6.896 Probability and Computation February 7, 2011

Lecture 2

Lecturer: Constantinos Daskalakis *Scribe: Alessandro Chiesa & Zeyuan Zhu*

NOTE: The content of these notes has not been formally reviewed by the lecturer. It is recommended that they are read critically.

Administrivia

- There are no scribe notes from Lecture 1. Please refer to the slides on the website.
- Ankur Moitra (`moitra@mit.edu`) is the TA for the class.
- Please do not forget to register for the class if you are listening.
- Exercises are due in class **one week** after they are given in lecture.

1 The MCMC Paradigm

We are interested in studying a general class of sampleability problems:

- **input:** a (large) finite set Ω and a positive weight function $w \colon \Omega \to \mathbb{R}^+$;
- **goal:** sample an element x in Ω with probability $\pi(x)$ that is proportional to its weight $w(x)$, i.e., with probability

$$\pi(x) \stackrel{\text{def}}{=} \frac{w(x)}{Z_w} \ ,$$

where $Z_w \stackrel{\text{def}}{=} \sum_{x \in \Omega} w(x)$ is called the *partition function* of the weight function w in Ω.

In the *Markov-Chain Monte-Carlo (MCMC) Paradigm*, we consider a general solution approach to this class of sampleability problems, where we construct a sequence of random variables $\mathcal{X} = (X_t)_{t \in \mathbb{N}}$ that converges to the target probability distribution π, i.e., for which

$$\lim_{t \to +\infty} \left(\Pr\left[X_t = y \middle| X_0 = x \right] \right) = \pi(y) \ .$$

Note that the limit probability distribution π is independent of the "starting point" X_0. We think of t as a "time" variable.

2 Basic Definitions

More formally:

Definition 1. *A* Markov chain *on a finite set Ω is a stochastic process $\mathcal{X} = (X_0, X_1, \dots, X_t, \dots)$ such that the following three conditions are satisfied:*

1. *for each time $t \in \mathbb{N}$, the random variable X_t takes on values within Ω;*
2. *for each time $t \in \mathbb{N}$, for every $x_0, x_1, \dots, x_{t+1} \in \Omega$, the probability that the random variable X_{t+1} takes on the value x_{t+1}, conditioned on X_i taking on the value x_i for $0 \leq i \leq t$, is equal to the probability that the random variable X_{t+1} takes on the value x_{t+1}, conditioned only on X_t taking on the value x_t, i.e.,*

$$\Pr\left[X_{t+1} = x_{t+1} \middle| X_t = x_t \, , \, X_{t-1} = x_{t-1} \, , \, \dots \, , \, X_0 = x_0 \right] = \Pr\left[X_{t+1} = x_{t+1} \middle| X_t = x_t \right] \ ; \text{ and}$$

3. for every two elements x and y in Ω, $\Pr\left[X_{t+1} = y \middle| X_t = x\right]$ is independent of t.

Informally, in a Markov chain, given the "present" state, "future" and "past" states are independent. In particular, it makes sense to define the transition probability from an element in Ω to another (possibly the same) element in Ω:

Definition 2. *For any two elements x and y in Ω, the* transition probability *from x to y in the Markov chain $\mathcal{X}$ is*

$$P_{\mathcal{X}}(x,y) \stackrel{\text{def}}{=} \Pr\left[X_{t+1} = y \middle| X_t = x\right] \ .$$

That is, given that at some time t, the chain is at element x, $P_{\mathcal{X}}(x,y)$ gives the probability that the chain will be at time y at the next time $t+1$. (Again, this probability is independent of time.)

The transition probabilities in $\mathcal{X}$, ranging over all pairs of elements in Ω, induce a transition matrix:

Definition 3. *The* transition matrix *of a Markov chain $\mathcal{X}$ is the $|\Omega| \times |\Omega|$ matrix with $[0,1]$ entries defined as follows:*

$$P_{\mathcal{X}} \stackrel{\text{def}}{=} \{P_{\mathcal{X}}(x,y)\}_{x,y \in \Omega} \ .$$

Observe that the transition matrix $P_{\mathcal{X}}$ of a Markov chain $\mathcal{X}$ is a *stochastic matrix*, because it satisfies the following two properties:

1. *non-negativity:* for every two elements x and y in Ω, $P_{\mathcal{X}}(x,y) \geq 0$; and
2. *stochasticity:* for every element in x in Ω, $\sum_{y \in \Omega} P_{\mathcal{X}}(x,y) = 1$ (i.e., each row "adds up to 1").

