CS 8803 MCMC – Markov Chain Monte Carlo Algorithms Professor: *Dana Randall*
January 29, 2010 Scribe: *Abhinav Shantanam*

Markov Chain Fundamentals (Contd.)

Recap from last lecture:

- Any starting probability distribution π_0 can be represented as a linear combination of the eigenvectors including v_0 (stationary distibution) and hence its evolution π_t with t as shown below,

$$\pi_0 = v_0 + \Sigma_{i\neq 0} a_i v_i ; \pi_t = v_0 + \Sigma_{i\neq 0} a_i \lambda_i^t v_i$$

and since $|\lambda_{i\neq 0}| < 1$, the $v_{i\neq 0}$ components diminish as t grows larger, making $\pi_t \to v_0$.

- The multiplicity/coefficient of $v_0(a_0)$ in any such linear combination is always 1.

Facts on Markov-Matrices

Fact 1: Any irreducible and ergodic Markov chain with a symmetric transition matrix has uniform stationary distribution.
Proof. Using $P(x, y) = P(y, x)$ in the condition for detailed balance, the claim follows.

Note that the dominant eigenvector (corresponding to $\lambda_0 = 1$) in such a case becomes $[\frac{1}{N}, \frac{1}{N}, ..., \frac{1}{N}]$, where $N = |\Omega|$.

Fact 2: We can define an other transition matrix P' such that $P'_{ij} = \pi(i)^{1/2} \cdot P(i, j) \cdot \pi(j)^{-1/2}$, where π is the stationary distribution corresponding to P. More generally,

$$P' = D_\pi^{1/2} \cdot P \cdot D_\pi^{-1/2} ; D_\pi = \begin{bmatrix} \pi(1) & 0 & ... & 0 \\ 0 & \pi(2) & ... & 0 \\ ... & ... & ... & ... \\ 0 & 0 & ... & \pi(N) \end{bmatrix}$$

Note that:

- P' is not necessarily a stochastic matrix.

- P' is symmetric whenever P represents a reversible Markov chain.
Proof. From the condition for detailed balance we have

$$\begin{gathered} \pi(i) \cdot P(i, j) = \pi(j) \cdot P(j, i) \\ \pi(i)^{1/2} \cdot P(i, j) \cdot \pi(j)^{-1/2} = \pi(j)^{1/2} \cdot P(j, i) \cdot \pi(i)^{-1/2} \\ P'_{ij} = P'_{ji}. \end{gathered}$$

- P' has the same eigenvalues as P and hence the same spectral gap.
Proof.

$$P' = D_\pi^{1/2} \cdot P \cdot D_\pi^{-1/2} \Rightarrow D_\pi^{-1/2} \cdot P' \cdot D_\pi^{1/2} = P$$

Let λ be an eigenvalue of P and x the corresponding eigenvector. Then

$$P \cdot x = \lambda x \Rightarrow D_\pi^{-1/2} \cdot P' \cdot D_\pi^{1/2} \cdot x = \lambda x \Rightarrow P' \cdot D_\pi^{1/2} \cdot x = \lambda D_\pi^{1/2} \cdot x$$

shows that P' too has an eigenvalue λ with $D_\pi^{1/2} \cdot x$ as the corresponding eigenvector.

- When P' is symmetric, it has an orthogonal basis of eigenvectors and the columns of D_π above form just such a basis.

- Since $(P')^t = D_\pi^{1/2} \cdot P^t \cdot D_\pi^{-1/2}$, convergence time for the P'- Markov chain is similar to that for the (possibly asymmetric) P-Markov chain.

