CSE 525: Randomized Algorithms **Spring 2023**

Lecture 17: Coupling Cont'd

Lecturer: Shayan Oveis Gharan *May 18th*

Disclaimer: *These notes have not been subjected to the usual scrutiny reserved for formal publications.*

17.1 Simple Random Walk on a Hypercube $\{0,1\}^n$

Consider the n dimensional hypercube with 2^n vertices, where every vertex is labelled with an n bit string. Two vertices are adjacent if their n bit strings differ in exactly one bit. Consider the following Markov chain.

i) At any vertex x, with probability 1/2 do nothing (self-loop)

ii) Otherwise, pick a uniformly random coordinate and flip it.

Equivalently, we can consider the following Markov chain: At any vertex x choose a uniformly random coordinate i and substitute it with a uniformly random bit $b \in \{0,1\}$.

Having the second description, considering the following coupling between X_t, Y_t: X_t and Y_t choose the same coordinate i and the same bit b.

Observe that this is a valid coupling, because each X_t, Y_t is following the same random walk. Furthermore, observe that $T_{X,Y}$ is at most the time by which each coordinate is chosen at least once. This is because once we choose a coordinate i from that moment X_t, Y_t will agree on that coordinate.

So, it is enough to find the expected time that it takes to choose each coordinate at least once, and then we can use the Markov's inequality.

This problem is known as the *coupon collector* problem: At each time step, the collector gets one out of n coupons uniformly at random. His aim is to continue till he has seen every coupon at least once. Let T_k be the time it takes to see k coupons assuming he has already seen $k-1$ coupons. Obviously $T_1 = 1$, and $\sum_{i=1}^n T_i$ is the time to take all coupons. It turns out that for each k, $\mathbb{E}[T_{k+1}] = \frac{1}{1-k/n}$. This is because he has already seen k coupons; so the next coupon will be new only with probability $1-k/n$. This is a geometric random variable, so its mean is $\frac{1}{1-k/n}$. By linearity of expectation we get

$$\mathbb{E}[T_1 + \cdots + T_n] = \sum_{k=1}^n \frac{1}{1-(k-1)/n} = \sum_{k=1}^n \frac{n}{n-k+1} = nH_n.$$

In general, it is not hard to see that the coupon collector time is highly concentrated around its expectation, in the sense that

$$\mathbb{P}[T_1 + \cdots + T_n > n\ln n + cn] \le e^{-c}.$$

We prove a weaker bound: For any i, the probability that coupon i is not collected by time $t = cn\ln n$ is at most $(1-1/n)^t \le n^{-c}$. So, by union bound, the probability that all coupons are not collected by time $2n\ln n$ is at most $1/n$.

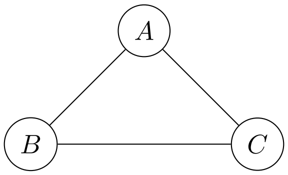
$$\mathbb{P}[T_1 + \cdots + T_n > cn\ln n] \le n^{-c+1}.$$

17.2 Graph Coloring

Let G be a graph with maximum degree Δ. Obviously, any such graph has a coloring with $q = \Delta + 1$ colors: Greedily color every vertex with a color with is not used by any of its neighbors. There is a well-known Brooks theorem which says that any graph has a coloring with Δ colors unless it has a $\Delta + 1$ clique or $\Delta = 2$ and it has an odd cycle. For $q < \Delta$ even the decision problem is NP-hard.

So, let us focus on the case the case where $q \geq \Delta + 1$. We want to see if the Metropolis chain mixes rapidly. Note that the chain is very simple to describe: At each time we choose a uniformly random vertex and a uniformly random color and we color that vertex with that color if possible.

It turns out that if $q = \Delta + 1$, then the above chain is not irreducible as shown below: Even though the graph has 6 proper coloring we cannot move out of the above state. In other words, the Metropolis chain is not irreducible in this case.