Example 1 (Card Shuffling)**.** *When considering the problem of shuffling cards, say, via the riffle shuffle, we can set:*

- $\Omega \stackrel{\text{def}}{=} \{$ *all possible permutations of the deck* $\}$,
- $w(x) \stackrel{\text{def}}{=} 1$ *for every deck permutation x (so that $\pi(x)$ is the uniform distribution over all the deck permutations),*
- X_0 *is the initial configuration of the deck (which can be arbitrary), and*
- X_t *is the state of the deck after the t-th riffle, for $t \in \mathbb{N}$.*

Note that, indeed, X_{t+1} is independent of $X_{t-1}, \ldots, X_0$ when conditioning on X_t.

For each element $x \in \Omega$ and time $t \in \mathbb{N}$, we will denote by the row vector $p_x^{(t)} \in \mathbb{R}_+^{1 \times |\Omega|}$ the distribution of X_t when conditioned on $X_0 = x$. The following relations hold for every time $t \in \mathbb{N}$:

$$\begin{aligned} p_x^{(t+1)} &= p_x^{(t)} P_{\mathcal{X}} \ \text{ and} \\ p_x^{(t)} &= p_x^{(0)} P_{\mathcal{X}}^t \ . \end{aligned}$$

These two relations characterize how the Markov chain evolves over time.

3 Markov Chains as Graphs

It is often useful to give a combinatorial interpretation of a Markov chain. Specifically, given a Markov chain $\mathcal{X}$, we can define a weighted directed graph $G(\mathcal{X})$ defined as follows:

- the vertex set V is identified with the element set Ω;
- the edges set E contains the directed edge (x,y) if and only if $P_{\mathcal{X}}(x,y) > 0$ (and, if so, this edge is assigned weight equal to $P_{\mathcal{X}}(x,y)$, i.e., the transition probability from x to y).

Note that self-loops are allowed (i.e., when $P_{\mathcal{X}}(x,x) > 0$). While we have defined the graph $G(\mathcal{X})$ as a weighted graph, much of the theory of Markov chains only depends on the *topology* of $G(\mathcal{X})$, rather than on the particular weights on its edges. Indeed, we will even call a graph $G(\mathcal{X})$ *undirected* if both $P_{\mathcal{X}}(x,y) > 0$ and $P_{\mathcal{X}}(y,x) > 0$ for every two elements x and y in Ω, ignoring the potential difference between the weights of the two edges.

Example 2 (Card Shuffling)**.** *When considering the problem of shuffling cards, say, via the riffle shuffle, $G(\mathcal{X})$ consists of a (huge!) weighted directed graph where the set of vertices consists of every possible permutation of the deck, and the directed edges indicate whether it is possible (i.e., with positive probability) to go from a certain deck configuration to another deck configuration through a single riffle.*

Note that not every edge is present! For example, the vertices 1234 *and* 4132 *do not share any edges, because there is no way to move between* 1234 *and* 4132 *(at least two "cuts" are needed). Also, while there is an edge from* 123456 *to* 142536*, there is not an edge to "come back" (again, for at least two cuts would be needed).*

4 Basic Properties

We now introduce basic but very important characterizations of Markov chains.

Definition 4. *Let $\mathcal{X}$ be a Markov chain. For every time $t \in \mathbb{N}$ and every two elements x and y in Ω, we define $P^t_{\mathcal{X}}(x,y)$ to be the probability that the chain will be at y at time $t_0 + t$, conditioned on the chain being at x at time t_0, for some time t_0. (Again, this probability does not depend on t_0.)*

Definition 5. *A Markov chain $\mathcal{X}$ is* irreducible *if (and only if), for every two elements x and y in Ω, there exists some time $t \in \mathbb{N}$ such that $P^t_{\mathcal{X}}(x,y) > 0$.*

Note that, equivalently, a Markov chain $\mathcal{X}$ is irreducible if and only if its graph $G(\mathcal{X})$ is strongly connected. (And, in case the graphical representation is undirected, strong connectivity "collapses" to connectivity.)

