CS265/CME309: Randomized Algorithms and Probabilistic Analysis
Lecture #15: Mixing Times, Strong Stationary Times, and Coupling

Gregory Valiant*

November 20, 2019

1 Introduction

Last lecture we discussed stationary distributions, and also saw the Metropolis Algorithm for constructing a Markov chain that has a desired stationary distribution. This prompts the following fundamental question: *How long must we run a Markov chain until the state of the chain is close to being drawn from the stationary distribution?* The answer corresponds to the notion of *mixing time*. Before defining the mixing time, we begin by giving a few different definitions of total-variation distance, which will be a natural metric for measuring the distance between distributions.

1.1 Total Variation Distance

Definition 1. *The* total variation distance *(also referred to as* statistical distance*) between two distributions, D_1, D_2 over some countable domain, S, is defined as one half the L_1 distance:*

$$\|D_1 - D_2\| = \frac{1}{2}\sum_{s \in S} |D_1(s) - D_2(s)| = \max_{A \subset S}\left(\Pr_{D_1}[A] - \Pr_{D_2}[A]\right),$$

where $D_1(s)$ denotes the probability that distribution D_1 assigns to element s, and $\Pr_{D_1}[A] = \sum_{s \in A} D_1(s)$.

The above definition is equivalent to the following *dual* definition of total variation distance, defined in terms of any joint distribution $J_{1,2}$ over pairs (X, Y), such that the marginal distribution of X is D_1 and the marginal distribution of Y is D_2.

Fact 2. *For any such joint distribution, $J_{1,2}$ over pairs X, Y where the marginal of X is D_1 and the marginal distribution of Y is D_2, it holds that*

$$\|D_1 - D_2\| \leq \Pr[X \neq Y],$$

and there exists a joint distribution $J^*_{1,2}$ for which these quantities are equal.

Before giving a proof of the above fact, we provide an intuitive illustration of the above.

Example 3. *Suppose D_1 corresponds to a fair coin flip, and D_2 corresponds to flipping a coin that lands h with probability 0.6 and t with probability 0.4. $\|D_1 - D_2\| = \frac{1}{2}(|0.5-0.6|+|0.5-0.4|) = 0.1$. We can also define a joint distribution over pairs of outcomes X, Y as follows:*

$$\Pr[X = h, Y = h] = 0.5, \Pr[X = h, Y = t] = 0, \Pr[X = t, Y = h] = 0.1, \Pr[X = t, Y = t] = 0.4.$$

This joint distribution respects the marginals, as $\Pr[X = h] = 0.5$, and $\Pr[Y = h] = 0.5 + 0.1 = 0.6$. Additionally, $\Pr[X \neq Y] = 0.1 = \|D_1 - D_2\|$, which is consistent with the above Fact. In this example, it is also clear that we cannot modify the joint distribution to decrease $\Pr[X \neq Y]$ any more without changing the marginal probabilities.

Proof of Fact 2. Given distributions D_1 and D_2 over a countable domain, W, (the same proof holds more generally, but is a bit fussier since then one needs integrals, etc.), define $p = \sum_x \min(D_1(x), D_2(x))$. First, we claim that

$$1 - p = \|D_1 - D_2\|.$$

To see this, let $A, B \subset X$ be defined such that $A = \{x : D_1(x) \geq D_2(x)\}$, and $B = \{x : D_1(x) < D_2(x)\}$.

$$\|D_1 - D_2\| = \sum_{x \in A}(D_1(x) - D_2(x)) = \sum_{x \in B}(D_2(x) - D_1(x)),$$

and hence

$$1 - p = 1 - \sum_{x \in A} D_2(x) - \sum_{x \in B} D_1(x) = \left(\sum_{x \in A} D_1(x) + \sum_{x \in B} D_1(x)\right) - \sum_{x \in A} D_2(x) - \sum_{x \in B} D_1(x).$$

