Counting and Sampling **Fall 2017**

Lecture 3: Introduction to Markov Chains

Lecturer: Shayan Oveis Gharan *October 4th*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

In this lecture we define Markov chains and discuss their properties. We also discuss several important classes of Markov chains. We start by a real-world application of Markov chains in text decoding. See section 1 of Persi's Markov chain Monte Carlo Revolution monograph.

There is a huge interests in studying and using Markov chains in sampling tasks. Here are a few reasons:

- They are typically very easy to implement as most of the Markov chains follow one of the well-known rules, e.g., Metropolis, or Heat-Bath.
- They use very small amount of memory. One only needs to remember the configuration of the last visited state.
- Typically, they "mix" very fast, and they return a near random sample.

In practice one may use Markov chains to generate random inputs for a programming task, or to study typical configurations in a physical system.

3.1 Markov Chains

One can think of a Markov chain as a stochastic process on a set of states Ω. Say X_t represent the location of the process at time t. If $X_t = x$ for some $x \in \Omega$, the Markov chain makes a transition to the next step according the a probability distribution $K(x, .)$. That is, for all states y,

$$\mathbb{P}\left[X_{t+1} = y | X_t = x\right] = K(x, y).$$

The most important property of a Markov chain is the *Markov property.* That is the location of the process at time $t+1$ only depends on the location at time t; it is independent of the rest of the history:

$$\mathbb{P}\left[X_{t+1} | X_0, \dots, X_t\right] = \mathbb{P}\left[X_{t+1} | X_t\right].$$

The operator K is usually called the *Markov Kernel.* It can be seen as a $|\Omega| \times |\Omega|$ stochastic matrix.

We can represent a Markov chain as a weighted directed graph where there is a vertex x for each state x. For any pair of states x, y, if $K(x, y) > 0$, there is an edge from x to y with weight $K(x, y)$. In this sense the Markov chain can be seen as a Random walk in this directed graph where at every vertex x we choose one of the incident edges pointing out of x with probability proportional to its weight.

Let p be a probability distribution vector over Ω. It follows that if we sample $X_0 \sim p$, then X_1 is distributed according to pK, i.e., for all x

$$\mathbb{P}\left[X_1 = x | X_0 \sim p\right] = \sum_y p(y) K(y, x) = p^T K(x).$$

With this notation, the evolution of the Markov chain can be dened in terms of matrix-vector equations. In particular, for all $t \geq 0$,

$$\mathbb{P}\left[X_t = y | X_0 \sim p\right] = p^T K^t.$$

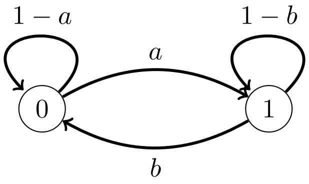
In particular, $K^t(x, y)$ is the probability that a walk started at x goes to y after t steps.

3.2 Stationary Distribution

Definition 3.1 (Stationary Distribution)**.** *We say a probability distribution π is a stationary distribution of the kernel K if $\pi K = \pi$. In other words, π is a stationary distribution if for all states x,*

$$\pi(x) = \sum_y \pi(y) K(y, x).$$

Example 3.2. *Let us give an example: Consider the following graph The stationary distribution is $\pi(0) =$*



$\frac{b}{a+b}, \pi(1) = \frac{a}{a+b}$.

In the following lemma we show that every Markov chain has a stationary distribution.

Lemma 3.3. *Any markov chain with kernel K has at least one stationary distribution.*

Proof. We need to show that there is a nonnegative vector π such that $\pi K = \pi$. Note that if $\sum_x \pi(x) \neq 1$, we can just normalize the sum to 1.

First of all, since K is a stochastic matrix, $K\mathbf{1} = \mathbf{1}$, i.e., 1 is an eigenvalue of K. Therefore, 1 is also an eigenvalue of K^T (because the characteristic polynomials of K and K^T are equal). It follows that there exists a vector p such that $K^T p = pK^T = p$.

We claim that the vector $p_+ = \max\{p, 0\}$ is also an eigenvector of K^T with eigenvalue 1. Firstly, observe that for any x,

$$p_+(x) = |p(x)| = \left| \sum_y p(y) K(y, x) \right| \leq \sum_y |p(y)| K(y, x) = \sum_y p_+(y) K(y, x). \tag{3.1}$$

We show that all of the above inequalities must be equality. By stochasticity of K,

$$\sum_y p_+(y) = \sum_y p_+(y) \left(\sum_x K(y, x) \right) = \sum_x \sum_y p_+(y) K(y, x) \geq \sum_x p_+(x).$$

So, the last inequality is an equality and all inequalities in (3.1) must be equalities. □

We will see later that an important class of Markov chains have a unique stationary distribution. This property is called ergodicity.

Definition 3.4 (Reversible Markov Chains)**.** *We say a Markov chain is reversible if there exists a nonnegative weight function $\pi : \Omega \to \mathbb{R}_+$ such that for all $x, y \in \Omega$,*

$$\pi(x)K(x,y) = \pi(y)K(y,x).$$

It follows that if a Markov chain is reversible then π is the stationary distribution of the Markov chain. This is because for any state x,

$$\sum_y \pi(y)K(y,x) = \sum_y \pi(x)K(x,y) = \pi(x).$$

Note that as a special case if for all x, y, $K(x,y) = K(y,x)$, then $\pi(.)$ is the uniform distribution.

