Coupling

2/3/2010 and 2/5/2010

1 Introduction

Consider the *move to middle* shuffle where a card from the top is placed uniformly at random at a position in the deck. It is easy to see that this Markov Chain is simply a walk on permutations of the cards, that the state space is connected, and that the Markov Chain is aperiodic. This therefore implies that the stationary distribution, Π is uniform. We hope that we can demonstrate that this Markov Chain is rapidly mixing, that is that the Markov Chain reaches the stationary distribution in a polynomial amount of time relative to the graph size, and exponentially quickly relative to the size of the state space, Ω.

Our goal is to discover a way to bound the mixing time of this Markov Chain. Unfortunately, simple analysis of the chain is very difficult. It is difficult do determine when the chain is sufficiently mixed. A tool we can used is called coupling.

Informally, *coupling* involves using not one, but two initial states in order to bound the mixing time of the Markov Chain. It is provable to say that the Markov Chains have reached stationarity when the two Markov Chains become identical. More formally, define $x, y \in \Omega$. We will use some mechanism to ensure that x and y perform coupled transitions. How we define our coupling and these transitions is critical to ensure that we accurately bound the mixing time of the Markov Chain.

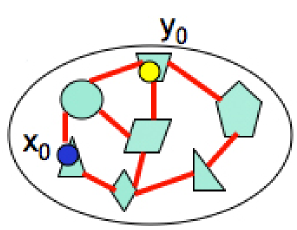


Figure 1: Starting position of coupling, x_0 and y_0 shown.

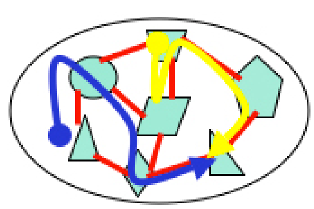


Figure 2: After some number of steps, the X and Y chains have coupled.

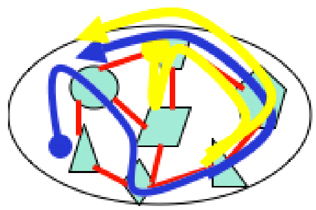


Figure 3: Once coupled, the X and Y chains do not separate.

One of the simplest ways to couple two random walks on the state space is to force them to perform identical moves (where possible). For example: starting at two random points, if point x moves up, force y to also move up. Obviously in this random walk the coupling does not force the states closer together, they remain the same distance from each other at all states. We will more cleverly define coupling in order to allow our states to move closer to one another.

2 Wacky, Informal Arguments

Let us return to the example of shuffling. Previously we were considering the mixing time of the *move to middle* shuffle. When trying to think about a coupling for this shuffle, it is very difficult to see how any coupling would bring states closer together. Instead, let us look at the reverse Markov Chain, where a card is chosen uniformly at random from the deck, and that card is moved to the top. We claim that this shuffle method has an identical mixing time to the *top to middle* shuffle and that there is an easy to define coupling for this chain.

Consider two initial configurations, $x_0, y_0 \in \Omega$. Let x_0 be any distribution, and y_0 be chosen from the uniform distribution ($y_0 \in \Pi$). On x_0, we choose a card uniformly at random from the deck and bring it to the top. Suppose that we choose the 27th card from x_0 to bring to the top, which happens to be $8\heartsuit$. For the coupling, instead of uniformly randomly choosing a card, we will search the deck for the $8\heartsuit$ and bring it to the top.

Initially it may seem that this does not preserve the Markov Chain because the move is chosen by x_0 instead of y_0 choosing its move uniformly randomly from the possible states. This concern, however, is unwarranted because each card is equally likely to be chosen by x_0, which implies that y_0's move is faithful to the original Markov Chain. The moves for all future y_i follow the same convention, but more specifically once a card is "coupled" between the two distributions, it always remains coupled. Clearly, because of the nature of the coupling, when a card is moved to the top, no uncoupled cards can be above any two coupled cards, and they always remain at the same height as the other from the top.