First, we argue that the Metropolis chain is irreducible if $q \geq \Delta + 2$. It is enough to show that we can go from any proper coloring to any fixed proper coloring, because the chain is reversible. We show that if $q \geq \Delta + 2$ Consider a vertex v and suppose it has a color c in the target coloring. If we can re-color v wi

The following is a major open problem in the field of counting/sampling:

Conjecture 17.1. *The Metropolis chain mixes in time $O(n \log n)$ if $q \geq \Delta + 2$.*

It turns out that even though there have been a huge amount of efforts we are still very far from prove the above conjecture. In this lecture and the next we will discuss several of the techniques to bound mixing time for different ranges of q.

Theorem 17.2. *For any $q \geq 4\Delta + 1$ the mixing time is $O(n \log n)$.*

Let us construct a coupling between two copies of the chain. Let X_t, Y_t be two colorings of graph G. Consider the natural metric where $d(X_t, Y_t)$ is equal to the number of vertices with different colors. We construct a coupling of X_{t+1}, Y_{t+1} such that

$$\mathbb{E}\left[d(X_{t+1}, Y_{t+1})\right] \leq \mathbb{E}\left[d(X_t, Y_t)\right] \left(1 - \frac{q - 4\Delta}{qn}\right). \tag{17.1}$$

Note that once we have the above inequality by a repeated application we can show that

$$\mathbb{E}\left[d(X_t, Y_t)\right] \leq d(X_0, Y_0) \left(1 - \frac{q - 4\Delta}{qn}\right)^t \leq e^{-\frac{(q-4\Delta)t}{qn}}.$$

Now, note that $d(X_0, Y_0) \leq n$ because in the worst case all vertices have different colors. Therefore, for $t = 8\Delta n \log n$, we have $\mathbb{E}\left[d(X_t, Y_t)\right] \leq 1/n$ which implies that $\mathbb{P}\left[X_t \neq Y_t\right] \leq 1/n$.

So, it remains to prove (17.1). We use the natural coloring, i.e., we use the same coin for both chains. Both X_t, Y_t will choose the same vertex v and try to color it with c if possible. Now, let us consider under what circumstances the distance of X_t, Y_t will decrease or increase.

- **Good Moves:** This is a move in which the distance between X_t, Y_t decreases. Suppose we choose a vertex of disagreement, say v, and we choose a color that belongs to none of the neighbors of v in X_t, Y_t. In such a case the distance decreases by 1. There are $d(X_t, Y_t)$ such vertices. Note that v has at most Δ neighbors so at most 2Δ colors are used on the neighbors of v. So, if the color c is different from these 2Δ, v will have the same color in X_{t+1}, Y_{t+1}, so $d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) - 1$. The probability of this event is $\frac{d(X_t,Y_t)(q-2\Delta)}{qn}$.

- **Bad Moves:** This is a move in which the distance between X_t, Y_t increases. Now, suppose we choose a vertex of agreement, say v and a color c. A bad event is if c is an invalid color for v in exactly one of the chains. In such a case the distance increases by 1. Observe that such an event happens if v is a neighbor of a disagreement vertex, say u, and c is the color of u in one of the two chains. The disagreement vertices have at most $\Delta d_t(X_t, Y_t)$ neighbors, and for any such neighbors there are at 2 bad colors. Therefore,

$$\mathbb{P}\left[d(X_{t+1}, Y_{t+1}) = d(X_t, Y_t) + 1\right] \leq \frac{2\Delta d_t(X_t, Y_t)}{qn}$$

- **Neutral Moves:** In any other move the distance remains invariant.

Putting these together we have,

$$\mathbb{E}\left[d(X_{t+1}, Y_{t+1}) | X_t, Y_t\right] \leq d(X_t, Y_t) - \frac{d(X_t, Y_t)(q - 2\Delta)}{qn} + \frac{d(X_t, Y_t)2\Delta}{qn}$$

Taking expectation from both sides proves (17.1).

