CS 5234 Algorithms at Scale **September 1, 2020**

Lecture 3: Markov Chain Monte Carlo

Lecturer: Arnab Bhattacharyya *Scribe: Jiun Wei, Changsheng, Yuchen, Shawn*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications. They may be distributed outside this class only with the permission of the Instructor.*

3.1 Introduction

Recall that in the last lecture, we saw sampling applied to:

1. Approximate Counting
2. Bayesian Inference

For approximate counting, we wanted to uniformly sample from the space of valid solutions to a computational problem (e.g., the 0-1 Knapsack Problem).

For Bayesian inference, we wanted to sample from a posterior distribution P such that:

$$P(x) = \frac{f(x)}{Z}$$

where $f(x)$ is known but Z is unknown.

3.1.1 Markov Chain Monte Carlo

Both sampling tasks can be accomplished using a method called Markov Chain Monte Carlo (MCMC).

Suppose P is the target distribution that we wish to sample from over Ω, the universe of all possible states.

The main idea is that we start from a known state X_0 at $t = 0$ in Ω and make random transitions to X_1 at $t = 1$, X_2 at $t = 2$, X_3 at $t = 3$ and so on. Each transition to the next state is also based solely on the current state and not on the previous states that came before.

For MCMC, we want that for large t, $Pr[X_t = x] \approx P(x)$.

Assuming certain conditions are met, the sequence of X_0, X_1, X_2 and so on is called a Markov Chain.

Definition 3.1 (Markov Chain) *A Markov Chain over Ω is a sequence of random variables $X_0, X_1, X_2, ...$ such that:*

1. *$\forall t, X_t$ takes values in Ω.*
2. *$\forall x_0, x_1, ..., x_{t+1} \in \Omega : Pr[X_{t+1} = x_{t+1} | X_0 = x_0, X_1 = x_1, ..., X_t = x_t] = Pr[X_{t+1} = x_{t+1} | X_t = x_t]$ (Informally: the next move only depends on current state, not past history.)*
3. *$Pr[X_{t+1} = y | X_t = x] = M(x, y)$ is independent of t.*

The last point also allows us to formulate an $|\Omega| \times |\Omega|$ **transition matrix** M.

3.1.2 Example: Card Shuffling

We can think of multiple card shufflings as a Markov Chain where each permutation or ordering of the standard 52 cards in a deck constitutes a state. Card shuffling is a random process, and we hope that after enough shuffles, the deck is equally likely to be any of the states in Ω, where $\Omega =$ all possible ordering of the cards (i.e., $|\Omega| = 52!$). This was also demonstrated using the Jupyter notebook via simulated riffle shuffles.

3.2 Stationary Distributions

3.2.1 Example: Finite State Machine

Markov chains can be represented by finite state machines. The idea is that a Markov chain describes a process in which the transition to a state at time $t + 1$ depends only on the state at time t. The main thing to keep in mind is that the transitions in a Markov chain are probabilistic rather than deterministic, which means that you can't always say with perfect certainty what will happen at time $t + 1$.

Thus Markov Chains are usually represented by **state transition diagrams**. Consider a Markov chain with three possible states $0, 1$ and 2, i.e. $\Omega = \{0, 1, 2\}$ and the following transition probabilities.

$$\mathbf{P}_r = \begin{bmatrix} \frac{1}{2} & \frac{1}{4} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{2} & \frac{1}{4} \\ \frac{1}{4} & \frac{1}{4} & \frac{1}{2} \end{bmatrix} \tag{3.1}$$