We claim that the mixing time of this Markov Chain is strictly less than the the coupling time of these coupled chains. Clearly, this is an example of the coupon collector's problem, and takes $O(n \log n)$ time before all the cards are selected with probability $P = 1 - \epsilon$. Because y_0 was selected from Π, when we stop after k steps, y_k is also from the uniform distribution. Because x_k and y_k are now coupled, we can presume that x_k is also sampled from the uniform distribution of Ω, and therefore this Markov Chain is rapidly mixing.
This chain has the same mixing time as the the *top to middle* shuffle.

3 Colorings

Consider the problem of sampling from colorings on a graph. Let $G = (V, E)$ be any graph with max degree Δ, and some number of colors $k \in \mathbb{Z}$. Our goal is to find a k coloring uniformly at random from the state space of possible colorings, Ω.

There is a problem with our goal, however. First, the k coloring may not exist, and deciding whether this coloring exists is NP-Complete. A corollary of this problem is that because the decision problem is NP-Complete, it is also NP-Complete to develop a starting configuration for our walk to begin on.

Because of the hardness of finding colorings for small values of k, we will be dealing exclusively with values where $k \geq \Delta + 1$, where it is trivial to assign a k-coloring using a greedy algorithm. Because k is strictly greater than the largest degree, it is always possible to assign a color to a vertex regardless of the colors of the surrounding (at most) Δ vertices.

We will define a Markov Chain to walk on the state space of these colorings as follows:

Markov Chain: M_{col}

1. Given $G = (V, E)$ at a valid coloring
2. Select $v \in_u V$
3. Select color $c \in_u 1, 2, ..., k-1, k$
4. Recolor v with c if it produces a valid k-coloring. Otherwise keep v colored as it is.

It is clear to see that any particular edge in the Markov Chain combined with a color is chosen with probability $\frac{1}{|V|k}$, which is important to ensure that the appropriate stationary distribution is reached. Unfortunately, when $k = \Delta$, the state space is disconnected. (Proof: Consider K_n where $k = n$. No valid coloring exists.) Additionally, on the torus, when $k = +1$, the state space is also disconnected in some instances. (Proof: Let the central node be yellow, its north neighbor be white, its east neighbor be blue, its south neighbor be red, and its west neighbor be red. Repeat this pattern for the torus. No valid transitions exist for this Markov Chain.) It is left to the reader to show that when $k \geq \Delta + 2$, M_{col} is ergodic.

We will now begin a series of proofs to show when this chain is rapidly mixing. The state of the art result [Jerrum] demonstrates that when $k \geq 2\Delta$, M_{col} is rapidly mixing. As a sidenote, Jerrum has shown that 8-colorings are rapidly mixing on the grid, a specialized case of general graphs. It has also been demonstrated that when $k = 6$ or $k = 7$, the Markov chain is also rapidly mixing. $k = 3$ has also been shown to be rapidly mixing by Randall. An open problem is whether 4 and 5 colorings are also rapidly mixing for the grid.

We will begin by demonstrating that the Markov Chain is rapidly mixing for $k \geq 4\Delta$. We will then introduce a technique called path coupling to prove rapid mixing for $k \geq 3\Delta$, and lastly we will improve on this bound to show that M_{col} is rapidly mixing when $k \geq 2\Delta$. First, however, we must demonstrate that the time for couplings to reach each other is greater than the mixing time for an algorithm.

4 Coalescence Time

The shuffling example we considered earlier presents us with a strange issue. In general, we do not wish to compare our random starting point, X, with the a sample from stationarity, $Y \in_u \Pi$. Instead, we want to begin with Y at a random starting point similar to X. We will define our coupling as follows:

Coupling:

$\hat{p}:\Omega^2 \to \Omega^2$ such that
$\sum_{y'\in\Omega} \hat{p}((x,y),(x',y')) = P(x,x')$ and
$\sum_{x'\in\Omega} \hat{p}((x,y),(x',y')) = P(y,y')$

This implies several things, the most important of which is that the marginal from $x \to x'$ and from $y \to y'$ is faithful to the original markov chain. Moreover, this implies that when $x = y$, $x' = y'$. In other words, steps of the coupled Markov Chains are still faithful to the original Markov Chain, and once two states are coupled, they remain coupled at all future steps.