17.3 Path Coupling

Path coupling is a technique that simplifies the construction of couplings, when the underlying metric is a shortest path on a graph. The main consequence is that while in a normal coupling we have to couple all pairs of states $x, y \in \Omega$, the path-coupling technique shows that it is enough to construct a coupling only between *neighboring states.* As we will see this simplifies the task of coupling and in many cases allows us to prove stronger mixing bounds.

Definition 17.3 (pre-metric)**.** *A pre-metric on Ω is a weighted connected undirected graph such that for any edge (x, y), the edge is a shortest path between x and y. In other words, the length of every other path from x to y is at least the weight of the edge (x, y).*

Note that we can naturally extend a pre-metric to a metric by considering the shortest path metric. The path-coupling technique says that when defining the coupling it is enough to define it only for adjacent states in the pre-metric.

Theorem 17.4 ([?])**.** *Suppose there exists a coupling $(X, Y) \to (X', Y')$ defined for all adjacent pair of states in the pre-metric such that for all adjacent X, Y*

$$\mathbb{E}\left[d(X', Y') | X, Y\right] \leq (1 - \alpha) d(X, Y)$$

where d is the shortest path metric. Then, this coupling can be extended to a coupling between all pair of states that also satisfy the above inequality.

Proof. Fix a pair of states (X,Y) and let us define a coupling between X,Y. Consider the shortest path from X to Y,

$$X = Z_0, Z_1, \ldots, Z_\ell = Y.$$

We construct a coupling by composing couplings of all pairs (Z_i, Z_{i+1}) along this path.

- First we map (Z_0, Z_1) to (Z_0', Z_1') according to the promised coupling.
- Iteratively, for each $i \geq 1$, we map (Z_i, Z_{i+1}) to (Z_i', Z_{i+1}') according to the promised coupling conditioned on the Z_i' chosen in the previous step of the iteration.

Observe that the above is a valid coupling. In particular, let us just look at the Z_1, Z_2 pair. First, observe that Z_1' is chosen in the very first step, so Z_1' is obtained by running the Markov chain from Z_1 for one step. Conditioned on the coin tosses that sends $Z_1 \to Z_1'$, we sample Z_2'. Note that it could be that conditioned on those coin tosses Z_2' is deterministically defined given Z_2. This is a valid coupling because in principal one can run any coupling this way; first run an honest step of the first chain and then by looking at the result of the coin tosses (and perhaps some additional coin tosses) move the second chain.

It follows that

$$\begin{aligned}
\mathbb{E}\left[d(X', Y')\right] &\leq \sum_{i=0}^{\ell-1} \mathbb{E}\left[d(Z_i', Z_{i+1}')\right] \\
&\leq \sum_{i=0}^{\ell-1} (1-\alpha) d(Z_i, Z_{i+1}) \\
&= (1-\alpha) d(X,Y).
\end{aligned}$$

The first inequality is because d is a shortest path distance, the second inequality follows by the promise of the theorem and the last equality follows by the definition of a pre-metric. □

Note that once we have the above theorem, similar to the coloring example, we can argue that the mixing time is $O(\frac{1}{\alpha}\log D)$ where D is the diameter of the shortest path metric.

17.4 Coloring using Path Coupling

In this section we prove the following theorem due to Jerrum [**?**].

Theorem 17.5. *If $q \geq 2\Delta + 1$, then the Metropolis rule mixes in time $O(n \log n)$.*

Let us start by defining a pre-metric. Ideally, similar to the previous proof we would like to let $d(X,Y)$ be the number of vertices that are colored differently. But, we have to make sure that this is a valid shortest path metric. Note that in general we may not be able to find a path from X to Y by correcting the color of every disagreement vertex. This is because along the way we may find non-proper colorings of G.

