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

Lecture 4: September 15

Lecturer: Prof. Alistair Sinclair *Scribes: Alistair Sinclair*

4.1 Recapitulation

Recall the following definitions from last time.

$$\begin{aligned}\Delta_x(t) &= \|p_x^{(t)} - \pi\| \\ \Delta(t) &= \max_x \Delta_x(t) \\ \tau(\epsilon) &= \min_t\{\Delta(t) \le \epsilon\} \\ \tau_{\text{mix}} &= \min_t\{\Delta(t) \le \frac{1}{2e}\}\end{aligned}$$

We also proved the following:

$$\begin{aligned}\tau(\epsilon) &\le \tau_{\text{mix}}\lceil \ln \epsilon^{-1}\rceil \\ \Delta(t) &\le e^{-\lfloor t/\tau_{\text{mix}}\rfloor}\end{aligned}$$

4.2 Strong stationary times

Definition 4.1 *A* **stopping time** *is a random variable $T \in \mathbb{N}$ such that the event $\{T = t\}$ depends only on $X_0, X_1, \ldots, X_t$. A stopping time T is a* **strong stationary time** *(SST) if*

$$\forall x,\ \Pr[X_t = x \,|\, T = t] = \pi(x)\ .$$

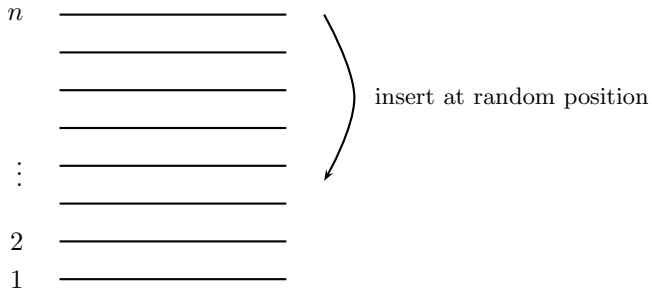
Given the definition of SST, the following claim should be intuitively reasonable:

Claim 4.2 *If T is a SST, then for any starting state x,*

$$\Delta_x(t) \le \Pr[T > t \,|\, X_0 = x]\ .$$

We will prove the claim in a moment. First, however, let us use the claim to bound the mixing time for the top-in-at-random shuffle introduced in a previous lecture.

4.2.1 Example: Top-in-at-random shuffle



Recall that this shuffle is ergodic and its stationary distribution is uniform.

Define

$$T = 1 + \text{time until the original bottom card first reaches the top.}$$

Clearly T is a stopping time. We claim that it is in fact an SST. To see this note that, at any time, all cards below the bottom card were once at the top. Hence each ordering of the cards below the original bottom card is equally likely. Once the bottom card reaches the top, we have a uniformly random permutation on all the other cards; in the next move this card will be inserted in a random position, and we get a random permutation. Hence T is indeed a SST.

We now analyze the tail probability of T. Write T as

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

where T_i is the number of steps required for the bottom card to rise from position i to $i+1$ (counting from the bottom of the deck). When the bottom card is at position i, the probability that the top card is inserted below it is $\frac{i}{n}$. So, T_i has a geometric distribution with parameter $\frac{i}{n}$.

Now we can relate this to the classical *coupon collector problem*. Recall that, at each time step, the collector gets one out of n coupons with equal probability. His aim is to continue till he has seen every coupon at least once. The time it takes to see the first coupon is 1. For the second, it is geometrically distributed with parameter $\frac{n-1}{n}$. In general, the number of time steps spent waiting for the kth new coupon is geometrically distributed with parameter $\frac{n+1-k}{n}$. The total time taken is the sum of all these times. It is now easy to see that our random variable T above has precisely this "coupon collector" distribution. The expectation of T is nH_n where H_n is the nth harmonic number. Moreover, it is well known that T is tightly concentrated around its expectation. The following inequality is left to the reader as a standard **exercise**:

$$\Pr\left[T > n \ln n + cn\right] \leq e^{-c} \ .$$

(Actually, for any fixed c, the probability tends asymptotically to $1 - e^{-e^{-c}}$.) Using Claim 4.2, we can therefore deduce that

$$\tau(\epsilon) \leq n \ln n + \lceil n \ln \epsilon^{-1} \rceil \ . \tag{4.1}$$

Thus in particular (setting $\epsilon = 1/2e$) we have the mixing time

$$\tau_{\text{mix}} \leq n \ln n + \lceil (1 + \ln 2) n \rceil \ .$$

Note that (4.1) is a better decay with ϵ than that implied by the general bound $\tau(\epsilon) \leq \tau_{\text{mix}} \lceil \ln \epsilon^{-1} \rceil$. In fact, it turns out that this Markov chain exhibits a so-called "sharp cutoff," in the sense that the variation

