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:

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

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

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

*©2019. Gregory Valiant. Not to be sold, published, or distributed without the authors’ consent.*
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. 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 $D_2$...

## 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 $\pi$, let

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

where $P_t^s$ 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 $\pi$, will be denoted $\tau_{mix}$, 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
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, its 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 \( \pi \), 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 \( \pi \), 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 \( s \), \( P_t^s \) as the weighted combination of this distribution conditioned on \( t \geq T \) and on \( t < T \):

\[
P_t^s = \pi \cdot \Pr[t \geq T] + q \cdot \Pr[t < T],
\]

for some distribution \( q \). Hence

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

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

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
To analyze $\mathbb{E}[T]$, 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

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

and by Markov’s inequality, $\Pr[T > 2e \log n] \leq 1/(2e)$, 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. $\Delta(t) = \max_s ||P^t_s - \pi|| \leq \max_{s,s'} ||P^t_s - P^t_{s'}|| \leq 2\Delta(t)$.

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

The idea behind couplings is to directly bound $\max_{s,s'} ||P^t_s - P^t_{s'}||$ 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 $P^t_s$ and $Y_t$ is drawn from $P^t_{s'}$ 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,
\[
\Delta(t) \leq \max_{s,s'} \|P^t_s - P^t_{s'}\| \leq \max_{s,s'} \Pr[X_t \neq Y_t | X_0 = s, Y_0 = s'] \leq \max_{s,s'} \Pr[T_{s,s'} \geq t],
\]
where the second-to-last inequality is from the dual definition of total variation distance given in Fact 2.

The punchline from the above proposition 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.

[TODO: examples—shuffling example and random walks over proper colorings]

3.1 Illustration I: Shuffling Revisited

3.2 Illustration II: Random Proper Colorings