the state transition diagram can be drawn as,

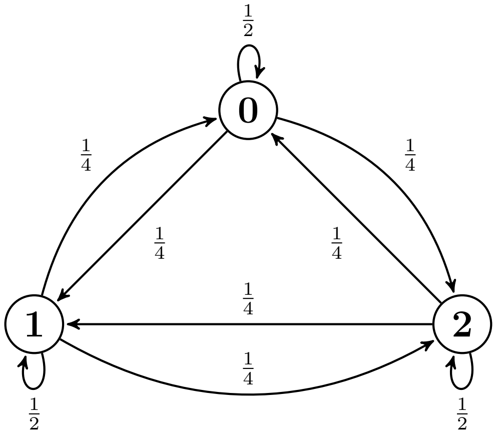
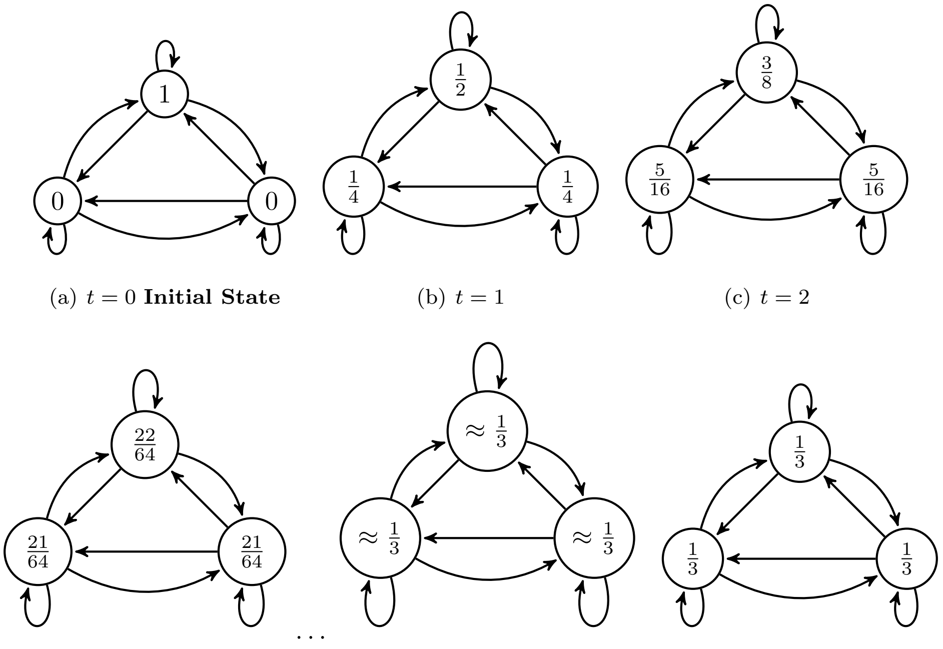


Figure 3.1: This graph shows an example of a finite state machine with 3 states. The edges denote the transition probabilities.

We eventually find that the probability of being at each state converges to $\frac{1}{3}$. This is called a **Stationary Distribution** because the resulting probability values are no longer changed by application of the transition probabilities.



(a) $t = 0$ **Initial State**

(b) $t = 1$

(c) $t = 2$

(d) $t = 3 \to t = n$

(e) **Stationary**

Figure 3.2: Above shows an example of the diffusion process of the above mentioned finite state machine. Suppose we start from state 0. In Figure 2(a), node 0 has an initial probability value 1 node 1 and node 2 have the initial value of 0, and in the next state, e.g. we can compute the value of node 0 by $\frac{1}{2} = \frac{1}{2} \cdot 1 + \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 0$. We can find that, as t advances and transitions take place, the probability values of being at each state start to converge to $\frac{1}{3}$ for all nodes.

Definition 3.2 (Stationary Distribution) *π is a stationary distribution if*

$$\forall y \in \Omega, \pi(y) = \sum_{x \in \Omega} \pi(x) \cdot M(x, y)$$

Question: As $t \to \infty$, does the distribution of X_t always converge to a stationary distribution?

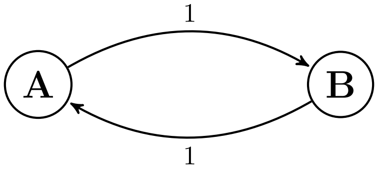


Figure 3.3: An example of a 2-state system. In this system, $\pi = [\frac{1}{2}\ \frac{1}{2}]$ is stationary.

The answer is no. For instance, consider a 2-state system that always transitions to the other state with probability 1. In this case, no convergence is ever achieved as $t \to \infty$. For instance, suppose we start at state A for $t = 0$. Then we will only be at state A for even values of t and state B for odd values of t. We never converge to $\pi = [\frac{1}{2}\ \frac{1}{2}]$ as the probability values will continue to evolve in a periodic manner. To ensure convergence, we need to prevent such periodicity from occurring.

Definition 3.3 (Regular) *A Markov Chain is* ***regular*** *if* $\exists\, T_0 > 0$ *such that for any two states* x *and* y*, there is a path of length* T_0 *from* x *to* y*.*

Theorem 3.4 *A finite regular Markov Chain has a unique stationary distribution* π *that it converges to if*

$$\forall x \in \Omega, \lim_{t \to \infty} Pr[X_t = x] = \pi(x)$$

All finite Markov Chains have stationary distributions, but not all converge to it.

MCMC Strategy: Design a Markov chain whose stationary distribution is the one you want to sample from.