We wish to have some intuition about the amount of time that it takes for two chains to couple so we have a valid understanding of the mixing times.

Definition: Coalescence time

A function t such that $\forall x, y \in \Omega, \epsilon > 0$
$Pr[x_{t(\epsilon)} \neq y_{t(\epsilon)} | x_0 = x, y_0 = y] \leq \epsilon$
This implies that even in the worst possible starting positions x, y, the coalescence time is bounded by this error probability.

This implies that even in the worst possible starting positions x, y, the coalescence time is bounded by this error probability. Understanding these worst case bounds is critical to proving the mixing times for our Markov Chains in all cases.

Theorem:

$\tau(\epsilon) < t(\epsilon)$

Proof:
Recall that $||p^t(x_0) - \Pi||_{tv} = \frac{1}{2}\sum_{w\in\Omega} |p^t(x,w) - \Pi(w)| = \max_{A\subseteq\Omega} \sum_{w\in A} [p^t(x,w) - \Pi(w)]$.

Let $A \subseteq \Omega$ be any subset at $t(\epsilon)$.

$$p^t(x,A) = \sum_{w\in A} [p^t(x,w)] = Pr[x_t \in A]$$
$$\geq Pr[x_t = y_t \wedge y_t \in A]$$
$$= 1 - Pr[x_t \neq y_t \vee y_t \notin A]$$
$$\geq 1 - Pr[x_t \neq y_t] + Pr[y_t \notin A]$$

We know

$$Pr[x_{t(e)} \neq y_{t(e)} | x_0 = x \wedge y_0 = y] \leq \epsilon \Rightarrow Pr[x_{t(e)} \neq y_{t(e)} | x_0 = x \wedge y_0 \in_\pi \Omega] \leq \sum_{y\in\Omega} \Pi(y) \cdot \epsilon = \epsilon$$

From this we can conclude

$$1 - Pr[x_t \neq y_t] + Pr[y_t \notin A] \geq Pr(y_t \in A) - \epsilon = \Pi(A) - \epsilon \Rightarrow ||P^t, \Pi|| < \epsilon \Rightarrow \tau(\epsilon) \leq t(\epsilon) \quad \square$$

In many cases, this proof will have a $log\epsilon$ factor associated with it because often the definition of coalescance time is about expected time for X and Y to meet. In this case we were bounding the probability by an ϵ, so this factor is not present.

4.1 Walk on the hypercube

To demonstrate coupling with a simple example, consider walks on the hypercube $b = \{0,1\}^n$ We will define our Markov Chain as follows:

Markov Chain:

1. Pick an index $i \in_u \{1, ..., n\}$
2. Pick $j \in 0, 1$ 3. Set $b_i = j$. This implies that there is a $\frac{1}{2}$ probability of self-loop.

We will show, using coupling, that $\tau(\epsilon) = O(n \log n)$.

Start with arbitrary $x = (x_1, ..., x_n), y = (y_1, ..., y_n)$. Instead of blindly hoping that our two Markov Chains coalesce, we will pick i, j to be the same for both x, y. Using this coupling, it is clear that once the two chains agree on a bit, they will always agree on that bit, and because of this fact we can assume that the chain converges to Π in coupon collector's time, $O(n \log n)$. It is important to note that if we flipped bit i instead of picking j as the chosen result for bit i, then the chains would never meet because any differences would always be different and the chains would always be at constant distance from one another.

5 K-Colorings

Our original motivation for looking at coupling was to determine when we can efficiently sample from k-colorings in a graph. We have shown that analyzing the coupling time gives a good bound on the mixing time of a Markov Chain, and we have given some strategies about how to couple two chains. Recall our definition of the Markov Chain:

Markov Chain: M_{col}

1. Given $G = (V, E)$ at a valid coloring
2. Select $v \in_u V$
3. Select color $c \in_u 1, 2, ..., k-1, k$
4. Recolor v with c if it produces a valid k-coloring. Otherwise keep v colored as it is.