Mixing Time and Spectral Gap

Theorem: Let P be an ergodic, symmetric Markov chain with N states and spectral gap δ. Then its mixing time is bounded above by

$$\tau_\varepsilon < \frac{\ln(N\varepsilon^{-1})}{\delta}.$$

Proof. Let us first write the initial distribution π_0 as a linear combination of P's eigenvectors:

$$\pi_0 = \pi_{eq} + \pi_{neq}, \text{ where } \pi_{neq} = \textstyle\sum_{i\in[N-1]} a_k v_k$$

Then the distribution after t steps is given by

$$\pi_t = \pi_0 \cdot P^t = \pi_{eq} + \pi_{neq} \cdot P^t = \pi_{eq} + \textstyle\sum_{i\in[N-1]} a_k \lambda_k^t v_k$$

Also, the total variation distance is given by

$$||\pi_t - \pi_{eq}||_{TV} = \tfrac{1}{2}||P^t \cdot \pi_{neq}||_1, \text{ where the subscript 1 denotes the } L1 \text{ norm.}$$

For any N-dimensional vector v, the *Cauchy-Schwartz* inequality relates its $L1$ norm to its $L2$ norm by

$$||v||_2 \leq ||v||_1 \leq \sqrt{N}\ ||v||_2$$

Thus

$$||\pi_t - \pi_{eq}||_{TV} \leq \tfrac{1}{2}\sqrt{N}\ ||P^t \cdot \pi_{neq}||_2$$

Since P is symmetric, its eigenvectors v_k are orthogonal. Then Pythagoras' theorem and $|\lambda_{k\geq 1}| \leq 1-\delta$, gives

$$\begin{aligned} ||P^t \cdot \pi_{neq}||_2 = \sqrt{\textstyle\sum_{i\in[N-1]} |a_k|^2\ |\lambda_k|^{2t}\ ||v_k||_2^2} &\leq (1-\delta)^t \sqrt{\textstyle\sum_{i\in[N-1]} |a_k|^2\ ||v_k||_2^2} \\ &= (1-\delta)^t ||\pi_{neq}||_2 \\ &= (1-\delta)^t \sqrt{||\pi_0||_2^2 - ||\pi_{eq}||_2^2} \\ &\leq (1-\delta)^t, \text{ since } ||\pi_0||_2^2 \leq 1 \end{aligned}$$

which in turn gives

$$||\pi_t - \pi_{eq}||_{TV} \leq \tfrac{1}{2}\sqrt{N}\ (1-\delta)^t \leq \tfrac{1}{2}\sqrt{N}\ e^{-\delta t}.$$

Recall that

$$\tau_\varepsilon = Min_{\forall P_0}\{t : ||\pi_t - \pi||_{TV} < \varepsilon\}$$

so that setting $||\pi_t - \pi||_{TV} = \varepsilon$ gives us an upper bound on the mixing time,

$$\tau_\varepsilon < \tfrac{\ln(\sqrt{N}\varepsilon^{-1}/2)}{\delta}$$

which implies he weaker bound stated in the theorem. □

Mixing Time from ”First Principles”

We will learn some formal methods of bounding the mixing time of a Markov chain (canonical paths, coupling, etc.) in the following lecture. Let us now try to bound it from ”first principles” in the examples below.

Walking on the hypercube

A random walk on an n-dimensional hypercube can be used to generate random n-bit strings. As shown below, an n-dimensional hypercube has 2^n vertices, and every pair that has a *hamming distance* of 1 between them forms a pair of neighbors along an edge of the hypercube.

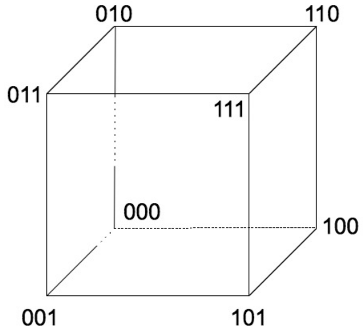


Figure 1: The 3-dimensional hypercube

To sample a random n-bit string, we can define a Markov chain on the hypercube using the *Metropolis* method as follows. Choose an index-bit pair $(i, b) \in_u \{1, 2, ..., n\}$ x $\{0, 1\}$ and change the i-th bit of the string at the current vertex (σ) to b to move to a (possibly) new vertex (τ). This defines a Markov kernel over the vertices of the hypercube as given below:

$$P(\sigma, \tau) = \begin{cases} \frac{1}{2n}; \delta_H(\sigma, \tau) = 1 \\ \\ \frac{1}{2} \; ; \sigma = \tau \; (\delta_H(\sigma, \tau) = 0) \\ \\ 0 \; ; \delta_H(\sigma, \tau) > 1 \end{cases}$$

Clearly, the state space is connected and consists of self-loops (aperiodicity), which makes the resulting Markov chain ergodic.

