CS294 Markov Chain Monte Carlo: Foundations & Applications **Fall 2009**

Lecture 2: September 8

Lecturer: Prof. Alistair Sinclair *Scribes: Anand Bhaskar and Anindya De*

2.1 Markov Chains

We begin by reviewing the basic goal in the Markov Chain Monte Carlo paradigm. Assume a finite state space Ω and a weight function $w : \Omega \to \Re^+$. Our goal is to design a sampling process which samples every element $x \in \Omega$ with the probability $\frac{w(x)}{Z}$ where $Z = \sum_{x \in \Omega} w(x)$ is the normalization factor. Often times, we don't know the normalization factor Z apriori, and in some problems, the real goal is to estimate Z.

With the aforementioned motivation, we now define a Markov chain.

Definition 2.1 *A Markov chain on Ω is a stochastic process $\{X_0, X_1, \ldots, X_t, \ldots\}$ with each $X_i \in \Omega$ such that*

$$\Pr(X_{t+1} = y \mid X_t = x, X_{t-1} = x_{t-1}, \ldots, X_0 = x_0) = \Pr(X_{t+1} = y \mid X_t = x) =: P(x, y)$$

Clearly, the Markov chain in the above matrix can be described by a $|\Omega| \times |\Omega|$ matrix P whose $(x, y)^{th}$ entry is $P(x, y)$. Hence, we sometimes blur the distinction between the Markov chain and its transtion matrix P. We now observe the following two important properties of the matrix P:

- P is non-negative, i.e., $\forall\ x, y \in \Omega,\ \ P(x, y) \geq 0$;
- $\forall x \in \Omega,\ \ \sum_{y \in \Omega} P(x, y) = 1$.

A matrix with the above two properties is called a *stochastic* matrix. We now describe some more notation which will be helpful in the course of our discussion. Let $p_x^{(t)} \in \Re^{1 \times |\Omega|}$ be the row vector corresponding to the distribution of X_t when the Markov chain starts at x, i.e. $X_0 = x$. With this notation, the evolution of the Markov chain can be defined in terms of matrix-vector equations. In particular:

- $p_x^{(t+1)} = p_x^{(t)} P$;
- $p_x^{(t)} = p_x^{(0)} P^t$.

[Note that while we assumed that the starting distribution $p_x^{(0)}$ is a point distribution—i.e., the starting vertex is x with probability 1—the same equations hold even if we start with a general distribution. However, it is usually sufficient to consider point distributions since any distribution is a convex combination of them. Moreover, for quantities such as the mixing time it is clear that the worst case is to start with a point distribution.]

2.1.1 Graphical representation

A Markov chain can also be represented graphically. Consider a Markov chain with state space Ω and transition matrix P. Then a corresponding graphical representation is the weighted graph $G = (V, E)$, where $V = \Omega$ and $E = \{(x, y) \in \Omega \times \Omega \mid P(x, y) > 0\}$. Also, edge (x, y) has weight $P(x, y) > 0$. Note that self-loops are allowed since we can have $P(x, x) > 0$.

Note that an edge is present between x and y if and only if the transition probability between x and y is non-zero. Much of the theory of Markov chains is not critically dependent upon the exact entries $P(x, y)$ but rather on whether a particular entry is zero or not. In terms of the graph, the critical thing is thus the structure of the graph G rather than the values of its edge weights.

Many natural Markov chains have the property that $P(x, y) > 0$ if and only if $P(y, x) > 0$. In this case the graph G is essentially undirected (except for the values of the edge weights). A very important special case is when the Markov chain is *reversible*:

Definition 2.2 *Let $\pi > 0$ be a probability distribution over Ω. A Markov chain P is said to be* reversible *with respect to π if $\forall\ x, y \in \Omega$, $\pi(x)P(x, y) = \pi(y)P(y, x)$.*

Note that any symmetric matrix P is trivially reversible (w.r.t. the uniform distribution π).

