1 Variation Distance

If we run a Markov Chain for a finite time, how do we know if it has reached its stationary distribution. For example, think of shuffling cards by choosing a card independently and uniformly at random and putting this card on the top of the deck. This process is a Markov Chain. (You can check that the chain is finite, irreducible, aperiodic, and that the stationary distribution is the uniform distribution).

How many steps before we obtain a shuffle which is close to uniformly distributed? We introduce the following distance measure, to quantify what 'close to the stationary distribution' means:

Definition 1.1 (Variation Distance)**.** *The **variation distance** between two distribution, D_1 and D_2 on a countable space S is given by*

$$\|D_1 - D_2\| = \frac{1}{2}\sum_{x\in S} |D_1(x) - D_2(x)| \tag{1}$$

Lemma 1.1. *For any $A \subset S$, let $D_i(A) = \sum_{x\in A} D_i(x)$ for $i = 1, 2$. Then:*

$$\|D_1 - D_2\| = \max_{A\subset S} |D_1(A) - D_2(A)| \tag{2}$$

Proof. Let $S^+ \subset S$ be the set of states s.t $D_1(x) \geq D_2(x)$, and let $S^- \subset S$ be the set of states s.t $D_2(x) \geq D_1(x)$. We have

$$\max_{A\subset S} D_1(A) - D_2(A) = D_1\left(S^+\right) - D_2\left(S^+\right) \tag{3}$$

and

$$\max_{A\subset S} D_2(A) - D_1(A) = D_2\left(S^-\right) - D_1\left(S^-\right) \tag{4}$$

But $D_1(S) = D_2(S) = 1$, we have

$$D_1\left(S^+\right) - D_1\left(S^-\right) = D_2\left(S^+\right) - D_2\left(S^-\right) = 1 \tag{5}$$

which implies that

$$D_1\left(S^+\right) - D_2\left(S^+\right) = D_2\left(S^-\right) - D_1\left(S^-\right) \tag{6}$$

so,

$$\max_{A\subset S} |D_2(A) - D_1(A)| = |D_1\left(S^+\right) - D_2\left(S^+\right)| = |D_1\left(S^-\right) - D_2\left(S^-\right)| \tag{7}$$

so,

$$D_1\left(S^+\right) - D_2\left(S^+\right)| + |D_1\left(S^-\right) - D_2\left(S^-\right)| \tag{8}$$

$$= \sum x \in S|D_1\left(x\right) - D_2\left(x\right)| \tag{9}$$

$$= 2\left\|D_1 - D_2\right\| \tag{10}$$

taking the maximum completes the proof. □

Suppose we take a 52-card deck and shuffle all the cards - but leaving the ace of spades on top. We can thus bound the variation distance between the resulting distance between the resulting distribution D_1 and the uniform distribution U by considering the set of states B where the ace of spades is on the top of the deck:

$$\left\|D_1 - U\right\| = \max_{A \subset S}|D_1\left(A\right) - U\left(A\right)| \geq \tag{11}$$

$$|D_1\left(B\right) - U\left(B\right)| = 1 - \frac{1}{52} \tag{12}$$

Definition 1.2 (Mixing time)**.** *Let π be the stationary distribution of a Markov Chain with state space S. Let p_x^t be the distribution of the state of the chain starting at state x after t steps. We define*

$$\Delta_x\left(t\right) = \left\|p_x^t - \pi\right\| \tag{13}$$

$$\Delta\left(t\right) = \max_{x \in S}\Delta_x\left(t\right) \tag{14}$$

We also define

$$\tau\left(\epsilon\right) = \min\left\{t : \Delta\left(t\right) \leq \varepsilon\right\} \tag{15}$$

and

$$\tau\left(\varepsilon\right) = \max_{x \in S}\tau\left(\varepsilon\right) \tag{16}$$

That is (13) is the variation distance between the stationary distribution and p_x^t and (14) is the maximum of these values over all states x. (15) is the first step at which the variation distance between the stationary distribution and p_x^t is less than ε. (16) is the maximum of these values over all states x, it is called the ***mixing time*** *of the Markov Chain.*

A chain is called rapidly mixing if $\tau\left(\varepsilon\right)$ is polynomial in $\log\frac{1}{\varepsilon}$ and the size of the problem (e.g. the number of cards in a deck).