We will now define a coupling to try to bound the mixing time with different values of k.

5.1 $k > 4\Delta$ is rapidly mixing

When $k \geq 4\Delta$, we will use the following coupling: first, choose a node to change the color of in both graphs. Second, choose a color to set both graphs to. Third, recolor the nodes as allowed by the Markov Chain. The probability of going from any one state to any other state is $\frac{1}{nk}$, and the chain is ergodic. We hope we can prove this chain to be rapidly mixing.

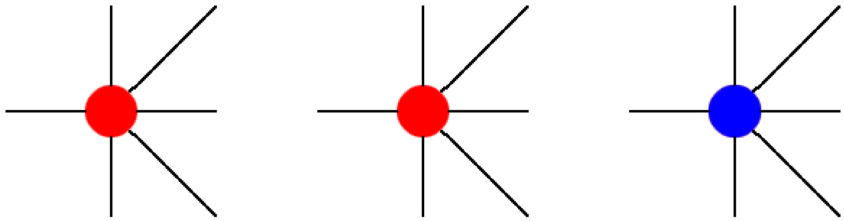
Theorem: M_{col} is rapidly mixing when $k > 4\Delta$.

Proof: Given two separate colorings, σ, τ on G, let d be the number of vertices where the colorings differ. We claim that as long as the distance between σ and τ is nonzero, the expected change in distance is negative (i.e. the chance of the two chains getting closer is greater than their chance of getting further apart). Decompose moves into ones where the two chains are brought closer and one where the chains get futher apart.

The good moves are the ones where v is chosen to be a point where σ and τ differ and a good color is picked. The probability of any particular good move being chosen is $\frac{k-2\Delta}{nk}$, and since there are d places at which the chains differ, the probability of choosing a good move is $d \cdot \frac{k-2\Delta}{nk}$.

Bad moves for this coupling are ones in which v is chosen to be a place where σ and τ agree and the process forces them to disagree. Consider a node, v chosen from G which is colored blue in both σ and τ. In σ, v's neighbors have colors $\{b_1, ..., b_n\}$ and in τ v's neighbors have colors $\{c_1, ..., c_n\}$ with $\forall i, j\ b_i \neq c_j$.

There are at most 2Δ bad color choices, assuming that v has the maximum degree of the graph, Δ, and all the neighbors in σ and τ are distinct. This gives us a probability of bad moves at any vertex $\leq \frac{2\Delta}{nk}$, so the probability of bad moves across the graph $P[bad] \leq \frac{2d\Delta}{nk}$.



(a) (b) (c)

Figure 4: In a), σ and τ both have $c(v)$ red. σ has a blue neighbor, τ does not. In b), σ cannot change at v because blue was chosen and it has a blue neighbor. In c), τ changes because it does not have a blue neighbor

Given these probabilities, we can now show that the chain is in expectation rapidly mixing.

$$E[\Delta d] = -1(\frac{dk-2\Delta d}{nk}) + \frac{2d\Delta}{nk} = \frac{4d\Delta - dk}{nk}.$$
$$\text{Let } A = \frac{k-4\Delta}{k}.$$
$$\frac{4d\Delta - dk}{nk} = -dA.$$
$$E[d] \leq n(1-A)^t \leq ne^{\frac{-At}{n}}$$
$$\Rightarrow ||P_t - \Pi||_{TV} \leq E[d] \leq ne^{\frac{-At}{n}} \Rightarrow \tau_\epsilon \leq \frac{1}{A} n \ln(n\epsilon^{-1})$$

This is sufficient to show the rapid mixing of the chain.

5.2 $k > 3\Delta$ is rapidly mixing

To demonstrate that the Markov Chain is rapidly mixing when $k > 3\Delta$, we will need to introduce the concept of *path coupling*. Path coupling allows us to demonstrate that as long as two states can be represented as having unit distances between them, and we can prove that these unit distances in expectation contract, then the chain as a whole contracts.