A reversible Markov chain can be completely represented by an undirected graph with weight $Q(x, y) := \pi(x)P(x, y) = \pi(y)P(y, x)$ on edge $\{x, y\}$ (without specifying P or π explicitly; and any fixed multiple of Q will do as well). To see this, note that the transition probability $P(x, y)$ can be computed from $P(x, y) = \frac{Q(x,y)}{\sum_z Q(x,z)}$. In fact, as we shall see below, for a reversible Markov chain π must in fact be its stationary distribution, and this can be computed from the $Q(x, y)$ also (using the fact that $\frac{\pi(x)}{\pi(y)} = \frac{P(y,x)}{P(x,y)}$). This is one reason why reversible chains are particularly nice to deal with. For any Markov chain, the quantity $\pi(x)P(x, y)$ is called the *ergodic flow* from x to y, i.e., the amount of probability mass flowing from x to y in stationarity. Reversibility says that the ergodic flows from x to y and from y to x are equal; for this reason, the condition in Definition 2.2 is known as the "detailed balance" condition. Of course, by conservation of mass we always have $\pi(S)P(S, \overline{S}) = \pi(\overline{S})P(\overline{S}, S)$ for any subset of states $S \subseteq \Omega$ (where $\overline{S} = \Omega \setminus S$). Detailed balance says that this also holds *locally*, for every pair of states.

2.1.2 Mixing of Markov chains

We now discuss some definitions and theorems which are important in the context of mixing of Markov chains.

Definition 2.3 *A probability distribution π over Ω is a* stationary distribution *for P if $\pi = \pi P$.*

Definition 2.4 *A Markov chain P is* irreducible *if for all x, y, there exists some t such that $P^t(x, y) > 0$. Equivalently, the graph corresponding to P (denoted by $G(P)$) is strongly connected. In case the graphical representation is an undirected graph, then it is equivalent to $G(P)$ being connected.*

Definition 2.5 *A Markov chain P is* aperiodic *if for all x, y we have* $\gcd\{t : P^t(x, y) > 0\} = 1$.

The following simple exercises are recommended to understand the content of these definitions:

Exercise 2.6 *For a Markov chain P, if $G(P)$ is undirected then aperiodicity is equivalent to $G(P)$ being non-bipartite.*

Exercise 2.7 *Define the period of x as $\gcd\{t : P^t(x,x) > 0\}$. Prove that for an irreducible Markov chain, the period of every $x \in \Omega$ is the same. [Hence, if $G(P)$ is undirected, the period is either 1 or 2.]*

Exercise 2.8 *Suppose P is irreducible. Then P is aperiodic if and only if there exists t such that $P^t(x,y) > 0$ for all $x, y \in \Omega$.*

Exercise 2.9 *Suppose P is irreducible and contains at least one self-loop (i.e., $P(x,x) > 0$ for some x). Then P is aperiodic.*

We now state a theorem which gives a necessary and sufficient condition for convergence of a Markov chain to its stationary distribution regardless of the initial state.

Theorem 2.10 (Fundamental Theorem of Markov Chains) *If a Markov chain P is irreducible and aperiodic then it has a unique stationary distribution π. This is the unique (normalized such that the entries sum to 1) left eigenvector of P with eigenvalue 1. Moreover, $P^t(x,y) \to \pi(y)$ as $t \to \infty$ for all $x, y \in \Omega$.*

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

We shall give an elementary probabilistic proof of the above theorem in the next lecture. However, today we sketch an algebraic proof in the special case where the Markov chain is reversible.

In preparation for this, let us verify that, if P is reversible w.r.t. π, then π is a stationary distribution.