2 Coupling

Definition 2.1 (Coupling)**.** *A* ***coupling*** *of a Markov Chain M_t with state space S is a Markov Chain $Z_t = (X_t, Y_t)$ on the state space $S \times S$ such that:*

$$\mathbb{P}\left(X_{t+1} = x' \mid Z_t = (x,y)\right) = \mathbb{P}\left(M_{t+1} = x' \mid M_t = x\right) \tag{17}$$

$$\mathbb{P}\left(Y_{t+1} = y' \mid Z_t = (x,y)\right) = \mathbb{P}\left(M_{t+1} = y' \mid M_t = y\right) \tag{18}$$

A coupling is simply two copies of the same MC running simultaneously - for example two independent runs of the same MC. We will be interested couplings that bring two copies of the same chain to the the same state, and then keep them there.

Lemma 2.1 (Coupling Lemma)**.** *Let $Z_t = (X_t, Y_t)$ be a coupling for a Markov Chain M_t on a state space S. Suppose $\exists$ a T s.t, for every $x, y \in S$*

$$\mathbb{P}\left(X_T \neq Y_T \mid X_0 = x, Y_0 = y\right) \leq \varepsilon \tag{19}$$

Then

$$\tau\left(\varepsilon\right) \leq T \tag{20}$$

Proof. Consider the coupling when Y_0 is chosen according to the stationary distribution and X_0 takes any arbitrary value. For the give T and ε and for any $A \subset S$,

$$\begin{aligned}
\mathbb{P}\left(X_T \in A\right) &\geq \mathbb{P}\left(\left(X_T = Y_t\right) \cap \left(Y_T \in A\right)\right) && (21)\\
&= 1 - \mathbb{P}\left(\left(X_T \neq Y_t\right) \cup \left(Y_T \notin A\right)\right) && (22)\\
&\geq \left(1 - \mathbb{P}\left(Y_T \notin A\right)\right) - \mathbb{P}\left(X_T \neq Y_T\right) && (23)\\
&\geq \mathbb{P}\left(Y_T \in A\right) - \varepsilon && (24)\\
&= \pi\left(A\right) - \varepsilon && (25)
\end{aligned}$$

The second line follows from the union bound, the third because

$$\mathbb{P}\left(X_T \neq Y_T\right) \leq \varepsilon \tag{26}$$

for **any** initial states X_0, Y_0. $\mathbb{P}\left(Y_T \in A\right) = \pi\left(A\right)$ as Y_T is distributed according to the stationary distribution. Repeating the argument for the set $S - A$ shows that $\mathbb{P}\left(X_T \notin A\right) \geq \pi\left(S - A\right) - \varepsilon$, or $\mathbb{P}\left(X_T \in A\right) \leq \pi\left(A\right) + \varepsilon$. It follows that,

$$\max_{x \in A} |p_x^T\left(A\right) - \pi\left(A\right)| \leq \varepsilon \tag{27}$$

so by Lemma 1.1 the variation distance from the stationary distribution after the chain runs for T steps is bounded above by ε. □

2.1 Variation distance is non-increasing

Lemma 2.2. *Given distributions σ_x, σ_y on a space S, let $z = (X, Y)$ be a random variable on $S \times S$ where X is distributed according to σ_x and Y is distributed according to σ_y. Then*

$$\mathbb{P}(X \neq Y) \geq \|\sigma_x - \sigma_y\| \tag{28}$$

Moreover, there exists a joint distribution Z where equality holds.

Proof. For each $s \in S$ we have

$$\mathbb{P}(X = Y = x) \leq \min\left(\mathbb{P}(X = x), \mathbb{P}(Y = x)\right) \tag{29}$$

So, summing over $x \in S$,

$$\mathbb{P}(X = Y) \leq \sum_{x \in S} \min\left(\mathbb{P}(X = x), \mathbb{P}(Y = x)\right) \tag{30}$$

so,

$$\mathbb{P}(X \neq Y) \geq 1 - \sum_{x \in S} \min\left(\mathbb{P}(X = x), \mathbb{P}(Y = x)\right) \tag{31}$$

$$= \sum_{x \in S} \mathbb{P}(X = x) - \min\left(\mathbb{P}(X = x), \mathbb{P}(Y = x)\right) \tag{32}$$

