CS271 Randomness & Computation **Fall 2011**

Lecture 24: November 22

Lecturer: Alistair Sinclair *Based on scribe notes by: Daniel Chen and Kaushik Ravindran*

In this lecture we discuss the Markov chain Monte Carlo method and some simple applications. We then define mixing time and study bounds on mixing time for some Markov chains.

24.1 Random Walks on Undirected Graphs

As in the last lecture, let $G = (V, E)$ be a connected, undirected graph. We perform a random walk on G starting from some vertex u. Let $P_u^{(t)}(v)$ denote the probability that the walk is at vertex v at time t. Then, provided G is not bipartite,

$$P_u^{(t)}(v) \stackrel{t\to\infty}{\longrightarrow} \frac{d(v)}{2|E|}.$$

In other words, the probability of being at any given vertex tends to a well-defined limit, independent of the initial vertex u. This limiting probability distribution is called the "stationary distribution" of the chain. We study this property in the more general setting of Markov chains.

24.2 Markov Chains

A *Markov chain* is a sequence of random variables $\left(X_t\right)_{t=0}^{\infty}$ with the Markov property:

$$\begin{aligned}\Pr[X_t = y | X_{t-1} = x, X_{t-2}, \ldots, X_0] &= \Pr[X_t = y | X_{t-1} = x] \\ &\stackrel{def}{=} P(x, y).\end{aligned}$$

Thus the probability that the process is in a particular "state" y at time t depends only on its state x at time $t-1$; this probability is denoted $P(x, y)$, and is independent of t. The values $P(x, y)$ are called the *transition probabilities* of the chain. We can also think of a Markov chain as a process on a directed graph $G = (V, E)$, where V is the range of values taken by the X_t (assumed discrete), and for each x, y for which $P(x, y) > 0$ there is a directed edge of weight $P(x, y)$ from x to y.

For random walk on a graph as in the previous section, we have $P(x, y) = \frac{1}{d(x)}$ whenever y is a neighbor of x, and $P(x, y) = 0$ otherwise.

A finite Markov chain is completely specified by the matrix P of transition probabilities. Note that P is a *stochastic matrix*, i.e., all its entries are non-negative and all its row sums are 1 ($\sum_y P(x, y) = 1$ for all x). We let $p_x^{(t)}$ denote the distribution of the state X_t at time t, given that the initial state was $X_0 = x$. Note that we have

$$p_x^{(t)} = p_x^{(0)} P^t,$$

where $p_x^{(0)} = (0, \ldots, 1, 0, \ldots)$ is the unit vector with probability mass concentrated at x. Hence the time evolution of the Markov chain can be viewed as iterative multiplication of the probability vector by the transition matrix P.

24.2.1 Fundamental theorem of Markov Chains

Definition 24.1 (irreducible) *P is "irreducible" if $\forall x, y,\ \exists t$ s.t. $p_x^{(t)}(y) > 0$.*

I.e., P is irreducible if every state can be reached from every other in a finite number of steps.

Definition 24.2 (aperiodic) *P is "aperiodic" if $\forall x, y,\ \gcd\{t : p_x^{(t)}(y) > 0\} = 1$.*

I.e., P is aperiodic if the transition probabilities do not exhibit "periodic" behavior. The simplest example of periodic behavior is random walk on a *bipartite* graph: clearly the state of the walk will have the same parity as its initial state at even time steps, and the opposite parity at odd time steps. (In fact, random walk on a graph is aperiodic *if and only if* the graph is not bipartite.) Obviously a periodic Markov chain cannot converge to a stationary distribution because the "parity" of its state at time t will depend on its initial state. We now state the fundamental property of Markov chains: any Markov chain that is irreducible and aperiodic always converges to a unique stationary distribution, regardless of the initial state:

Theorem 24.3 (Fundamental theorem) *If P is irreducible and aperiodic, then it converges to a unique stationary distribution π, i.e.,*

$$p_x^{(t)}(y) \stackrel{t\to\infty}{\longrightarrow} \pi(y),\ \forall x, y.$$

Moreover, π is the unique left eigenvector of P with eigenvalue 1 (i.e., $\pi P = \pi$).

