CSE 525: Randomized Algorithms **Spring 2023**

Lecture 15: Introduction to Markov Chains

Lecturer: Shayan Oveis Gharan *May 16th*

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 recommend Persi's Markov chain Monte Carlo Revolution monograph. for real-world applications of Markov Chains.

15.1 Markov Chains

A Markov chain is 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 $K(x, .)$, called the Markov kernel. 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:

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

K can be seen as a $|\Omega| \times |\Omega|$ stochastic matrix.

Definition 15.1 (Reversible Markov Chains)**.** *We say a Markov chain is reversible if there exists a non-negative weight function $w : \Omega \to \mathbb{R}_+$ such that for all $x, y \in \Omega$,*

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

In this course we only work with reversible Markov chains.

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.

Definition 15.2 (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,*

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

It follows that if a Markov chain is reversible then w is proportional to 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.

Reversible Markov Chains are the same as Random Walks Suppose K is a reversible Markov chain on a state Ω with weight function w. We claim that we can construct an undirected weighted graph $G = (V, E, c)$ for $c : E \to \mathbb{R}_{\geq 0}$ and simulate the chain by following a random walk on G, where at every vertex x we choose next vertex y with probability $\frac{c(x,y)}{\sum_z c(x,z)}$. Note that here $c(x,y) = c(y,x)$.

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

$$c(x,y) = w(x)K(x,y) = w(y)K(y,x).$$

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

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

There is a general idea to make any given (not necessarily reversible) 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.

Let K be the Markov Kernel for a reversible Markov chain corresponding to a weighted graph $G = (V, E, w)$. It turns out that the stationary distribution of the chain is,

$$\pi(v) = d_w(v)/d_w(V)$$

15.2 Mixing Time of Markov Chains

Definition 15.3 (Ergodic Markov Chains)**.** *A reversible Markov chain is ergodic if the underlying graph is connected and not bipartite.*

Theorem 15.4 (Fundamental Theorem of Markov Chains)**.** *Any irreducible reversible Markov chain has a unique stationary distribution, π. 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 $\|K^t(x,.) - \pi\|_{TV} \leq \epsilon$, where for two probability distribution μ, ν we write

$$\|\mu - \nu\|_{TV} := \sum_{x:\mu(x)>\nu(x)} (\mu(x) - \nu(x)) = \frac{1}{2}\|\mu - \nu\|_1 = \max_{\mathcal{A}\subseteq\Omega} \mu(\mathcal{A}) - \nu(\mathcal{A}).$$

One of the major questions about Markov chains is how long does it take to mix, i.e., to get ϵ-close to the stationary distribution in total variation distance.

For a state x let

$$\tau_x(\epsilon) = \min\{t : \|K^t(x,.) - \pi\|_{TV} \leq \epsilon\}$$

be the first time that the total variation distance of the walk started at x from the stationary distribution drops below ϵ. Now, define

$$\tau(\epsilon) = \max_x \tau_x(\epsilon).$$

Definition 15.5 (Mixing Time)**.** *For any Markov chain the mixing time is defined as $\tau(1/2e)$. In other words, this is the time that the total variation distance of the walk started at the worst possible starting point drops be low $1/2e$.*

The choice of constant $1/2e$ is for algebraic convenience but as we will see this will only change the mixing time up to a constant.

15.3 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 $\text{Bin}(n, 1/2)$.
- 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.

15.4 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 15.6. *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 $\pi(y) \leq \pi(x)$, then,

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

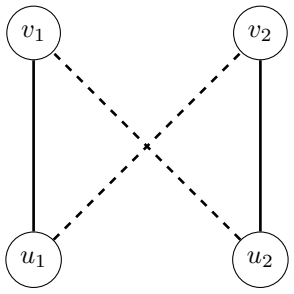
If $\pi(y) < \pi(x)$, a similar proof works out. □

Applications There is a huge interests in studying and using Markov chains, specifically the Metropolis rule 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.

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.

15.5 Coupling

Definition 15.7 (Coupling)**.** *Let μ, ν be probability distributions over Ω. A coupling between μ, ν is a probability distribution π on $\Omega \times \Omega$ that preserves the marginals of μ, ν respectively. In particular, for all*

$x \in \Omega$,

$$\sum_y \pi(x,y) = \mu(x) \text{ and } \sum_y \pi(y,x) = \nu(x).$$

Let us give an example: Consider the following two distributions over $\{1,2,3\}$. Consider the following

	1	2	3
$\mu(.)$	0.2	0.5	0.3
$\nu(.)$	0.3	0.4	0.3

coupling $\pi(1,1) = 0.2, \pi(2,2) = 0.4, \pi(3,3) = 0.3, \pi(2,1) = 0.1$. Observe that all marginal probabilities are satisfied; for example $\pi(2,2) + \pi(2,1) = 0.5 = \mu(2)$. Furthermore, if (X,Y) is a sample of π, we have that $\mathbb{P}_\pi[X = Y] = 0.9$. You can compare this coupling with an independent coupling in which $\tilde{pi}(i,j) = \mu(i)\nu(j)$. In that case we would have $\mathbb{P}_{\tilde{\pi}}[X = y] = \pi(1)\nu(1) + \pi(2)\nu(2) + \pi(3)\nu(3) = 0.35$.

Lemma 15.8 (Coupling Lemma)**.** *Let μ and ν be probability distributions on Ω, and let X and Y be random variables with distributions μ and ν, respectively. Then*

1. $\mathbb{P}[X \neq Y] \geq \|\mu - \nu\|_{TV}$.
2. *There exists a coupling between μ and ν such that $\mathbb{P}[X \neq Y] = \|\mu - \nu\|_{TV}$.*

Proof. We prove the 2nd part, that is we construct the optimal coupling between μ and ν. The coupling will be very similar to the above example. For any i in the support of these distributions, we let $\pi(i,i) = \min\{\mu(i), \nu(i)\}$. For all other pairs $i \neq j$ we sequentially choose $\pi(i,j)$. Obviously, there is a way to match the remaining mass in the distributions such that all marginals are preserved. For example, when we process $i \neq j$ we define $\pi(i,j) = \min\{\mu(i) - \pi(i,.), \nu(j) - \pi(.,j)\}$.

Note that obviously, this coupling has the highest probability that $X = Y$ because for any i, we must have

$$\mathbb{P}_{(X,Y)\sim\pi}[X = i, Y = i] \leq \min\{\mu(i), \nu(i)\}$$

in order to satisfy the marginal of i. Therefore,

$$\mathbb{P}[X \neq Y] = 1 - \sum_i \pi(i,i) = \sum_i \mu(i) - \min\{\mu(i), \nu(i)\} = \|\mu - \nu\|_{TV}.$$

□

References