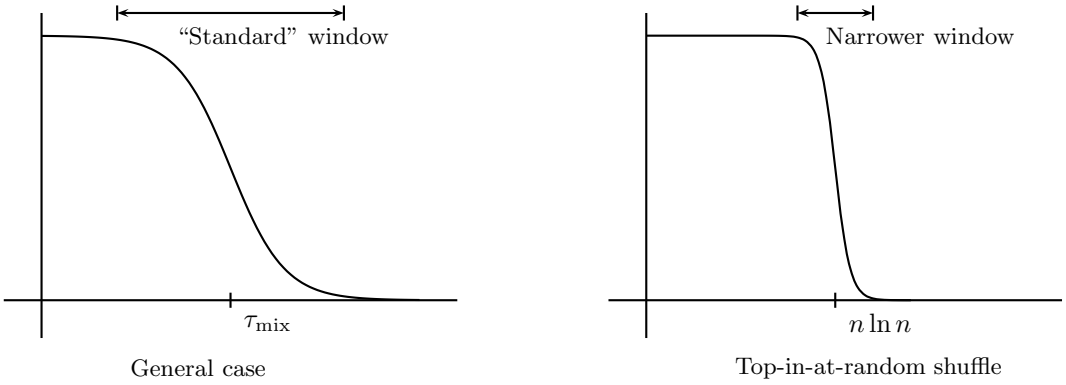


Figure 4.1: The sharp cutoff phenomenon

distance $\Delta(t)$ switches suddenly from close to 1 to close to 0 around the value $t = n \ln n$. I.e., the window over which $\Delta(t)$ decays is of smaller order than τ_{mix} itself (see Figure 4.1).

To say this precisely, write

$$t = n \ln n + n\alpha(n).$$

Then if the function $\alpha(n)$ tends to infinity with n (arbitrarily slowly), we have $\Delta(t) \to 0$ as $n \to \infty$. And if the function $\alpha(n)$ tends to minus infinity (arbitrarily slowly), then $\Delta(t) \to 1$.

The first of these facts is already immediate from (4.1). To see the second fact, we need to prove a *lower* bound on the mixing time, i.e., to prove that the total variation distance from the stationary distribution is large when t is smaller than $n \ln n$. For this, we need to find an event that has very different probabilities under the distribution of the Markov chain at time t and the stationary distribution. Denote the positions of the original bottom k cards by $C_1, \ldots, C_k$. Define the event

$$A_k = \text{“}C_1 < C_2 < \ldots < C_k\text{”} \ ;$$

that is, the cards are in the same relative order as they were initially (though other cards may have been inserted inbetween them). This event certainly holds if the card originally at the kth position from the bottom has not yet reached the top and been reinserted. Using our previous notation, the time it takes for this card to reach the top has the same distribution as

$$T_k + T_{k+1} + \ldots + T_{n-1} \ .$$

So, after t steps of the Markov chain,

$$\begin{aligned} \Pr\left[A_k \text{ holds at time } t\right] &\geq \Pr\left[\text{card at position } k \text{ hasn't been reinserted by time } t\right] \\ &= \Pr\left[T_k + T_{k+1} + \ldots + T_{n-1} + 1 > t\right] \\ &= \Pr\left[\text{coupon collector has } k \text{ or more coupons still to collect after } t \text{ steps}\right]. \end{aligned} \tag{4.2}$$

Now let $t = n \ln n - n\alpha(n)$, where $\alpha(n) \to \infty$ as $n \to \infty$ (arbitrarily slowly). A simple second moment argument (**exercise!**) then shows that, for any fixed k, the probability on the rhs of (4.2) tends to 1 as $n \to \infty$. (I.e., for any fixed k, after time only a little bit less than $n \ln n$ the coupon collector will almost surely still have k coupons left to collect.)

On the other hand, under the uniform distribution π we clearly have $\pi(A_k) = \frac{1}{k!}$. Putting these together gives, for $t = n \ln n - n\alpha(n)$,

$$\|p_x^{(t)} - \pi\| \geq 1 - \frac{1}{k!} - o(1) \quad \text{as } n \to \infty.$$

Since this holds for any fixed k, by taking k large enough and then n large enough (depending on k), we see that $\Delta(t) \to 1$ as $n \to \infty$.

Let us now go back and supply the proof of Claim 4.2.

