COMP 580: Probabalistic Algorithms and Data Structures **(04/30/2019)**

Markov Chains, Stationary distributions, MCMC 2

Instructor: Dr. Anshumali Srivastava *Name:* Advait Balaji, *Netid:* ab114

1. Markov Chain

As introduced in the previous class markov chains describe transformations of a set of random variables over a set of states Ω. A stochastic process is a collection of random variables X_i indexed over some set A, i.e. $\{X_i : i \in A\}$. Let $X_0, X_1, \cdots$ be a stochastic process. A Markov chain is a sequence of random variables such that the next state X_{i+1} depends only on the current state X_i exhibiting a memoryless property.

$$P(X_{i+1} = y | X_i = x_i, \cdots, X_0 = x_0)$$
$$= P(X_{i+1} = y | X_i = x_i).$$

The probability $P_{xy} = P(X_{i+1} = y | X_i = x)$ is called the *transition kernel*. A matrix consisting of transition kernels $\forall x, y$ is called a *transition matrix*. Note that $\sum_y P_{xy} = 1$. The initial distribution for X_0 determines the distribution for any n-th state.

2. Markov Chains as Graphs

Without loss of generality we can represent an arbitrary kernel at x, y as $P(x \to y)$ as $P(x, y)$. This allows us to represent Markov chains as directed graph $G(V, E)$. Here we can represent each $P(x, y)$ as an edge and each state in Ω as node in the graph. This edge can also be visualized as carrying a "weight" equal to $P(x, y)$ indicating the transition probabilities between the given nodes. Each outgoing edge from State X has effectively normalized weights as a result as $\sum_y P_{xy} = 1$. Figure 1 shows an example of a Markov process as a directed graph

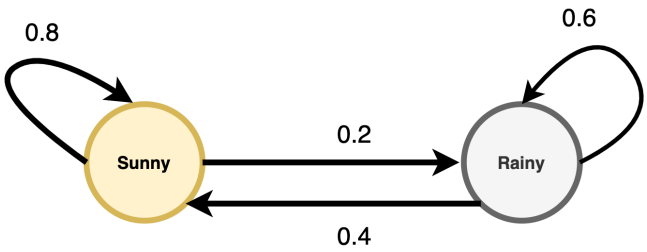

Figure 1: Markov Chain as a directed graph

3. Properties of the Transition Matrix

We denote the transition matrix as $\mathbb{P}$. Using the transition matrix and an initial distribution (probabilities) over the states we can then simulate running a markov chain in order to ascertain the value of the distributions after t iterations (i.e at time time t). In particular we can write;

$$\Pi_{t+1} = \Pi_t \times \mathbb{P}$$

Equivalently,

$$\Pi_t = \Pi_o \times \mathbb{P}^t$$

where Π_o is the stationary distribution at time t = 0 and $\mathbb{P}^t$ is just the t^{th} power of the transition matrix.

4. **Running a Markov Chain**

We can describe the process of simulating a Markov Chain briefly as:

- Start with an initial stationary distribution Π_o.
- Sample a state from the distribution, i.e choose $\mathbb{X}_i$
- We sample state $\mathbb{X}_j$ from row $\mathbb{P}_i$ of the transition matrix
- Repeat above step from $t = 0 \cdots t$

5. **Example of a Markov Chain**

An example of a Markov Chain simulation is the Random Shuffling Algorithm. The problem states that given an initial ordering of cards, at every iteration we randomly pick two cards and swap their positions. We then continue this for t iterations and argue that this indeed a Markov chain.

Algorithm 1: Random Shuffling Problem

Input: Array of N numbers/cards
Result: A random permutation of N numbers
Initial state 1..n cards;
for $t \leftarrow 0$ **to** t **do**
    select i, j[1...N];
    swap the i$^{\text{th}}$ element and j$^{\text{th}}$ element;

Starting from an initial ordering, there are at most ${}^{\text{n}}\text{C}_2$ states that can be reached each with a probability of $\frac{1}{{}^{\text{n}}C_2}$ and satisfies the Markov property.

6. **Chain Convergence**

Based on the above discussions it is possible to claim that the state $\mathbb{X}_i$ sampled at time t from the distribution Π_t is a unique stationary distribution. In other words, given sufficient t ($t \to \infty$), it is possible to reason, $\Pi_{t-1} \approx \Pi_t$ and is independent of Π_o.

Proof: We know that for a given eigenvector λ_i and eigenvalue e_i, we can decompose $\mathbb{P}$ as $\lambda_i e_i =_i \mathbb{P}$. We need to prove that $\Pi_o p^t \to e_i$. Let the leading eigenvector be 1 and corresponding eigenvalue be e_1. By definition we have $|\lambda_i| < 1, \forall i \in [2, 3..n]$. We have:

$$\Pi_o p^t = \sum_{i=1}^{n} \alpha_i e_i p^t$$

Normalizing,

$$= \frac{\sum_{i=1}^{n} \alpha_i \lambda_i^t e_i}{n_t}$$

$$= \alpha_1 e_1 + \sum_{i=2}^{n} \alpha_i \lambda_i^t e_i$$

We know that $|\lambda_i| < 0 \forall i > 1$, then the second term vanishes

$$= \alpha_1 e_1$$

Hence, we see that for a given transition matrix $\mathbb{P}$ we always arrive at $\Pi_t = \Pi_o \mathbb{P}^t = \alpha_1 e_1$ which is unique and independent of the Π_o we started with and will always converge.

7. Irreducible Markov Chains

A Markov Chain is irreducible if for any x,y $\in \Omega$, $\exists t$, such that

$$P(x, y)^t > 0$$

This denotes that the transition kernel for a given state is positive for at least one of the other state i.e there exists no isolated state. If there was such a state then we could decompose or reduce the system to two (or those many) markov chains.

8. Aperiodic Markov Chains

A Markov Chain is aperiodic if $\forall x, y \ \exists \Omega$ such that

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

If a chain is aperiodic, then it is possible to get to to state x for all $t = 1, 2, 3....n$ In particular, this means that we can define an aperiodic Markov chain if and only if we ensure non-zeros along the diagonal of $\mathbb{P}$. In this case, the distributions will oscillate but not converge. Figure 2 shows an example of an Aperiodic Markov Chain.

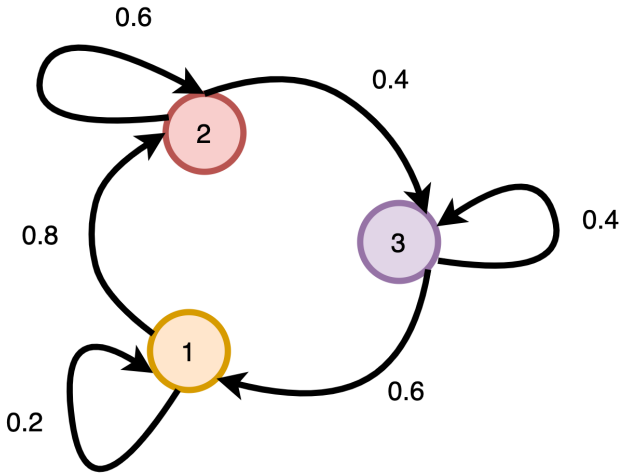

Figure 2: Aperiodic Markov Chain