Lemma 3.5 π *is a stationary distribution if it satisfies the "detailed balance" equation:*

$$\forall x, y : \pi(x) \cdot M(x,y) = \pi(y) \cdot M(y,x)$$

Proof: Suppose we have a Markov Chain with transition matrix M and a distribution π satisfying our hypothesis. Now consider the jth component of πM:

$(\pi M)_j = \sum_{i \in \Omega} \pi_i M(i,j) = \sum_{i \in \Omega} \pi_j M(j,i) = \pi_j \sum_{i \in \Omega} M(j,i) = \pi_j$

This implies that $\pi M = \pi$. ■

Corollary 3.6 *A Markov Chain has uniform stationary distribution if* $M(x,y) = M(y,x)\ \forall x, y$*.*

Proof: For a uniform distribution, $\pi(x) = \pi(y)$. Hence, by Lemma 3.5, for a uniform stationary distribution:

$\pi(x) \cdot M(x,y) = \pi(y) \cdot M(y,x) \implies M(x,y) = M(y,x)\ \forall x, y$ ■

3.3 Application to Knapsack Problem $KP(L)$

- Weights: $w_1, w_2, ..., w_n$
- Limit: L
- Ω = Solutions to KP(L)
- $X_0 = \emptyset$

And then we define the transition from t. Suppose we have X_t, and here is how we generate X_{t+1}.

At step t, choose $i \in [n]$ uniformly at random.

$$X_{t+1} = \begin{cases} X_t \setminus \{i\} & \text{if } i \in X_t \\ X_t \cup \{i\} & \text{if } i \notin X_t \text{ and } \sum_{k \in X_t \cup \{i\}} w_k \leq L \\ X_t & \text{otherwise} \end{cases}$$

We can easily check that this is regular. Firstly, check that we can go from any state to any other state, i.e. a path should exist. Secondly, there's a self-loop transition somewhere, i.e. $X_{t+1} = X_t$ for some states. Otherwise, we never violate the weights constraint, which means that $\sum_{k \in X_t \cup \{i\}} w_k \leq L$ will always be true.

According to Corollary 3.6, this Markov Chain's transition probabilities from state S to any adjacent state S' and vice versa are the same at $M(S,S') = M(S',S) = \frac{1}{n}$. And the Markov Chain converges to a uniform stationary distribution. Hence, it is possible to sample uniformly from the population of KP(L) solutions using the MCMC method.

A question to think about (but won't be discussed in this class): The claimable convergence is $t \to \infty$, but we will only running the Markov Chain on a smaller number of sets. So how close do we get to the uniformly distribution?

3.4 Metropolis-Hastings

Recall the Bayesian inference setup, where we want to generate samples from a distribution P, such that $P(x) = \frac{f(x)}{Z}, f : \Omega \to \mathbb{R}^{\geq 0}$. The function $f(x)$ can be thought of as a score that determines how desirable it is to be in state x, while Z is some normalizing constant that is prohibitively expensive to compute.

The Metropolis-Hastings algorithm allows us to design a Markov Chain with stationary distribution P, thus providing us with a method of sampling from P without having to compute Z. Suppose you have a regular Markov Chain on Ω with transition matrix M. We define the Metropolis-Hastings Markov Chain transitions as follows:

1. Pick Y with probability $M(X_t, Y)$.
2. Set $\alpha = \min\{1, \frac{P(Y)\cdot M(Y,X_t)}{P(X_t)\cdot M(X_t,Y)}\}$, using $\frac{P(Y)}{P(X)} = \frac{f(Y)}{f(X)}$
3. With probability α, $X_{t+1} \leftarrow Y$ and with probability $1-\alpha$, $X_{t+1} \leftarrow X_t$.

Theorem 3.7 *The stationary distribution π of the Metropolis-Hastings Markov Chain is P*

Proof: Let M be the transition matrix of the original Markov chain, and let Q be the transition matrix of the Metropolis-Hastings Markov chain. Then:

$$\begin{aligned} P(i)\cdot Q(i,j) &= P(i)\cdot M(i,j)\cdot \min\{1, \frac{P(j)\cdot M(j,i)}{P(i)\cdot M(i,j)}\} = \min\{P(i)\cdot M(i,j), P(j)\cdot M(j,i)\} \\ &= P(j)\cdot M(j,i)\cdot \min\{\frac{P(i)\cdot M(i,j)}{P(j)\cdot M(j,i)}, 1\} = P(j)\cdot Q(j,i) \end{aligned}$$

By the detailed balance lemma, P is the stationary distribution with respect to Q ■