Proof: [of Claim 4.2]. We want to show that

$$\Delta_x(t) \leq \Pr[T > t \,|\, X_0 = x] \ .$$

Recall the equivalent definition of total variation distance from Lecture 3:

$$\Delta_x(t) \equiv \| \, p_x^{(t)} - \pi \, \| = \max_{A \subseteq \Omega} | \, p_x^{(t)}(A) - \pi(A) \, | \ .$$

In what follows, we denote by T_x the SST for the chain started at state x. Now

$$\begin{aligned}
p_x^{(t)}(A) &\equiv \Pr[X_t \in A] \\
&= \Pr[X_t \in A, T_x > t] + \sum_{t' \leq t} \Pr[X_t \in A, T_x = t'] \\
&= \Pr[X_t \in A \,|\, T_x > t] \Pr[T_x > t] + \pi(A) \sum_{t' \leq t} \Pr[T_x = t'] \\
&= \Pr[X_t \in A \,|\, T_x > t] \Pr[T_x > t] + \pi(A) \left(1 - \Pr[T_x > t]\right) \\
&= \pi(A) + \Pr[T_x > t] \left(\Pr[X_t \in A \,|\, T_x > t] - \pi(A)\right)
\end{aligned}$$

The third line here follows from the definition of a SST. Since the last term in parentheses is the difference of two probabilities, it is bounded in absolute value by 1. Hence $| \, p_x^{(t)} - \pi(A) \, | \leq \Pr[T > t \,|\, X_0 = x]$, as required. ■

4.2.2 Example: Riffle shuffle

Recall from Lecture 2 the Gilbert-Shannon-Reeds model of the riffle shuffle. We now show how a SST can be used to establish an upper bound on τ_{mix} for this shuffle.

Recall that the shuffle proceeds as follows: Cut a pack of n cards into two stacks according to the binomial distribution with parameter $\frac{1}{2}$. Thus, the probability that the cut occurs exactly after k cards is given by $\frac{\binom{n}{k}}{2^n}$. Now choose an interleaving of left and right stacks uniformly at random (u.a.r.). This can be done by dropping cards one at random from the stacks with probability proportional to size of stack at that time (that is, card from left stack is dropped with probability $\frac{|L|}{|L|+|R|}$). Note that the cards of each stack always maintain their relative order. A simple calculation shows that any particular interleaving has probability of $\frac{1}{\binom{n}{k}}$ of occurring. Hence, the probability of any cut followed by any possible interleaving is $\frac{1}{2^n}$.

It turns out that to analyze riffle shuffle, it is easier to work with the *inverse shuffle*:

- Label all cards with 0's and 1's independently and uniformly at random.
- Pull out all the cards with label 0 while maintaining the relative order of the resulting stacks.
- Place the stack with label 0 on top of the stack with label 1.

As an **exercise**, you should check that this permutation is exactly the inverse of the original one, and that both have the same probability $\frac{1}{2^n}$.

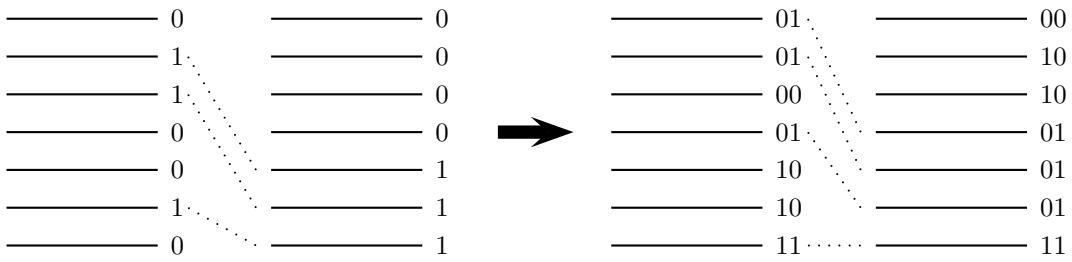


Figure 4.2: Two sequential inverse riffle shuffles

We now claim that it is enough to analyze the inverse shuffle. To justify this, we note that the riffle shuffle (and indeed any card shuffling scheme) is a random walk on a group (in this case, the symmetric group S_n on n elements). We now argue that for a general random walk on a group, $\Delta(t) = \Delta^{inv}(t)$, where Δ^{inv} denotes the variation distance for the inverse walk. More precisely, suppose the original random walk is specified by a set of *generators* $\{g_1, \ldots, g_k\}$ for the group; at each step, a generator is chosen from some fixed probability distribution (so that each generator has non-zero probability of being chosen) and applied to the current state. The inverse random walk is specified in exactly the same way, but using instead the generators $\{g_1^{-1}, \ldots, g_k^{-1}\}$ with the same probability distribution. It is easy to check (**exercise**!) that both the original walk and the inverse walk are doubly stochastic, so both have the uniform stationary distribution.