Claim 2.11 *If a Markov chain P is reversible w.r.t. π, then π is a stationary distribution for P.*

Proof:

$$(\pi P)(y) = \sum_x \pi(x)P(x,y) = \sum_x \pi(y)P(y,x) = \pi(y)$$

■

Exercise 2.12 *Let P be reversible w.r.t. π. Show that P is similar to a symmetric stochastic matrix, under transformation by the diagonal matrix $\mathrm{diag}(\sqrt{\pi(x)})$.*

Our proof of Theorem 2.10 for reversible chains will make use of the following classical theorem:

Theorem 2.13 (Perron-Frobenius) *Any irreducible, aperiodic stochastic matrix P has an eigenvalue $\lambda_0 = 1$ with unique associated left eigenvector $e_0 > 0$. Moreover, all other eigenvalues λ_i of P satisfy $|\lambda_i| < 1$.*

Proof:(of Theorem 2.10, sketch for reversible case) The fact that P is reversible w.r.t. π means that it is similar to a symmetric matrix (see exercise above). Hence the eigenvalues of P are real, and we can choose a basis of $\Re^{|\Omega|}$ among its eigenvectors. Let the eigenvalues be $\lambda_0 \geq \lambda_1 \geq \ldots \geq \lambda_{|\Omega|-1}$, and the corresponding eigenvectors be $e_0, e_1, \ldots, e_{|\Omega|-1}$. By Theorem 2.13, $\lambda_0 = 1$, and $|\lambda_i| < 1$ for all $i > 0$; also, $e_0 = \pi$ is the unique stationary distribution. To see the convergence property, note that we can write the initial distribution $p^{(0)}$ as a linear combination of eigenvectors, i.e., $p^{(0)} = \sum_i \alpha_i e_i$. But then $p^{(t)} = p^{(0)}P^t = \sum_i \alpha_i \lambda_i^t e_i$. Since $|\lambda_i| < 1$ for all $i > 0$, this implies that $p^{(t)} \to \alpha_0 e_0$, which is the same as e_0 up to a scalar factor. However, by conservation of mass we must have $\alpha_0 = 1$, so the distribution converges to π. ■

If P is not reversible then the Perron-Frobenius theorem still applies but the proof of Theorem 2.10 is a bit more complicated; see, e.g., [Se06] for details.

If P is irreducible (but not necessarily aperiodic), then π still exists and is unique, but the Markov chain does not necessarily converge to π from every starting state. For example, consider the two-state Markov chain with $P = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$. This has the unique stationary distribution $\pi = (1/2, 1/2)$, but does not converge from either of the two initial states. Notice that in this example $\lambda_0 = 1$ and $\lambda_1 = -1$, so there is another eigenvalue of magnitude 1, contradicting the Perron-Frobenius theorem. However, the Perron-Frobenius theorem does generalize to the periodic setting, with the weaker conclusion that the remaining eigenvalues satisfy $|\lambda_i| \leq 1$.

In this course we will not spend much time worrying about periodicity, because of the following simple observation (proof: **exercise!**).

Observation 2.14 *Let P be an irreducible (but not necessarily aperiodic) stochastic matrix. For any $0 < \alpha < 1$, the matrix $P' = \alpha P + (1-\alpha)I$ is stochastic, irreducible and aperiodic, and has the same stationary distribution as P.*

This operation corresponds to introducing a self-loop at all vertices of $G(P)$ with probability $1 - \alpha$. The value of α is usually set to $1/2$. P' is usually called a "lazy" version of P.

In the design of MCMC algorithms, we usually do not worry about periodicity since, instead of simulating the Markov chain P, the algorithm can simulate the lazy version P'. This just has the effect of slowing down the dynamics by a factor of 2.