The same can also be understood in the following manner. At each step, with probability 1/2 we choose randomly from one of the n neighbors (i.e. flip the bit σ_i), and with probability 1/2 we stay where we are. Because we leave σ_i alone or flip it with equal probability, its new value is equally likely to be 0 or 1, regardless of what its previous value was. This means that each bit of σ becomes random whenever we touch it, so that the entire string will become random when we have touched each bit at least once. This is nothing but the *Coupon-Collector's* problem!

Thus the expected number of bits that we haven't touched at the end of t steps can be treated as a measure of how far we are then from the stationary distribution. Since the probability that a particular bit is still untouched at the end of t steps is $(1-1/n)^t$, we have

$$||\pi_t - \pi_{eq}||_{TV} \leq E[\# \text{ untouched bits}] \leq n \cdot (1-1/n)^t \leq ne^{-t/n}$$

using $1 - x \leq e^{-t}$. Setting it equal to ε gives us the following upper bound on the mixing time

$$\tau_\varepsilon < n \ln(n\varepsilon^{-1}) = O(n \ln(n)) \text{ for a constant } \varepsilon.$$

Thus the Markov chain defined above mixes rapidly.

Riffle-shuffling a deck of cards

Another interesting Markov chain is shuffling a deck of cards. We will analyze *riffle-shuffling* due to Edgar N. Gilbert, Claude Shannon, and Jim Reeds (henceforth GSR), following the statistician David Aldous and the former magician-turned-mathematician Persi Diaconis (who proved that 7 shuffles get us "reasonably" close to a truly random deck of cards), but before that let's take a look at a familiar analysis.

Imagine that the cards are numbered $1-52$ (instead of the colorful suits) and arranged so initially from top to bottom. Now take the card on the top and insert it uniformly into one of the 52 positions among the cards below it. If we do this repeatedly, then at any stage, all the cards that have been removed from the top would be uniformly distributed throughout the deck. Thus, as soon as the original bottom card is removed from the top of the deck (which, like the foregoing example of walking on the hypercube, should happen in $O(n \log(n))$ steps), the whole deck becomes random. For a deck of 52 cards, this amounts to around 200 such *top-in-at-random* shuffles.

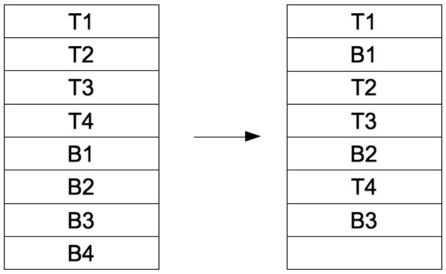
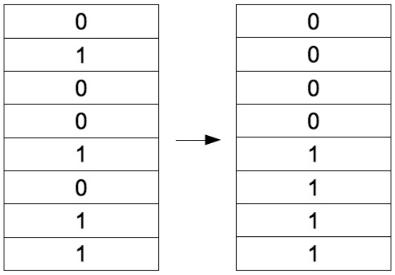


Figure 2: Riffle shuffle

In a riffle shuffle (illustrated in figure 2), the deck is split into two roughly equal parts, which are then interleaved by dropping cards from the bottoms of the two parts in a random fashion. In the GSR model of the riffle shuffle, if at any time the two parts have x and y cards remaining, the cards are dropped from the two parts with probabilities $\frac{x}{x+y}$ and $\frac{y}{x+y}$ respectively. This is better understood backwards: each card has an equal and independent chance of being pulled back into one of the two parts; the equivalent *inverse riffle shuffle* is illustrated in the figure below.



0	0
1	0
0	0
0	0
1	1
0	1
1	1
1	1

Figure 3: Inverse riffle shuffle

In the inverse riffle shuffle, we assign a bit to each card in the deck uniformly at random (u.a.r.). After all the cards have been assigned a bit, the cards labeled 0 are moved to the top and those labeled 1 are moved to the bottom. We then again assign each card a bit u.a.r. and repeat the movement. In effect, if each card were assigned a t-bit long string in the beginning itself, it would end up at a position in the final deck determined by the lexicographic order of the string assigned to it in the set of all strings assigned, unless of course two cards are assigned the same string which will lead to a collision.