Most of the Markov chains that we study in this course are reversible. So, it is very easy to find their stationary distribution. In fact, as we will see soon there are recipes to construct Markov chains for any given space Ω such that the stationary distribution is the desired distribution $w(x)/Z$.

Reversible Markov Chains are the same as Random Walks Suppose K is a reversible Markov chain on a state Ω. We claim that we can construct an undirected weighted graph $G = (V, E, w)$ and simulate the chain by following a random walk on G, where at every vertex x we choose next vertex y with probability $\frac{w(x,y)}{\sum_z w(x,z)}$. Note that here $w(x,y) = w(y,x)$.

It is enough to define w. For all x, y, let

$$w(x,y) = \pi(x)K(x,y) = \pi(y)K(y,x).$$

We claim that condition on $X_t = x$, the law of X_{t+1} is the same as if we run a random walk on G. In particular, suppose the random walk is at a vertex x, the probability that it goes to y in the in the next step is

$$\frac{w(x,y)}{\sum_z w(x,z)} = \frac{\pi(x)K(x,y)}{\sum_z \pi(x)K(x,z)} = \frac{K(x,y)}{\sum_z K(x,z)} = K(x,y).$$

3.3 Mixing of Markov Chains

Definition 3.5 (Irreducible Markov Chains)**.** *A Markov chain is irreducible if for all x, y, there exists some t such that $K^t(x,y) > 0$. Equivalently, the graph corresponding to K is strongly connected.*

Note that if the chain is reversible, and the corresponding graph is undirected then irreducibility corresponds to showing that the underlying graph is connected.

Definition 3.6. *A Markov chain is aperiodic if for all x, y we have* $\gcd\{t : K^t(x,y) > 0\} = 1$.

Lemma 3.7. *Let K be an irreducible, aperiodic Markov chain. There exists $t > 0$ such that for all x, y,*

$$K^t(x,y) > 0.$$

We leave this lemma as an exercise. This lemma will be crucially used in the proof of the following fundamental theorem of Markov chains.

Theorem 3.8 (Fundamental Theorem of Markov Chains)**.** *Any irreducible and aperiodic Markov chain has a unique stationary distribution. Furthermore, for all x, y,*

$$K^t(x,y) \to \pi(y)$$

as t goes to infinity. In particular, for any $\epsilon > 0$ there exists $t > 0$ such that $\|K^t(x,.) - \pi\|_{TV} \leq \epsilon$.

In light of this theorem we shall refer to an irreducible and aperiodic Markov chain as *ergodic*.

There is a general idea to make any given Markov chain aperiodic. All we need to do is to add self-loop at all states; in other words, at any state x we stay with probability $1/2$ and we follow the transition kernel K with probability $1/2$. This new Markov chain is called the lazy chain.

3.4 Examples of Markov Chains

Random Walks on Undirected Graphs Consider an undirected graph $G = (V, E)$. A random walk on G, at any vertex v move to a uniformly random neighbor. The following facts are immediate:

- The chain is irreducible if G is connected.
- The chain is aperiodic if G is not bipartite.
- The chain is reversible with distribution $\pi(v) = d(v)/2|E|$.

These chains have had many applications in theory of CS, e.g., to test if a vertex $v \in V$ is connected to a vertex $u \in V$ in logarithmic space [Rei08], or to find local clusters around a given vertex [ST13, AP09].

3.5 Card Shuffling

One of the very important real-world applications of Markov chain technique has been in card shuffling. Suppose we have a deck 52 cards and we want to shuffle them to make them "random". Ideally, we would like to have one permutation out of all 52!. The Markov chain techniques suggest to use the total variation distance as a measure of randomness. We will talk about the mixing time of a few card shuffling techniques. Let us here just introduce a few techniques.

Random Transposition: Pick two cards i and j uniformly at random (with replacement) and swap them.

This Markov chain is obviously irreducible and aperiodic. It is also reversible, because for all pair of states x, y, $K(x, y) = K(y, x)$, i.e., if we go from x to y we can go back to x by choosing the same transposition. So, it has a uniform stationary distribution.

Top to Random: Take the top card and insert it at one of the n positions in the deck chosen uniformly at random.

This walk is irreducible and aperiodic, but it is not reversible. In fact once we move the top card to a position there is no way to go back to the previous state. Nonetheless, we claim that its stationary distribution is uniform. This is because in the corresponding directed graph the indegree and outdegree of ever vertex is n and the probability of choosing each particular transition is $1/n$. In particular, observe that for any fixed permutation σ there are exactly n permutations that move to σ in one step of the Top-to-Random walk.

Riffle Shuffle. (Gilbert-Shannon-Reeds [Gi55,Re81]) The riffle shuffle is defined as follows:

- Split the deck into two parts according to the binomial distribution $\mathrm{Bin}(n, 1/2)$.