Definition 6. *A Markov chain $\mathcal{X}$ is* aperiodic *if (and only if), for every two elements x and y in Ω, it holds that*

$$\gcd\left\{t \in \mathbb{N} : P^t_{\mathcal{X}}(x,y) > 0\right\} = 1 \ .$$

Definition 7. *Let $\mathcal{X}$ be a Markov chain. For any element x in Ω, define the* period *of x as*

$$\gcd\left\{t \in \mathbb{N} : P^t_{\mathcal{X}}(x,x) > 0\right\} \ .$$

Lemma 1. *Let $\mathcal{X}$ be an irreducible Markov chain such that $G(\mathcal{X})$ is undirected. Then $\mathcal{X}$ is aperiodic if and only if $G(\mathcal{X})$ is non-bipartite.*

Proof: Suppose that $\mathcal{X}$ is bipartite. Then, choosing x and y "on the same side", we see that every path from x to y will have even length (and at least one such path exists because $\mathcal{X}$ is irreducible). Hence, $\gcd\{t \in \mathbb{N} : P^t_{\mathcal{X}}(x,y) > 0\} \geq 2 > 1$, thereby establishing that $\mathcal{X}$ is periodic.

Conversely, suppose that $\mathcal{X}$ is non-bipartite so that there exists some cycle C of odd length in $G(\mathcal{X})$. Consider any two elements x and y in Ω. Take any path p_x to from x to an element $c \in C$ and any path p_y from c to y. (Both paths exist because $\mathcal{X}$ is irreducible.) We can now move from x to y in at least two ways. In the first way, we can follow p_x, then go around C, and then follow p_y; the total length is $|p_x| + |C| + |p_y|$. In the second way, we can follow p_x, then go half-way around C and back, and then follow p_y; the total length is now $|p_x| + 2 \cdot \frac{|C|-1}{2} + |p_y|$. However, the greatest common divisor of $|p_x| + |C| + |p_y|$ and $|p_x| + (|C| - 1) + |p_y|$ must be equal to 1, because they are consecutive integers. As x and y were arbitrary elements in Ω, we conclude that $\mathcal{X}$ must be aperiodic. □

Lemma 2. *Let $\mathcal{X}$ be a Markov chain. If $\mathcal{X}$ is irreducible, then all the elements in Ω have the same period.*

Exericse (1pt): Prove Lemma 2.

Lemma 3. *Let $\mathcal{X}$ be an irreducible Markov chain. Then $\mathcal{X}$ is aperiodic if and only if there exists some time $t \in \mathbb{N}$ such that $P^t_{\mathcal{X}}(x,y) > 0$ for every two elements x and y in Ω.*

Exericse (1pt): Prove Lemma 3. (Hint: Review the Coin Problem and the Frobenius number.)

Lemma 4. *Let $\mathcal{X}$ be a Markov chain. If $\mathcal{X}$ is irreducible and contains at least one self-loop (i.e., $P_{\mathcal{X}}(x,x) > 0$ for some $x \in \Omega$), then $\mathcal{X}$ is aperiodic.*

Proof: For any two elements x and y in Ω, any self-loop on an element z easily allows for creating two paths from x to y with two respective lengths that are consecutive integers, forcing their greatest common divisor to be equal to 1, and thus forcing $\mathcal{X}$ to be aperiodic. (And the condition of irreducibility is used to ensure the existence of such paths.) □

5 Mixing of Markov Chains

We begin analyzing the mixing properties of a Markov chains. First, we introduce the notion of a stationary distribution:

Definition 8. *Let $\mathcal{X}$ be a Markov chain. A probability distribution π over Ω is a* stationary distribution *for $\mathcal{X}$ if $\pi = \pi P_{\mathcal{X}}$.*

Next, we state an important basic limit theorem:

Theorem 1 (Fundamental Theorem of Markov Chains)**.** *If a Markov chain $\mathcal{X}$ is irreducible and aperiodic, then it has a unique stationary distribution π. In particular, π is the unique (ℓ_1-normalized) left eigenvector of $P_{\mathcal{X}}$ corresponding to the eigenvalue 1. Moreover, $\lim_{t\to+\infty} P^t_{\mathcal{X}}(x,y) = \pi(y)$ for every two elements x and y in Ω.*

Because of this theorem, we shall sometimes refer to an irreducible, aperiodic Markov chain as *ergodic*. In the next lecture, we will give a combinatorial proof of the theorem. Today, we give an algebraic proof of a slightly weaker theorem, where we make the additional assumption that the Markov chain $\mathcal{X}$ is *reversible* with respect to a given distribution π:

Definition 9. *Let $\mathcal{X}$ be a Markov chain and let $\pi > 0$ be a probability distribution over Ω. The Markov chain $\mathcal{X}$ is* reversible *with respect to π if, for every two elements x and y in Ω,*

$$\pi(x)P_{\mathcal{X}}(x,y) = \pi(y)P_{\mathcal{X}}(y,x) \ .$$

Note that any Markov chain $\mathcal{X}$ where its transition matrix $P_{\mathcal{X}}$ is symmetric is trivially reversible (with respect to the uniform distribution π). Intuitively, in a Markov chain that is reversible with respect to π, the probability mass that flows from node x to node y is the same as the one that flows from y back to x. Therefore, we expect that π is stationary under the transition matrix $P_{\mathcal{X}}$:

Lemma 5. *If a Markov chain $\mathcal{X}$ is reversible with respect to a probability distribution π over Ω, then π is a stationary distribution for $\mathcal{X}$.*

Proof: Using the definition of reversibility, for any $y \in \Omega$,

$$\sum_{x\in\Omega} \pi(x)P_{\mathcal{X}}(x,y) = \sum_{x\in\Omega} \pi(y)P_{\mathcal{X}}(y,x) = \pi(y)\sum_{x\in\Omega} P_{\mathcal{X}}(y,x) = \pi(y) \ .$$

We therefore conclude that $\pi P_{\mathcal{X}} = \pi$. □

Reversible Markov chains can be represented also using *ergodic flows*:

Definition 10. *The* ergodic flow *between nodes x and y relative to a probability distribution π is defined to be the amount of probability mass flowing between x and y under π:*

$$Q(x,y) \stackrel{\text{def}}{=} \pi(x)\pi(x,y) \equiv \pi(y)P_{\mathcal{X}}(y,x) \ .$$

Thus, we can always write the transition probabilities of a reversible Markov chain $\mathcal{X}$ using ergodic flows:

$$P_{\mathcal{X}}(x,y) = \frac{\pi(y)P_{\mathcal{X}}(y,x)}{\pi(y)} = \frac{\pi(y)P_{\mathcal{X}}(y,x)}{\sum_{x\in\Omega}\pi(y)P_{\mathcal{X}}(y,x)} = \frac{Q(x,y)}{\sum_{x\in\Omega}Q(x,y)} \ .$$

We can also write the stationary distribution of $\mathcal{X}$ in terms of the transition probabilities, which is implied directly by the reversibility:

$$\frac{\pi(x)}{\pi(y)} = \frac{P_{\mathcal{X}}(y,x)}{P_{\mathcal{X}}(x,y)} \ .$$

We now turn to the proof of the theorem, in the special case of reversible Markov chains:

Proof: [Proof of Theorem 1] Consider a diagonal matrix $D = \mathrm{diag}\left\{\sqrt{\pi(1)}, \sqrt{\pi(2)}, \ldots, \sqrt{\pi(|\Omega|)}\right\}$. Because of the reversibility, $DP_{\mathcal{X}}D^{-1}$ is now symmetric and all of its eigenvalues are real. Let them be $\lambda_0, \lambda_1, \ldots, \lambda_{|\Omega|-1}$.

We invoke the Perron-Frobenius theorem, which states that if a Markov chain is aperiodic and irreducible, then $\lambda_0 = 1$ with multiplicity 1, and its corresponding eigenvector e_0 is coordinate-wise positive; moreover, all other eigenvectors satisfy $|\lambda_i| < 1$.

Now we can write any starting distribution $p_x^{(0)}$ using the basis of eigenvectors (guaranteed by the Spectral Theorem): $p_x^{(0)} = \sum_i \alpha_i e_i$. Then, we get:

$$p_x^{(1)} = p_x^{(0)}P_{\mathcal{X}} = \sum_i \alpha_i e_i P_{\mathcal{X}} = \sum_i \alpha_i D^{-1}(e_i D P_{\mathcal{X}} D^{-1})D = \sum_i \alpha_i D^{-1}\lambda_i e_i D = \sum_i \alpha_i \lambda_i e_i \ .$$

We have used the fact that for any vector, left multiplying it by D^{-1} and right multiplying it by D does not change it. In general, we can write $p_x^{(t)} = \sum_i \alpha_i \lambda_i^t e_i$. Using the fact that $|\lambda_i| < 1$ for $i > 0$, we conclude that $p_x^{(t)} \to \alpha_0 e_0$ as $t \to +\infty$. □

Note that:

- When $\mathcal{X}$ is reducible, then there may exist *multiple* stationary distributions. (Different components of the graph may have different stationary distributions, and there is no probability flow between them.)
- When $\mathcal{X}$ is irreducible (but not necessarily aperiodic), then the stationary distribution π still exists and is unique, but the Markov chain *does not necessarily converge to π from every starting vertex.* For example, consider the two-state Markov chain with transition matrix $P_{\mathcal{X}} = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$: the stationary distribution is the uniform one, yet it cannot be reached from any vertex; indeed, if we start at vertex 1, we go to vertex 2 after a single step, and then back and forth between 1 and 2 forever. (In this example, $\lambda_0 = 1$ and $\lambda_1 = -1$, so there is *another* eigenvalue of magnitude 1.)