The proof of this theorem is not presented here. Please refer to any standard probability text for a proof.

The significance of this theorem is that it provides an algorithmic approach to sampling from probability distributions over large sets, generally known as the "Markov chain Monte Carlo" method. Namely, given a set Ω and a target probability distribution π over Ω, the basic paradigm is to (i) design an (irreducible, aperiodic) Markov chain on Ω with stationary distribution π; then (ii) simulate the Markov chain, starting from any initial state, for a large number of steps and output the final state. Asymptotically, the distribution of this final state will be exactly π; in practice, we will need to analyze in advance how many steps are necessary for the distribution to be "close to" π—this is known as the *mixing time* of the chain.

24.2.2 Determining the stationary distribution

The Fundamental Theorem says that the stationary distribution is the unique left eigenvector of P with eigenvalue 1. Here we list some useful observations that help in determining the value of π in many important cases. (We assume throughout that P is irreducible and aperiodic.)

Observation 1: If P is *symmetric* (i.e., $P(x, y) = P(y, x)\ \forall x, y$), then π is uniform.

To see this, by the Fundamental Theorem we just need to check that $\pi(x) = 1/N\ \forall x$ (where N is the number of states) satisfies $\pi P = \pi$:

$$(\pi P)(x) = \sum_y \pi(y)P(y, x) = \frac{1}{N}\sum_y P(x, y) = \frac{1}{N} = \pi(x).$$

Observation 2: If P is *doubly stochastic* (i.e., $\sum_x P(x, y) = 1\ \forall y$), then π is uniform. (This generalizes Observation 1.) Exercise: Verify this.

Observation 3: If P is *reversible* with respect to some distribution π, (i.e., $\pi(x)P(x,y) = \pi(y)P(y,x)\ \forall x, y$), then the stationary distribution is π.

To check that π is stationary:

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

We can use Observation 3 to check that the stationary distribution for random walk on an undirected graph is indeed $\pi(x) = \frac{d(x)}{2|E|}$, as claimed earlier. To do so, we just need to show that the random walk is reversible w.r.t. π: for neighbors x, y we have

$$\pi(x)P(x,y) = \frac{d(x)}{2|E|} \times \frac{1}{d(x)} = \frac{1}{2|E|}.$$

Since this is independent of x, y, we must have $\pi(x)P(x,y) = \pi(y)P(y,x)$. And for non-neighbors x, y, we have $\pi(x)P(x,y) = 0 = \pi(y)P(y,x)$. Hence reversibility holds.

24.3 Examples

24.3.1 Card Shuffling

Given a set of n cards, the sample space Ω is the set of possible permutations (so $|\Omega| = n!$). The goal of card shuffling is to pick a permutation uniformly at random, so the distribution π is uniform. The following three natural Markov chains achieve a uniform stationary distribution.

(i) Random Transpositions:
Choose any two cards at random in the deck of n cards and transpose their positions.

If two states (permutations of the n cards) differ by a single transposition, the transition probabilities in both directions are identical (both are $\frac{1}{n^2}$). Hence the stochastic matrix P is symmetric, so by Observation 1 π is uniform.

(ii) Top-in-at-Random:
Remove the top card and insert it at random into any of the n positions in the deck.

The transition probability between adjacent states is $1/n$; also, each state can be reached by a transition in exactly n ways. Hence the sum of transition probabilities into any state is 1. Therefore P is doubly stochastic so, by Observation 2, π is uniform.

(iii) Riffle Shuffle:
The Gilbert-Shannon-Reeds process is a fairly good model for the way real people shuffle cards:

- Split the deck into 2 parts (R and L) according to a binomial distribution (i.e., the number of cards in the first part, R, has distribution $\text{Bin}(n, 1/2)$).
- Drop cards one by one from R and L, with the next card coming from R (from L) with probability proportional to the number of cards currently in R (in L). Conditioned on the choice of R and L, this

is equivalent (Exercise!) to picking, uniformly at random, any interleaving of the cards in R and L (i.e. an ordering of the full deck that preserves the order of both R and L).