- Drop cards in sequence, where the next card comes from the left hand L (resp. right hand R) with probability $\frac{|L|}{|L|+|R|}$ (resp. $\frac{|R|}{|L|+|R|}$).

Similar to the Top-to-Random this walk is irreducible, aperiodic, but not reversible. The stationary distribution is uniform because the outdegree and indegree of all vertices are equal. We will say more about this walk later. We will see this walk mixes much faster than the previous two walks.

3.6 The Metropolis Rule

Suppose we have a state space Ω together with a weight function $w : \Omega \to \mathbb{R}_+$. We would like to sample from the distribution $\pi(x) = w(x)/Z$, where as usual Z is the partition function. The Metropolis rule is a general recipe to construct an ergodic Markov chain with stationary distribution $\pi(.)$. To construct the Metropolis chain we need to ingredients:

Neighborhood Structure: The first requirement is a connected undirected graph $G = (\Omega, \mathcal{E})$ on the state space. Typically, two elements $x, y \in \Omega$ are connected if they different by some local changes. For example, if Ω represents all matchings, two matchings are connected if they different in a single or two edges. Note that we need G to be undirected to get a reversible Markov chain and connected to get an irreducible chain.

Proposal Distribution At any vertex x we require a *proposal* distribution, $p(x, .)$ satisfying the following properties:

- $p(x, y) > 0$ only if y is a neighbor of x.
- $p(x, y) = p(y, x)$ for all y.
- $\sum_y p(x, y) = 1$.

As we elaborate below, at the state x we try to move by choosing a neighbor y based on the proposal distribution. But this proposal may not be accepted.

Now, we are ready to define the Metropolis chain:

- At a state x we choose a neighbor y with probability $p(x, y)$, and we propose to move to y.
- We accept this proposal with probability $\min\{1, \frac{\pi(y)}{\pi(x)}\}$ and we reject and stay at x with the remaining probability.

Having this idea in mind the Metropolis rule is reminiscent of the simulated annealing ideas used in numerical optimization. At a state x we choose a neighbor y by a local move; if y has a higher probability we always move to y, otherwise we move to y with the ratio of probability y to x.

In the next lemma we show that the Metropolis chain is always irreducible.

Lemma 3.9. *For any Ω and any connected undirected graph $G = (\Omega, \mathcal{E})$, the Metropolis chain is reversible with stationary distribution π.*

Proof. It is enough to show that for any pair of states x, y,

$$\pi(x)K(x, y) = \pi(y)K(y, x).$$

First of all, if y is not a neighbor of x in G then we never move to y, so $K(x,y) = 0$. Since G is undirected, $K(y,x) = 0$ as well. Now, consider a y that is a neighbor of x. We have

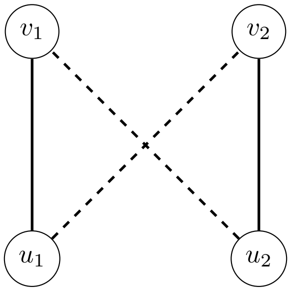
$$K(x,y) = p(x,y)\min\left\{1, \frac{\pi(y)}{\pi(x)}\right\}.$$

Now, we consider two cases. If $\pi(y) \le \pi(x)$, then,

$$\begin{aligned}
\pi(x)K(x,y) &= \pi(x)p(x,y)\frac{\pi(y)}{\pi(x)} \\
&= \pi(y)p(y,x) && (p(x,y) = p(y,x)) \\
&= \pi(y)p(y,x)\min\left\{1, \frac{\pi(x)}{\pi(y)}\right\} && (\pi(y) \le \pi(x)) \\
&= \pi(y)K(y,x).
\end{aligned}$$

If $\pi(y) < \pi(x)$, a similar proof works out. □

Perfect Matchings using Random Transposition Walk Suppose we have a bipartite graph $G = (V,E)$ and we want to construct a Markov chain to generate a uniformly random perfect matching in G. Consider the following Metrpolis Rule: For every matching M, a matching M' is in the neighborhood of M if we can obtain M' from M by a random transoposition, that is we substitiute two edges $(u_1, v_1), (u_2, v_2) \in M$ with edges $(u_1, v_2), (v_1, u_2)$. It follows that this chain is reversible with π be the uniform distribution. But unfortunately, the chain is not necessarily irreducible for any bipartite graph G. In particular, if G is the cycle C_n, for n even, it has exactly two perfect matchings and these matchings are disjoint. So, there is no local move to move between these two matchings. One has to change all $n/2$ edges of a matching to move from one to the other.



References

[AP09] Reid Andersen and Yuval Peres. Finding sparse cuts locally using evolving sets. In *STOC*, pages 235–244, 2009. 3-4

[Rei08] Omer Reingold. Undirected connectivity in log-space. *J. ACM*, 55(4):1–24, September 2008. 3-4

[ST13] Daniel A. Spielman and Shang-Hua Teng. A Local Clustering Algorithm for Massive Graphs and Its Application to Nearly Linear Time Graph Partitioning. *SIAM J. Comput.*, 42(1):1–26, 2013. 3-4