Rearranging terms, this equals

$$\sum_{x \in A}(D_1(x) - D_2(x)) = \|D_1 - D_2\|.$$

Now we will show that we can construct a joint distribution over pairs (X, Y) whose marginal distributions, respectively, are D_1 and D_2, and where $\Pr[X \neq Y] = 1 - p$. To do this, consider the following joint distribution: with probability p, select element x with probability $\frac{\min(D_1(x), D_2(x))}{p}$, and set $X = Y = x$. [From the definition of p, it is clear that these probabilities define a distribution.] With probability $1 - p$, we draw X from the distribution supported on set A that puts probability $\frac{D_1(x) - D_2(x)}{\sum_{y \in A}(D_1(y) - D_2(y))}$ on element x, and we draw Y from the analogous distribution supported on set B, that puts probability $\frac{D_2(x) - D_1(x)}{\sum_{y \in B}(D_2(y) - D_1(y))}$ on element x. This is a valid distribution, the marginal distributions are D_1 and D_2, respectively, and $\Pr[X = Y] = p$, by construction.

To see that no joint distribution can have $\Pr[X = Y] > p$, note that in the above construction, for each x, $\Pr[X = Y = x] = \min(D_1(x), D_2(x))$, and hence there is no element x for which this probability can be increased while maintaining the marginal distributions. Since this is true for all x, it is impossible for $\Pr[X = Y] = \sum_x \Pr[X = Y = x]$ to be any larger than p for any valid joint distribution. $\square$

2 Mixing Times

The following quantity, $\Delta(t)$, will measure the worst-case distance of a Markov chain to the stationary distribution, where the "worst-case" is with respect to selecting the starting state.

Definition 4. *Given a finite, irreducible, aperiodic Markov chain $\{X_t\}$ with stationary distribution π, let*

$$\Delta(t) = \max_s \|\pi - P_s^t\|,$$

where P_s^t denotes the distribution of X_t, conditioned on $X_0 = s$.

We are now ready to define the mixing time of a Markov chain. The mixing time will be the first time, t, such that no matter what state one starts the chain in, the distribution of the state at time t is close, in total variation distance, to the stationary distribution.

Definition 5. *The mixing time, of a Markov chain with stationary distribution π, will be denoted τ_{mix}, and is defined as*

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

The choice of constant $\frac{1}{2e}$ in the definition of mixing time is somewhat arbitrary, and in some cases, people just replace that constant by $\frac{1}{2}$. The reason it doesn't matter is captured by the following fact, whose proof we will see after we develop an understanding of "couplings".

Fact 6. *For any finite, irreducible, aperiodic Markov chain, $\Delta(t)$ is non-increasing, namely for all t, $\Delta(t+1) \leq \Delta(t)$. Additionally, for any constant $c \geq 1$,*

$$\Delta(c \cdot \tau_{mix}) \leq \frac{1}{e^c}.$$

2.1 Bounding the Mixing Time via Strong Stationary Stopping Times

The following approach to bounding the mixing time is cool when it works, though there are many Markov chains for which it doesn't work. If you're asked to bound the mixing time of a chain, it is often worth spending a few moments thinking whether you can use this technique, but don't expect it to always work.

Definition 7. *Given a Markov chain $\{X_t\}$ with stationary distribution π, a* strong stationary stopping time *is a random variable T defined in terms of the random variables $X_0, X_1, \ldots,$ with the property that the event that $T = t$ depends only on $X_0, \ldots, X_t$, and that for all states s,*

$$\Pr[X_t = s | t \geq T] = \pi(s).$$

The condition in the definition that the event that $T = t$ depends only on $X_0, \ldots, X_t$, makes sure that you wouldn't need to "look into the future" of a chain to figure out whether T has already happened or not. The condition that $\Pr[X_t = s | t \geq T] = \pi(s)$ means that once T has happened, the chain is completely mixed, in the sense that the chain is at the stationary distribution.