Here it is not hard to check (Exercise!) that P is again doubly stochastic, so by Observation 2 π is uniform.

24.3.2 Graph Coloring

Given a graph $G = (V, E)$, the goal is to sample (proper) colorings of G with k colors (we shall see in a later lecture some motivations for doing this). Note that this is presumably harder than the standard decision problem of simply determining whether G has a k-coloring. The following random walk achieves a uniform stationary distribution over k-colorings:

- Start with an arbitrary k-coloring.
- Pick a vertex v and a color c uniformly at random. Recolor v with c if this is legal, else do nothing.

This Markov chain on k-colorings is irreducible provided $k \geq \Delta + 2$, where Δ is the maximum degree of G. (Exercise: Check that, with this assumption, it is possible to pass from any k-coloring to any other by recoloring one vertex at a time, always remaining within the set of legal colorings.) It is aperiodic because there is a probability at least $1/k$ at each step that the state remains unchanged (if the color chosen is the current color of v). Note also that, by a simple greedy argument, G has at least one $(\Delta + 1)$-coloring, and moreover such a coloring can be found in polynomial time; thus finding an initial state for the Markov chain is not a problem.

The vertex and color are picked uniformly at random. Hence, the transition probability between two states is the same in either direction. Hence P is symmetric, and the stationary distribution π is uniform as desired.

24.3.3 Metropolis Process

The examples above are Markov chains converging to a uniform stationary distribution. However, as previously indicated, we sometimes may want to sample using a specified non-uniform distribution. The *Metropolis Process* is a very general method for achieving this; it is named after one of its inventors [MR+53].

Formally, given a large set Ω and a weight function $w : \Omega \to \mathbf{R}^+$, we want to design a Markov Chain with stationary distribution $\pi(x) = w(x)/Z$, where $Z = \sum_{x \in \Omega} w(x)$ is a normalizing factor. Note that we do *not* assume that Z is known; indeed, in many applications Z is precisely what we will be aiming to compute!

The ingredients of the Metropolis Process are:

- A connected "neighborhood structure" (undirected graph) on Ω.
- A "proposal distribution" κ such that $\kappa(x, y) > 0$ iff x, y are neighbors, and $\kappa(x, y) = \kappa(y, x)$. (I.e., κ specifies a rule for choosing a random neighbor of any state.)

Then, we construct the following Markov Chain:

1. In state x, pick a neighbor y with probability $\kappa(x, y)$.
2. With probability $\min\{1, w(y)/w(x)\}$, go to y; else stay at x.

Claim 24.4 *The Markov Chain constructed by the Metropolis Process is reversible w.r.t.* $\pi(x) = w(x)/Z$.

Proof: If x, y are not neighbors, then $\pi(x)P(x,y) = \pi(y)P(y,x) = 0$. If x, y are neighbors, then assume w.l.o.g. that $w(x) \geq w(y)$. Then we have

$$\pi(x)P(x,y) = \frac{w(x)}{Z} \times \kappa(x,y)\frac{w(y)}{w(x)} = \frac{1}{Z}w(y)\kappa(x,y),$$

and

$$\pi(y)P(y,x) = \frac{w(y)}{Z} \times \kappa(y,x) = \frac{1}{Z}w(y)\kappa(y,x).$$

Since $\kappa(x,y) = \kappa(y,x)$, these two expressions are equal and the Markov chain is reversible. ■

Note: The assumption $\kappa(x,y) = \kappa(y,x)$ is not really necessary. We can allow an arbitrary proposal distribution κ (such that $\kappa(x,y) > 0$ iff x, y are neighbors) by modifying the acceptance probability $\min\{1, w(y)/w(x)\}$ to $\min\{1, (w(y)\kappa(y,x))/(w(x)\kappa(x,y))\}$.