2.2 Examples of Markov Chains

2.2.1 Random Walks on Undirected Graphs

Definition 2.15 *Random walk on an undirected graph $G(V, E)$ is given by the transition matrix*

$$P(x,y) = \begin{cases} 1/deg(x) & \text{if } (x,y) \in E; \\ 0 & \text{otherwise.} \end{cases}$$

Proposition 2.16 *For random walk P on an undirected graph, we have:*

- *P is irreducible iff G is connected;*
- *P is aperiodic iff G is non-bipartite;*
- *P is reversible with respect to $\pi(x) = deg(x)/(2|E|)$.*

Proof: The first part is straightforward. The second part is a previous exercise. For the third part, we check directly that $\pi(x)P(x,y) = \frac{1}{2|E|} = \pi(y)P(y,x)$. ■

2.2.2 Ehrenfest Urn

In the Ehrenfest Urn, we have 2 urns and n unlabelled balls, where there are j balls in the first urn and $n-j$ balls in the other. Define the state space $\Omega = \{0, 1, ..., n\}$, denoting the number of balls in the first urn. At each step of the Markov chain, we pick a ball u.a.r. and move it to the other urn.

The non-negative entries of the transition matrix are given by

$$\begin{aligned} P(j\ , j+1) &= (n-j)/n, \\ P(j\ , j-1) &= j/n. \end{aligned}$$

The Markov chain is clearly irreducible, and we can check that $\pi(j) = \binom{n}{j}/2^n$ is a stationary distribution (**exercise**). However, P is not aperiodic since the time to return to any given state is even.

2.2.3 Card Shuffling

In card shuffling, we have a deck of n cards, and we consider the space Ω of all permutations of the cards. Thus $|\Omega| = n!$. The aim is to sample from the distribution given by the uniform weight $w(x) = 1\ \forall x \in \Omega$, i.e., to sample a permutation of the cards u.a.r. Thus, in the Markov chain setting, we want the stationary distribution π be uniform.

We look at three different shuffling schemes:

Random Transpositions

Pick two cards i and j uniformly at random with replacement, and switch cards i and j.

This is a pretty slow way of shuffling. The chain is irreducible (any permutation can be expressed as a product of transpositions), and also aperiodic (since we may choose $i = j$, so the chain has self-loops). Since the random transpositions are invertible, we have $P(x, y) = P(y, x)$ for every pair of permutations x, y, so P is symmetric. This implies immediately that its stationary distribution is uniform (since it is reversible w.r.t. the uniform 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 shuffle is again irreducible (**exercise**) and aperiodic (due to self-loops). However, note that it is not symmetric (or even reversible): If we insert the top card into (say) the middle of the deck, we cannot bring the card back to the top in one step.

However, notice that every permutation y can be obtained, in one step, from exactly n different permutations (corresponding to the n possible choices for the identity of the previous top card). Since every non-zero transition probability is $\frac{1}{n}$, this implies that $\sum_x P(x, y) = 1$; thus the matrix P is *doubly stochastic* (i.e., its column sums, as well as its row sums, are 1). It is easy to show that the uniform distribution is stationary for doubly stochastic matrices; in fact (**exercise**), π is uniform *if and only if* P is doubly stochastic.

Riffle Shuffle (Gilbert-Shannon-Reeds [Gi55,Re81])

- *Split the deck into two parts according to the binomial distribution $Bin(n, 1/2)$.*
- *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. $\frac{|R|}{|L|+|R|}$).*

Note that the second step of the shuffle is equivalent to choosing an interleaving of the two parts uniformly at random (**exercise**).

The chain is irreducible (**exercise**), aperiodic (due to self-loops), and doubly stochastic, and hence its stationary distribution is uniform.

Note: This shuffle is quite different from the "perfect shuffle" performed by professional magicians, who split the deck exactly in half and then perfectly interleave the two halves. The perfect shuffle has no randomness, so the configuration of the deck after any given number of shuffles is known exactly.

References

[Se06] E. Seneta, *Non-negative matrices and Markov chains*, 2nd ed. (revised printing), Springer-Verlag, New York, 2006.

[Gi55] E. Gilbert, "Theory of shuffling," Technical Memorandum, Bell Laboratories, 1955.

[Re81] J. Reeds, Unpublished manuscript, 1981.