More formally, consider all pairs σ, τ such that their hamming distance, $d_h(\sigma, \tau) = 1 \Rightarrow E[\Delta(d_h(\sigma, \tau)) = 1] < 0$. Our claim is that if this holds, and we can define all pairs as having a path consisting of these pairs at distance 1, then their distance in expectation also shrinks.

Consider a pair, $x, y \in \Omega$. Define $dist(x, y) = \sum_{i=0}^{d-1} dist(z_i, z_{i+1})$ where $x = z_0, z_1, ..., z_d = y$ and $d(z_i, z_{i+1}) = 1$. Our proof is very similar to before, only this time we only have to consider entire graphs which differ by one point. Realizing this, we can enhance the probability that our chains are in expectation decreasing. We can show that the rate of good moves is $\frac{k-\Delta}{nk}$. To show this, consider two chains σ, τ which differ at only one point. Let v be this point. When the Markov Chain "selects" v, assuming v has degree Δ, there are then Δ points which are invalid selections. There are therefore $k - \Delta$ valid selections. As the two chains differ by only one point, the probability of good moves is $\frac{k-\Delta}{nk}$.

As before, there are at most 2Δ bad moves in the graph when distance is 1. To acheive a bad move, you must pick a neighbor of the node that is different between σ and τ. Let us call this node w, and w is any neighbor of v. For each w, there are 2 bad choices, choosing the color *blue* or choosing the color *red*. Either of these choices will cause the σ and τ to become further apart because the change is valid in one of $\sigma(w)$

or $\tau(w)$, but not in the other. This implies that the probability of bad moves is $\frac{2\Delta}{nk}$.

Following the logic from the proof in the previous section, the mixing time is $\tau_\epsilon \leq \frac{1}{A} n \ln(n\epsilon^{-1})$ where $A = \frac{k-3\Delta}{d}$.

5.3 $k > 2\Delta$ is rapidly mixing

A method Jerrum uses allows us to prove that when $k > 2\Delta$, the chain is rapidly mixing. We essentially use the same mechanism for the $k > 3\Delta$ proof, in that we use path coupling's strengths to prove our Markov Chain's mixing rate. The additional change that allows to demonstrate fast mixing of $k > 2\Delta$ is a change to the coupling. We use the fact that the chains differ at one point to reduce the chance of error in the chain.

In this case, the only place error can be introduced when the neighbors of the uncoupled node, v, are selected. Let one of these nodes be w. Suppose v is colored blue in σ and colored red in τ. Suppose w is, without loss of generality, colored yellow. The two bad choices for color for w are red and blue, so the novel idea is to alternate selection of red and blue (or whichever colors are contained in v) between σ and τ. For instance, if blue is chosen for w, place blue as the color for $w \in \sigma$ and red as the color for $w \in \tau$. If red is chosen, do the opposite.

This presents two cases: Case 1) w has a color unique from v in both graphs. In this case, the distance between the two increases to 2. Case 2) w is equal in color to v in both graphs, and w does not get recolored. In this case there are the same number of good moves as in the $k > 3\Delta$ proof, but there is only one bad choice for each neighboring vertex. Perform the proof as before, but since there is only one bad move now, $\tau_\epsilon \leq \frac{1}{A} n \ln(n\epsilon^{-1})$ where $A = \frac{k-2\Delta}{d}$

6 Conclusion

It is important to remember that coupling is simply a thought experiment. When mixing Markov Chains, one does not execute two simultaneously and then output the result when the two are identical; instead it demonstrates provable bounds on when the Markov Chain has reached an arbitrarily close point to the stationary distribution.

Additionally, Mark Jerrum uses a different algorithm for performing the algorithms. His algorithm involves removing the color from a vertex and replacing the color instead of our Markov Chain, and his analysis is different and worth reading. Jerrum is also able to show that the chain is rapidly mixing for $k = 2\Delta$.