24.4 The Mixing Time

Although the Fundamental Theorem tells us that an irreducible, aperiodic Markov chain converges to a stationary distribution, it does not tell us how fast it converges. For example, how many shuffles are needed to obtain an (almost) uniform permutation of a deck of cards? This question is particularly important for algorithmic applications, where we need to ensure that a sample can be obtained in a fairly small number of simulation steps (even when the set Ω is of exponential size, as it is, e.g., for graph colorings above).

In order to study the rate of convergence of Markov chains, we first introduce a few definitions:

Definition 24.5 (Variation Distance) *The "variation distance" between probability distributions* μ, ξ *over* Ω *is defined as:*

$$||\mu - \xi|| = \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \xi(x)| = \max_{A\subseteq\Omega} |\mu(A) - \xi(A)|.$$

(Exercise: Check the above equality.) Note that this is just the same as the standard L_1 norm, with the extra $\frac{1}{2}$ factor introduced to keep the variation distance between 0 and 1.

Definition 24.6 *For an (irreducible, aperiodic) Markov chain, define the distance at time* t *by*

$$\Delta(t) = \max_{x\in\Omega} ||\pi - p_x^{(t)}||$$

The Fundamental Theorem says that $\Delta(t) \to 0$ as $t \to \infty$. The mixing time measures this rate of convergence:

Definition 24.7 (Mixing Time) *The "mixing time"* τ_{mix} *is defined as:*

$$\tau_{mix} = \min\left\{t : \Delta(t) \leq 1/2e\right\}.$$

(The constant $1/2e$ is chosen for algebraic convenience only; any other constant less than $1/2$ would do.) The reason we define the mixing time this way is the following:

Fact: $\Delta\left(\tau_{mix}\lceil \ln \epsilon^{-1}\rceil\right) \leq \epsilon$. (This follows from the submultiplicativity of Δ, i.e., $\Delta(kt) \leq (2\Delta(t))^k$.)

Thus once the mixing time has been reached, the variation distance at subsequent times decays exponentially fast: in order to achieve a variation distance of any desired $\epsilon > 0$, we need to run the chain for at most $\tau_{mix}\lceil \ln \epsilon^{-1}\rceil$ steps.

24.5 Strong Uniform Times

We now look at a couple of simple analyses of mixing times. These use the concept of a "strong stationary time."

Definition 24.8 (Strong Stationary Time) *A "strong stationary time" is a stopping time T s.t.* $\Pr[X_t = y|T = t) = \pi(y)$.

It should be intuitively clear, and is not hard to prove (Exercise!), that $\Delta(t) \leq \Pr[T > t]$ for any strong stationary time T. Thus any strong stationary time will give us an upper bound on the mixing time τ_{mix}.

We use strong stationary times to analyze two of the card shuffling examples we saw earlier.

Theorem 24.9 *The mixing time of Top-in-at-Random is $O(n \log n)$.*

Proof: Consider the card B initially at the bottom of the deck. After some number of steps, B will have moved up to some higher position (without yet having reached the top of the deck). We claim that, conditioned on the position of B and the identities of the cards below B, the permutation of the cards below B is uniformly random. This is true because whenever a card is inserted below B, this is done u.a.r. Therefore, if we set T to be the time until B reaches the top of the deck and is reinserted for the first time, then T is a strong stationary time. Thus we just need to analyze the distribution of T.

For $1 \leq i \leq n-1$, Let T_i be the time for B to move up from position i from the bottom to position $i+1$ from the bottom (where the bottom of the deck corresponds to $i = 1$, and the top to $i = n$). Then

$$T = T_1 + T_2 + T_3 + \ldots + T_{n-1} + 1.$$

Note that T_i is a geometric r.v. with expectation n/i (because card B moves iff the top card is inserted in one of the i positions below B, so T_i is the time for the first success in a sequence of Bernoulli trials with success probability i/n). Hence, by linearity of expectation,

$$E[T] = \sum_i (n/i) = O(n \log n).$$

But by Markov's inequality $\Pr[T > 2eE[T]] \leq 1/2e$, so $\tau_{mix} \leq O(n \log n)$. ■

Note: The above bound is actually tight up to constants. In fact, $\tau_{mix} = n \ln n + O(n)$.