On the other hand, for *any* irreducible Markov chain $\mathcal{X}$ (that is not necessarily aperiodic), for any $\alpha \in (0,1)$, the Markov chain $\mathcal{X}'$ whose transition matrix is $P'_{\mathcal{X}} \stackrel{\text{def}}{=} \alpha P_{\mathcal{X}} + (1-\alpha)I$ is irreducible and *aperiodic*; moreover, $\mathcal{X}'$ has the same stationary distribution as $\mathcal{X}$. The operation of going from $\mathcal{X}$ to $\mathcal{X}'$ corresponds to introducing a self-loop at all vertices of $G(\mathcal{X})$, each with weight $1-\alpha$. We call $\mathcal{X}'$ the *lazy version* of $\mathcal{X}$.

6 Examples on Card Shuffling

We now argue that the three card shuffling methods that we mentioned in the previous lecture *all converge to the uniform distribution.* We are going to analyze the irreducibility, aperiodicity, and reversibility (with respect to the uniform distribution) of each Markov chain, and then invoke Theorem 1 for establishing the result.

Recall that the set Ω now consists of $n!$ vertices, each corresponding to a permutation of the n cards in the deck.

6.1 Random Transpositions

In the method of random transposition, in each iteration we pick uniformly at random two cards i and j and switch them. The Markov chain is irreducible because we can reach every permutation from any starting permutation; it is aperiodic because every state has a self-loop to itself (by choosing $i = j$); it is reversible with respect to the uniform distribution. Hence, by Theorem 1, random-transposition shuffling converges to the uniform distribution over all card permutations.

6.2 Top in at Random

In the method of top in at random, in each iteration we take the top card and insert it at one of the n positions uniformly at random. The Markov chain is irreducible because we can reach every permutation from any starting permutation; it is aperiodic because we can put the top card back to its top position, forming a self-loop. We now argue that it is also reversible with respect to the uniform distribution.

At every state x, the Markov chain goes to n new states each with probability $1/n$. Moreover, for every state y, it comes back from exactly n states each with probability $1/n$ also. (This is because with the same probability, we can choose a card in the middle and put it back to the top.) Therefore the transition matrix satisfies that every row and every column sums up to 1. We call such matrix *doubly stochastic*, and it is clear that it is reversible with respect to uniform:

$$\sum_x \frac{1}{n} P_{\mathcal{X}}(x, y) = \frac{1}{n} = \sum_y \frac{1}{n} P_{\mathcal{X}}(y, x) \ .$$

Hence, again by invoking Theorem 1, top-in-at-random shuffling converges to the uniform distribution over all card permutations.

Let us formulate the fact that we have used as a lemma:

Lemma 6. *If a Markov chain $\mathcal{X}$ has a transition matrix $P_{\mathcal{X}}$ that is doubly stochastic, then it is reversible with respect to the uniform distribution.*

6.3 Riffle Shuffle

In the method of riffle shuffle, in each iteration we: (1) split the deck into two parts according to the binomial distribution $\mathrm{Bin}(n, 1/2)$; and (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., $\frac{|R|}{|L|+|R|}$).

The Markov chain is irreducible because using riffle shuffle we can "simulate" top-in-at-random with positive probability; and we already established that top-in-at-random is irreducible. The aperiodicity of this chain still comes from self-loops, as we can always take the left deck to be the empty set.

We now argue that the transition matrix for this Markov chain is also doubly stochastic, and again this suffices to prove that the chain is reversible with respect to the uniform distribution (and thus it converges to the uniform distribution). Indeed, if we write down the transition matrix, every row sums up to 1 by definition. Because the shuffle is symmetric, no state is given priority over any others, therefore, the summation of probabilities in each column must be the same value, but this value cannot be any number other than 1. As a consequence, the transition matrix is doubly stochastic.

Exericse (2pt): Prove that the riffle shuffle induces a random interleaving of the two decks. That is, given two sequences $T = (t_1, \ldots, t_n)$ and $S = (s_1, \ldots, s_m)$, prove that the random process used to "mix" them in the riffle shuffle induces a uniform distribution over the space of all possible interleavings between S and T.