Fact 8. *Given a Markov chain with stationary distribution π, and a strong stationary stopping time T, for any time $t \geq 0$,*

$$\Delta(t) \leq \Pr[T > t].$$

Proof. We can re-express the distribution after time t, when starting in state s, P_s^t as the weighted combination of this distribution conditioned on $t \geq T$ and on $t < T$:

$$P_s^t = \pi \cdot \Pr[t \geq T] + q \cdot \Pr[t < T],$$

for some distribution q. Hence

$$\Delta(t) = \|\pi - P_s^t\| = \|\pi - \pi \cdot \Pr[t \geq T] - q \cdot \Pr[t < T]\| = \Pr[t < T]\|\pi - q\| \leq \Pr[t < T],$$

since the distance between any two distributions is at most 1. $\square$

The following example provides a nice illustration of how a strong stationary stopping time can be fruitfully used.

Example 9 (Top in at Random Shuffle)**.** *Consider the shuffling scheme where we have a stack of n cards, and iteratively take the top card, and insert it into a uniformly random position in the stack (and with probability $1/n$ we insert it into the top spot, in which case the ordering remains unchanged at that iteration). It is not hard to show that this chain is irreducible and aperiodic, and that the stationary distribution is the uniform distribution over the $n!$ orderings of the deck. How many times iteration must we run this shuffle until the deck is mixed (i.e. close to being at a random ordering)?*

Define the stopping time T to be one plus the first time at which the card that started at the bottom of the deck has reached the top. (I.e. if the Ace of Spades started at the bottom of the deck at time 0, the stopping time is the timestep after the first time that it reaches the top.) To see that this is a valid stationary stopping time, consider that, at any time before T in the shuffle, given the identities of the cards below this bottom card, their ordering is uniformly random. Hence, at the first time where this bottom card has reached the top, we have a uniformly random ordering of the $n-1$ other cards, with this bottom card at the top. At the next timestep, we have inserted this in a uniformly random index, and hence have a uniformly random ordering.

To analyze $\mathbf{E}[T]$, note that as we do the shuffle, the bottom card will monotonically rise in the deck, until its at the top. When it has index $i > 1$, the probability it moves up by one in the deck is $(n-i+1)/n$ and with the remaining probability, it stays in the same location. Hence

$$\mathbf{E}[T] = 1 + \frac{n}{1} + \frac{n}{2} + \frac{n}{3} + \ldots \approx n \log n,$$

and by Markov's inequality, $\Pr[T > 2en \log n] \leq 1/(2e)$, and hence by Fact 8 the mixing time of this shuffling is at most $(2e)n \log n$. On the homework, you will actually show that the mixing time is the range $[0.9n \log n, 1.1n \log n]$.

3 Coupling

To motivate the connection between couplings, the dual definition of total variation distance, and mixing time, consider the following basic fact about $\Delta(t)$:

Fact 10.

$$\Delta(t) = \max_s \|P_s^t - \pi\| \leq \max_{s,s'} \|P_s^t - P_{s'}^t\| \leq 2\Delta(t).$$

Proof. Recall that one property of the stationary distribution, π, is that if we select X_0 according to π, and run the chain for any number of steps, the distribution of the chain at time t is π. Hence $\pi = \sum_w \pi(w)P_w^t$, and $\max_s \|P_s^t - \pi\| = \max_s \|P_s^t - \sum_w \pi(w)P_w^t\|$. Since π is a weighted average of the P_w^t for different w's, this distance is at most the distance between the two furthest points, namely $\max_{s,s'} \|P_s^t - P_{s'}^t\|$. The final inequality in the statement of the fact is from the triangle inequality, since

$$\|P_s^t - P_{s'}^t\| \leq \|P_s^t - \pi\| + \|\pi - P_{s'}^t\|.$$

□

The idea behind couplings is to directly bound $\max_{s,s'} \|P_s^t - P_{s'}^t\|$ by showing that for any two initial states s, s' one can construct a joint distribution over pairs (X_t, Y_t) where X_t is drawn from P_s^t and Y_t is drawn from $P_{s'}^t$ such that $\Pr[X_t \neq Y_t]$ is as small as possible. Recalling the dual definition of total variation distance (Fact 2), if we can prove that $\Pr[X_t \neq Y_t]$ is sufficiently small for some value of t, then we will have bounded $\Delta(t)$, and hence the mixing time. The following definition formalizes the properties that we require of this joint distribution:

Definition 11. *Given a Markov process, defined by transition probabilities P, a* coupling *is a joint process $(X_0, Y_0), (X_1, Y_1), \ldots$ such that the following conditions hold:*

1. *The marginal distributions of $\{X_t\}$ and $\{Y_t\}$ correspond to the Markov process, namely for all states s, s',*

$$\Pr[X_t = s | X_{t-1} = s'] = P_{s,s'}, \Pr[Y_t = s | Y_{t-1} = s'] = P_{s,s'}.$$

2. *If $X_t = Y_t$, then $X_{t+1} = Y_{t+1}$, namely once the two chains meet/"couple", they stay together for good.*

Proposition 12. *Given a (valid) coupling $\{(X_t, Y_t)\}$ of a Markov chain, let $T_{s,s'} = \min\{t : X_t = Y_t | X_0 = s, Y_0 = s'\}$. Then*

$$\Delta(t) \leq \max_{s,s'} \Pr[T_{s,s'} \geq t].$$

Proof. From Fact 10,

$$\Delta(t) \leq \max_{s,s'} \|P_s^t - P_{s'}^t\| \leq \max_{s,s'} \Pr[X_t \neq Y_t | X_0 = s, Y_0 = s'] \leq \max_{s,s'} \Pr[T_{s,s'} \geq t],$$

where the second-to-last inequality is from the dual definition of total variation distance given in Fact 2. □

How do we make a valid coupling? One option is to have the chains $\{X_t\}$ and $\{Y_t\}$ evolve *independently*, up until the first time that $X_t = Y_t$, after which point they evolve together (according to the Markov process). This is a perfectly valid coupling.

The punchline from the above proposition, however, is that if we want a *good* bound on the mixing time, we need to design a coupling that gets $X_t = Y_t$ *as fast as possible*—namely we are trying to get X_t and Y_t to "couple" as soon as possible, no matter their initial states. The "independent" coupling, while valid, isn't trying to get the chain to meet especially quickly, and in many cases, such a coupling would give rather bad bounds on the mixing time, in comparison to the bounds that could be proved via more creative couplings.

3.1 Illustration I: Shuffling Revisited

As a first illustration, consider the shuffling protocol that is the "time reverse" process to the shuffling considered in Example 9. Specifically, consider the shuffling protocol where, at each step, a uniformly random card is selected, and then moved to the top of the deck. Given two initial orderings, s and s', consider defining the joint process (X_t, Y_t) as follows. $X_0 = s$ and $Y_0 = s'$, and at each step, the X chain chooses a uniformly random card and moves it to the top of the deck. The chain Y, *selects the same card* and moves that to the top. Here, when we say "selects the same card", we don't mean that the index of the card is the same, we mean that the value of the card is the same—if X selects the 8 of hearts to move to the top, then chain Y will also select the 8 of hearts, and move that to the top.

Why is this a valid coupling? Well, its clear that the marginal distributions are still consistent with the original chain, because the choice of the card is still uniformly random (its just the SAME randomness in the X and Y chain).

Why is this a good choice of coupling? Well letting S_t denote the set of cards that have been selected up through time t, the chains X_t and Y_t will both have the top $|S_t|$ cards on their decks being the cards in set S_t, and they will have the same order. Hence, the time until the chains couple is at most the time until all n cards are selected (no matter the initial orderings of the deck, namely for all s, s'). This time until all n cards have been selected at least once is *exactly* the coupon-collector problem, and hence we know that, with high probability, this time is bounded by, say $nlogn + o(n)$.

If, instead of this coupling, we used the naive coupling where the chains X_t and Y_t evolved independently until the first time they happen to be the same, the expected time until the chains couple would be $> n!$.

3.2 Illustration II: Random Proper Colorings

Given a graph with degree at most d, one can describe a simple Markov chain over proper k-colorings of the graph. (Recall that a coloring of the nodes with one of k colors is "proper" if no two neighboring nodes are colored with the same color.) Let X_0 denote some initial proper coloring. To obtain X_t from X_{t-1}, choose a vertex v uniformly at random from the vertices, and a color c uniformly at random from the k colors, and color vertex v with color c if that would be a valid coloring (given the assignment of colors to the other vertices as specified by X_{t-1}). If coloring v with c would not result in a proper coloring, then $X_t = X_{t-1}$.