Now for any given state x, there exists a bijective mapping $\tilde{f}$ between the set of paths of t steps starting at x in the original walk, and the set of paths of t steps starting at x in the inverse walk:

$$\tilde{f}(x \circ \sigma_1 \circ \ldots \circ \sigma_t) = x \circ \sigma_t^{-1} \circ \ldots \circ \sigma_1^{-1} \ .$$

Moreover, the bijection preserves the probabilities of the paths. And if two paths reach the same state, i.e., $x \circ \sigma = x \circ \tau$, then by the group property we must have $x \circ \sigma^{-1} = x \circ \tau^{-1}$, so the paths $\tilde{f}(x \circ \sigma)$ and $\tilde{f}(x \circ \tau)$ also reach the same state. This implies that $\tilde{f}$ induces another bijective mapping f between the set of states reachable from x in t steps of the original walk and the states reachable from x in t steps of the inverse walk (namely, $f(x\sigma) = x\sigma^{-1}$) such that $p_x^{(t)}(y) = p_x^{inv(t)}(f(y))$ for all y. This means that the distributions $p_x^{(t)}$ and $p_x^{inv(t)}$ are identical up to relabeling of the points. And, since the stationary distribution π of both the original walk and the inverse walk is uniform, we conclude that $\| \ p_x^{(t)} - \pi \ \| = \| \ p_x^{inv(t)} - \pi \ \|$. Hence $\Delta(t) = \Delta^{inv}(t)$, as claimed.

Let us return now to the riffle shuffle. Suppose that we perform repeated inverse shuffles starting from an original configuration as shown in Figure 4.2 and we retain the 0–1 labels on the back of cards. After t steps, each card will be labeled by a t-digit binary number. Since during each inverse shuffle the labels are assigned u.a.r., at any time the sets of cards with *distinct* labels are in uniform random relative order (but those with the same label are in the same order as they started in). This implies that the stopping time defined by

$$T = \min\{t : \text{all cards have distinct labels}\}$$

is a SST. After t steps of inverse shuffle, the label of each card is an independent random t-bit binary number. We now make use of the well-known "birthday problem": if n people have their birthdays independent of each other and randomly distributed over cn^2 days, then Pr[some pair has the same birthday] is asymptotically $1 - \exp(-1/2c) \approx 1/2c$. In our case, the number of "birthdays" is the number of t-digit binary numbers, which is 2^t, and we want the probability that some pair of cards has a common label to be small (in particular, less than $1/2e$ to get the mixing time). So we choose c to be a (modest) constant such that $1 - \exp(-1/2c) \leq 1/2e$ and then t so that $2^t \geq cn^2$, i.e., $t \geq 2\log_2 n + \Theta(1)$. Thus the mixing time is $\tau_{\text{mix}} \leq 2\log_2 n + \Theta(1)$.

Aldous [Al83] has shown that $\tau_{\text{mix}} \sim (3/2)\log_2 n$. Thus, we have obtained an almost tight bound. For further information on SST and other stopping rules, the reader is referred to Aldous and Diaconis [AD86] and Lovász and Winkler [LW95]. Finally, Bayer and Diaconis [BD92] have found a way to explicitly calculate $\Delta(t)$ for the riffle shuffle for any value of t and for any number of cards n. (For $n = 52$, the number of possible deck arrangements is $52! \approx 8.07 \times 10^{67}$ and hence a brute force approach is not feasible, so some ingenuity is required to do this.) In particular, for $n = 52$, the exact variation distances are shown in the following table:

t	≤ 4	5	6	7	8	9
$\Delta(t)$	1.00	0.92	0.61	0.33	0.17	0.09

Thus, they argue that for practical purposes like 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.

For most Markov chains used in practice, there is no obvious choice of a strong stationary time. Hence, more sophisticated methods are needed to bound the mixing time. We will begin to discuss these in the next lecture.

4.3 References

[AD86] D. Aldous and P. Diaconis, "Shuffling cards and stopping times," *American Mathematical Monthly*, **93** (1986), pp. 333–348.

[Al83] D. Aldous, "Random walks on finite groups and rapidly mixing Markov chains," *Séminaire de Probabilites XVII*, Springer Verlag, Lecture Notes in Mathematics **986** (1983), pp. 243–297.

[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 (1986), pp. 333–348.