For $\sigma_x < \sigma_y$ the second term in the sum is 0. For $\sigma_y < \sigma_x$ it is:

$$\mathbb{P}(X = x) - \mathbb{P}(Y = x) = \sigma_x - \sigma_y. \tag{33}$$

Following lemma 1.1 this is equal to $\|\sigma_x - \sigma_y\|$ □

3 Geometric Convergence

Theorem 3.1 (General Mixing)**.** *Let $\boldsymbol{P}$ be the transition matrix for a finite, irreducible, aperiodic Markov Chain. Let m_j be the smallest entry in the j^th column, and let $m = \sum_j m_j$. Then for all x and t,*

$$\left\|p_x^t - \pi\right\| \leq (1 - m)^t \tag{34}$$

Proof. If the minimum entry in column j is m_j, then in one step the chain reaches state j with probability m_j. Hence, a coupling can be created that achieves this. This holds for all j, so at each step the two chains can be made to couple with probability at m. Hence the probability they have not coupled after m steps is at most $(1 - m)^t$, using the coupling lemma. (Should this m be a t?). □

I.e. under very general conditions, Markov Chains converge quickly to their stationary distributions, with the variation distance converging geometrically in the number of steps. If the mixing time is bounded from above, we can refine this result.

Theorem 3.2 (Refined Mixing)**.** *Let $\boldsymbol{P}$ be the transition matrix for a finite, irreducible, aperiodic Markov Chain M_t with $\tau(c) \leq T$ for some $c \leq 1/2$. Then for this Markov Chain, $\tau(\varepsilon) \leq \lceil \ln \varepsilon / \ln 2c \rceil T$*

Proof. Consider any two initial states X_0, Y_0. We have $\left\| p_x^T - \pi \right\| \leq c$ by definition of $\tau(c)$. Thus, $\left\| p_x^T - p_y^T \right\| \leq 2c$. By lemma 2.2 $\exists$ a random variable $Z_{T,x,y} = (X_T, Y_T)$ with the marginals distributed according to $\left(p_x^T, p_y^T\right)$ s.t $\mathbb{P}(X_T \neq Y_T) \leq 2c$. Now consider a Markov Chain with transition matrix $\mathbf{P}^T$. $Z_{T,x,y}$ is a coupling for this chain. This guarantees that the probability that the two states have not coupled in one step is at most $2c$. For k steps it is $(2c)^k$. By the coupling lemma, this new chain is within variation distance ε of it's stationary distribution after k steps if:

$$(2c)^k \leq \varepsilon \tag{35}$$

I.e.

$$k \leq \lceil \frac{\ln \varepsilon}{\ln 2c} \rceil \tag{36}$$

steps until the new chain is within ε of it's stationary distribution. However, each T steps of the old chain corresponds to a single step of the new chain, so

$$\tau(\varepsilon) \leq \lceil \frac{\ln \varepsilon}{\ln 2c} \rceil T \tag{37}$$

□

4 Applications

4.1 Card Shuffling

Think of shuffling cards by choosing a card independently and uniformly at random and putting this card on the top of the deck. This process is a Markov Chain. We can choose a coupling first we obtain X_{t+1} from X_t by choosing a position j uniformly at random, and moving the j^{th} card to the top, denoting the value of the card c. We obtain Y_{t+1} from Y_t, by moving the card with value c to the top. In both chains, the probability that a specific card is moved to the top is $1/n$, and once a card is moved to the top it always remains in the same position in both copies of the chain. Now, to bound the number of steps until the copies become coupled we simply bound how many times cards must be chosen uniformly at random before every card is chosen once (coupon collecting). So, we know that if the Markov chain runs for m steps, the probability that a specific card has not been moved to the top is at most:

$$\left(1 - \frac{1}{n}\right)^m \tag{38}$$

Choosing m such that $m \geq n \ln n + cn$ the probability is:

$$\left(1 - \frac{1}{n}\right)^{n \ln n + cn} \leq e^{-\ln n + cn} \tag{39}$$

and so by the union bound, the probability that any card has not been moved to the top is at most e^{-c}. Choosing $c = \ln(1/\varepsilon)$, and the coupling lemma allows us to conclude that after $n \ln\left(\frac{n}{\varepsilon}\right)$ steps the variation distance between the uniform distribution and the state of the chain is bounded above by ε.

4.2 Routing on the Hypercube

4.3 Approximately counting proper coulorings