Theorem 24.10 *The mixing time of the Riffle Shuffle is upper bounded by $2 \log_2 n + O(1)$.*

Proof: To analyze the riffle shuffle, it is easier to work with the *inverse* process, which has the same mixing time as the original shuffle. (This follows from the fact that the shuffle is a random walk on a group.) The inverse shuffle works in the following manner:

1. Mark each card with a 0 or 1, independently and u.a.r.
2. Pull out the cards marked 0 (keeping them in the same relative order) and put them on top of the cards marked 1 (also kept in the same relative order).

Suppose that, as we perform the inverse shuffle, we retain the 0-1 labels on the backs of the cards. Thus after t steps each card will be labeled with a t-digit binary number. It is not hard to see that all sets of

cards with distinct labels are in random relative order. (This just follows from the fact that all labels were assigned u.a.r.) In other words, the only structure remaining from the initial ordering is that cards with the same label are in their original relative order. Thus if we let T be the first time at which all cards have distinct labels, then T is a strong stationary time.

How many steps are needed before this event occurs? After t steps, the label of each card is a random t-bit number drawn independently and u.a.r. By the well-known "birthday problem," if n people choose "birthdays" randomly from a set of cn^2 dates, then the probability that some pair have the same birthday is asymptotically $1 - \exp\{-1/2c\} \approx 1/2c$. In our application, the number of birthdays is 2^t (the number of t-digit binary numbers), and we want the probability that some pair of cards have a common label to be small — specifically, less than $1/2e$ to get the mixing time. So we choose c s.t. $1 - \exp(-1/2c) \leq 1/2e$ and then t s.t. $2^t \geq cn^2$, i.e., $t \geq 2\log_2 n + \Theta(1)$. Thus $\tau_{mix} \leq 2\log_2 n + \Theta(1)$. ■

Note: The above bound is almost tight; the true value of τ_{mix} is $\frac{3}{2}\log_2 n$ plus lower order terms.

For more information on strong stationary times, and other stopping rules, see the papers by Aldous and Diaconis [AD86] and by Lovász and Winkler [LW95].

Finally, Bayer and Diaconis [BD92] have found a clever way to explicitly calculate $\Delta(t)$ for the riffle shuffle for any value of t and any number of cards n, in particular for $n = 52$. (Note that as the space of possible deck arrangements has $52! \approx 8.07 \times 10^{67}$ states, a naïve brute-force calculation is not feasible.) For $n = 52$ the numbers are:

t	1	2	3	4	5	6	7	8
$\Delta(t)$	1	1	1	1	0.92	0.61	0.33	0.17

Table 24.1: Exact variation distance for deck of 52 cards (to 2 decimal places).

Based on these results, they argue that for practical purposes (e.g., for card games in a casino) seven riffle shuffles are sufficient to prevent even the best card player from exploiting any structure remaining in the deck.

It turns out that for most natural Markov chains, there is no obvious choice of a strong stationary time. We must therefore turn to more sophisticated methods to bound the mixing time. In the next lecture, we will see one of these (known as "coupling") and use it to bound the mixing time of the Markov chain on graph colorings.

References

[AD86] D. Aldous and P. Diaconis "Shuffling cards and stopping times," *Amer. Math. Monthly 93*, 1986, no. 5, pp. 333-348.

[BD92] D. Bayer and P. Diaconis, "Trailing the dovetail shuffle to its lair," *Annals of Applied Probability* **2** (1992), pp. 294–313.

[LW95] L. Lovász and P. Winkler. "Mixing of random walks and other diffusions on a graph." In *Surveys in Combinatorics* (P. Rowlinson, ed.), London Mathematical Society Lecture Notes Series 218, Cambridge University Press (1995), pp. 119–154.

[MR+53] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller and E. Teller, "Equation of state calculations by fast computing machines," *Journal of Chemical Physics*, vol. 21, no. 6, pp. 1087–1092, 1953.