $P_{i,j}$ is the weight of the directed edge between node i and node j. Therefore another way to represent a Markov chain is using a matrix. A Markov chain can be represented by an $N \times N$ matrix P, where $P_{x,y} = P(x, y)$ and $N = |\Omega|$. In other words, entry (i, j) in P gives us the probability of transitioning from state i to state j. This matrix is know as the stochastic matrix of a Markov chain. By representing Markov chains as matrices, we can use well-established linear algebra techniques in our analysis. In particular, we can easily compute the future distribution of states.

Given the distribution of states at time t, represented by a row vector π_t, where $\pi_t[i]$ gives us the probability of being in state i at time t, we can compute the distribution of states at time $t + k$ by:

$$\pi_{t+k} = \pi_t \times P^k$$

Here, P^k is the matrix product $P \times P \times ... \times P$, k times.

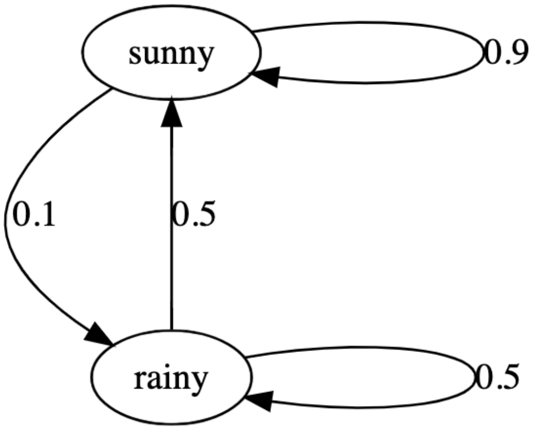

Figure 1: Graph representing Markov chain.

2.3 Example: Rainy vs. Sunny Weather

Let us consider a simple Markov chain which predicts if a day is sunny or rainy. Then, $\Omega = \{sunny, rainy\}$, and we can define $P(sunny, rainy) = 0.1$, $P(sunny, sunny) = 0.9$, $P(rainy, rainy) = 0.5$, $P(rainy, sunny) = 0.5$. This Markov chain can be represented by the graph shown in Figure 1. The matrix P for this Markov chain is:

$$P = \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}$$

Suppose we knew that on day t it was sunny, and we want to know the likelihood of it being sunny or rainy two days later. We can easily calculate this from the matrix representation of the Markov chain.

$$\pi_{t+2} = \pi_t * P^2$$

$$\pi_{t+2} = \begin{bmatrix} 1.0 & 0.0 \end{bmatrix} \begin{bmatrix} 0.9 & 0.1 \\ 0.5 & 0.5 \end{bmatrix}^2$$

$$\pi_{t+2} = \begin{bmatrix} 0.86 & 0.14 \end{bmatrix}$$

Therefore, we know that there is an 86% likelihood that it will be sunny in two days and a 14% likelihood that it will be rainy.

3 Stationary Distributions of Markov Chains

When we run a Markov chain, we can imagine that it might converge at some point to a stationary distribution after a long time. This is a useful property of Markov chains to have, and it is interesting for reasons we will see later. First, we will provide a more formal definition for a stationary distribution in the context of Markov chain analysis. A distribution π, is a stationary distribution of a Markov chain if:

$$\pi = \pi \times P$$

In other words, a distribution is a stationary distribution if propagating the state transition does not change the distribution. This is easy to see because if $\pi = \pi \times P$, then $\pi = \pi \times P^k$ for any positive integer k. We commonly want to know if a Markov chain has a stationary distribution. If it does have a stationary distribution, we may also want to know if it is unique regardless of which node we start on in the graph. The fundamental theorem of Markov chains helps us determine these properties. Before we introduce the theorem, we must define irreducible and aperiodic Markov Chains.

3.1 Irreducible Markov Chains

We say that a Markov Chain is irreducible if for any $x, y \in \Omega$,

$$\exists\, t, \quad s.t. \;\; P(x,y)^t > 0$$

Note: $P(x,y)^t$ here is an abuse of notation. It is the probability of reaching state y from state x in t steps (and not $P(x,y)$ raised to power t). In simple terms, irreduciblity basically means that you can reach any state no matter which state you start from.

3.2 Aperiodic Markov Chains

We say that a Markov Chain is aperiodic if for all $x, y \in \Omega$,

$$\text{GCD}\{t : p(x,y)^t > 0\} = 1$$

Here GCD stands for greatest common divisor. If a chain is aperiodic, then it is possible to return to state x for all $t = 1, 2, ...$ In particular, this means that aperiodic Markov chains have nonzeros along the diagonal of P.

3.3 The Fundamental Theorem of Markov Chains

The Fundamental Theorem of Markov Chains states that if a Markov Chain is both irreducible and aperiodic, then it is guaranteed to have a "unique" stationary distribution π. Moreover, the chain *converges*. That is

$$p(x,y)^t \to \pi[y] \;\; as \;\; t \to \infty$$

for any $x \in \Omega$.

4 Some Practical Considerations

In most practical scenarios we deal with in real life, the set of all states Ω is very big. For example, in the random permutations problem, the set consists of all possible permutations of N, which is exponential in N. Keeping that in mind, it is not feasible to "simulate" most Markov chains, since the matrix P would be $N! \times N!$ in size. This is far too large for most N from a practical standpoint. However, thinking about Markov Chains in terms of graphs and matrices is to understand the behavior of the system. In practice, we are usually primarily concerned with the convergence of a Markov Chain. We will see that there are other, more practical ways to get the stationary distribution even when the matrix is $N! \times N!$.