Since the graph in question has degree at most d, there exists a coloring with $k = d + 1$ colors, since for any assignment of colors to the neighbors of a node, v, there is at least $k - degree \geq 1$ choice of color one could use to color node v so that it respects its neighbors' colors.

The following conjecture states that, as long as their is one extra bit of flexibility beyond this, namely if $k \geq d + 2$, then the above Markov chain mixes in polynomial time. (Its not hard to show that the stationary distribution of this chain is the uniform distribution over proper colorings—since this chain has symmetric transition probabilities, and is aperiodic and irreducible). If this chain mixes fast, then this provides a natural way of efficiently answering questions like "What is the probability node 3 and node 7 have the same color in a random proper coloring?"

Conjecture 13. *Given a graph with maximum degree d, the above Markov chain over proper colorings will mix in time polynomial in the size of the graph, provided* $k \geq d + 2$.

We do not know whether the above conjecture is true of not. If $k \geq 2d+1$, we can prove the analog of the above conjecture, via the construction of a fairly fancy coupling. If $k \geq 4d+1$, as we show below, a relatively simple coupling will suffice to prove fast mixing.

Proposition 14. *Given a graph with n vertices with maximum degree d, the above Markov chain over proper colorings will mix in time polynomial in the size of the graph, provided $k \geq 4d+1$. Specifically, provided $k \geq 4d+1$, the mixing time is bounded by $kn(2+\log n)$.*

Proof. Consider the coupling $\{X_t, Y_t\}$ where X_t evolves independently, and, at each step, Y_t chooses *the same node v and the same color c*. Just because X and Y choose the same vertex and color does not mean that the vertex will be colored identically in X and Y after that step—there is some probability that X will "accept" that color, but Y cannot accept that color because of the colors of the neighboring nodes. We will show, however, that this sort of issue won't happen too often.

To that end, let Z_t denote the number of vertices in the graph for which X_t and Y_t assign different colors. Out of the nk possible choices of vertex v and color c, how many will result in Z_{t+1} decreasing by one? Well, we need to first choose one of the Z_t vertices for which the colors differ, and then we need to pick a color c that is not one of the $\leq d$ colors of a neighboring vertex in X_t or one of the $\leq d$ colors of a neighboring vertex in Y_t. Hence there ar at most $Z_t(k-2d)$ such good choices of v and c. What is the probability that we make negative progress, namely that $Z_{t+1} = Z_t + 1$? Well we need to pick a vertex that is a neighbor of one of the Z_t vertices on which the colorings differ, and a color which is the color of that neighbor in one of the two chains, hence there are $dZ_t \cdot 2$ such bad options, and the probability is $\frac{2dZ_t}{nk}$. Putting this together

$$\mathbf{E}[Z_{t+1}|Z_t] \leq Z_t + \frac{2dZ_t}{kn} - \frac{Z_t(k-2d)}{nk} = Z_t - Z_t\left(\frac{k-4d}{kn}\right).$$

Hence provided $k \geq 4d+1$, $\mathbf{E}[Z_{t+1}] \leq \mathbf{E}[Z_t](1-\frac{1}{kn})$, and inductively we have that

$$\mathbf{E}[Z_t] \leq Z_0(1-\frac{1}{kn})^t \leq ne^{-\frac{t}{kn}},$$

where we used the fact that $Z_0 \leq n$, and the fact that $1-\alpha \leq e^{-\alpha}$. Since Z_t is a non-negative integer,

$$\Pr[Z_t > 0] = \Pr[Z_t \geq 1] \leq \mathbf{E}[Z_t] \leq ne^{-\frac{t}{kn}},$$

and hence if $t \geq kn(2+\log n)$, this quantity is less than $1/e^2 < 1/2e$. Hence the mixing time is bounded by $kn(2+\log n)$. □