Thus, if every card is assigned a different bit-string the permutation of the deck is completely determined. Moreover, since the strings are generated via random coin flips (or a walk on the hypercube), the final permutation is a random permutation. But we are already getting ahead of ourselves, because we don't even know whether analyzing this seemingly easy-to-understand model would do us any good. Turns out it will.

We now take a little detour and claim that it is enough to analyze the inverse riffle shuffle. To justify this, observe that both the riffle shuffle and the inverse riffle shuffle (indeed any card shuffling scheme) are nothing but random walks on the same group (in this case, the symmetric group S_n of all possible permutations of the n cards). We want to argue that for a random walk on this group, $\varepsilon_t = \varepsilon_t^{inv}$, where ε^{inv} denotes the variation distance from stationary distribution for the inverse walk.

Let the original (corresponding to the riffle shuffle) random walk on the group be specified by a set of generators $\{g_1, g_2, ..., g_k\}$, such that at each step a generator chosen and applied according to a fixed probability distribution μ (so that each generator has non-zero probability of being chosen). The inverse random walk is then specified in exactly the same way, but using instead the generators $\{g_1^{-1}, g_2^{-1}, ..., g_k^{-1}\}$ with the same fixed probability distribution ($\bar{\mu}(g_i) = \mu(g_i^{-1})$). Since each element of S_n has a unique inverse (permutation) under the function composition operation ($\circ$), there exists for any given state x, a bijective mapping f between the set of t-step paths starting at x in the original walk, and the set of t-step paths starting at x in the inverse walk, i.e.

$$f(x \circ \sigma_1 \circ \sigma_2 \circ ... \circ \sigma_t) = x \circ \sigma_t^{-1} \circ \sigma_{t-1}^{-1} \circ ... \circ \sigma_1^{-1}$$

This bijection preserves the probabilities of the paths since the probability distribution over the two sets of generators is the same. Moreover, if two paths reach the same state, i.e. $x \circ \sigma = x \circ \tau$, then by the group property we must have $x \circ \sigma^{-1} = x \circ \tau^{-1}$, or that the paths $f(x \circ \sigma)$ and $f(x \circ \tau)$ also reach the same state. This implies that f induces another bijective mapping f' between the set of states reachable from x in t steps of the original walk, and the set of states reachable from x in t steps of the inverse walk ($f(x\sigma) = x\sigma^{-1}$). As a result, $\pi_x^t(y) = (\pi^{inv})_x^t(f(y)) \; \forall \; y$ and t, which is just another way of saying that the distributions π_x^t and $(\pi^{inv})_x^t$ are identical upto the relabeling of the states. And since the stationary distribution π_{eq} of both the original and the inverse walk is uniform (they are both doubly stochastic), we conclude that $||\pi_x^t - \pi_{eq}|| = ||(\pi^{inv})_x^t - \pi_{eq}||$, and hence $\varepsilon_t = \varepsilon_t^{inv}$, as claimed earlier.

Going back to inverse riffle shuffle, the question that now remains is, how long each string needs to be so that, if the strings are generated at random, no two strings are the same. Why that is only the *birthday paradox* where we need to figure out the total number of days (here 2^t) to allow n people (cards) to choose their birthdays (bit-strings) from, so that there are no collisions.

The chance that any of the potential $\binom{n}{2}$ pairs of cards are assigned the same string is $1/2^t$. Thus, from union bound, the deck has a collision with probability at most $\binom{n}{2} \; 2^{-t} (\leq n^2 \; 2^{-t})$, and setting this equal to its distance ε from the random deck, we get an upper bound on t, and hence the number of (inverse) riffle shuffles required.

$$\varepsilon = n^2 \; 2^{-t} \Rightarrow t = \log_2(n^2 \; \varepsilon^{-1}) = 2\log_2(n) + \log_2(\varepsilon^{-1})$$

It turns out that the right factor in front of $\log_2(n)$ is $3/2$, which for $n = 52$ suggests that we should shuffle a deck of cards 9 times.