The idea is to work with an enlarged state space. Let the state space of the Markov chain be all possible q^n colorings of G. Observe that given any state (with possibly a non-proper coloring) the Metropolis chain only assigns valid colorings to vertices. Therefore, after a number of steps we will reach a state with a proper coloring. This means that all states with non-proper coloring are *transient* states of this chain, meaning that they have probability 0 in the stationary distribution. So, the stationary distribution does not change and a coupling proof for the chain with extended state space gives an upper bound on the mixing time.

Now, let X, Y be two (not necessarily proper) coloring of G that differ in exactly one vertex v. Let $c_X(v), c_Y(v)$ be the color of v in X, Y respectively. Consider the following coupling: Pick the same vertex u in both chains, and if

i) $u = v$**:** both chains pick the same color c and color u with c (if possible).

ii) $u \in N(v)$**:** The X chain picks a color c. If $c = c_X(v)$, then Y chain picks $c_Y(v)$. If $c = c_Y(v)$, then Y chain picks $c_X(v)$, and otherwise the Y chain picks the color c.

iii) $u \notin N(v)$**:** In this case both chains pick the same color c.

Observe that this is clearly a valid coupling.

Note that in case (iii) the distance between X, Y remains invariant. This is because u and all neighbors of u have the same color in both chains; so this is the neutral move.

In the case (i) the distance between X, Y decreases if c is different from the color of all neighbors of v. Note that since v is the only point of disagreement, there are $q - \Delta$ good colors. So, with probability $\frac{q-\Delta}{qn}$ the distance decreases by 1.

Now, let us analyze case (ii): if $c = c_X(v)$, then both options are trying to color u with invalid colors; so both actions are rejected and the distance remains invariant. If $c = c_Y(v)$ in the X chain, X tries to color u with $c_Y(v)$ and Y tries to color u with $c_X(v)$. This can indeed happen and it implies that u will have different colors. In all other cases, i.e., if $c \neq c_X(v), c_Y(v)$ both chains will do exactly the same: Either both accept or both reject c, and the distance remains invariant. This implies that

$$\mathbb{P}\left[d(X', Y') = 2 | X, Y\right] \leq \frac{\Delta}{qn}.$$

It follows that

$$\mathbb{E}\left[d(X', Y') | X, Y\right] = 1 - \frac{q - \Delta}{qn} + \frac{\Delta}{qn} = 1 - \frac{q - 2\Delta}{qn} \leq 1 - \frac{1}{qn}$$

where the last inequality follows by assuming $q \geq 2\Delta + 1$. Therefore, the chain mixes in time $O(qnlogn)$.

17.5 Dirichlet Form

Consider a Markov chain with Kernel K on state space Ω and stationary π. For two functions $f, g \in \Omega \to \mathbb{R}$ define

$$\langle f, g \rangle_\pi = \sum_x f(x) g(x) \pi(x).$$

Fact 17.6. *for any pair of functions $f, g \in \Omega \to \mathbb{R}$,*

$$\langle Kf, g \rangle_\pi = \langle f, Kg \rangle_\pi.$$

Proof. We write

$$\begin{aligned}\langle Kf, g \rangle_\pi &= \sum_x \pi(x) g(x) \sum_y K(x, y) f(y) = \sum_x g(x) \sum_y K(y, x) \pi(y) f(y) \\ &= \sum_y f(y) \pi(y) \sum_x P(y, x) g(x) = RHS.\end{aligned}$$

□

So the Markov kernel K is self-adjoint, so it has real eigenvalues and we can apply the spectral theorem.

17.6 Mixing Time via Spectral Gap

Definition 17.7 (Variance)**.** *For a function $f : \Omega \to \mathbb{R}$ define*

$$\mathrm{Var}(f) = \langle f - \mathbb{E}f, f - \mathbb{E}f\rangle_\pi = \langle f, f\rangle_\pi - 2\mathbb{E}f\langle f, \mathbf{1}\rangle_\pi + (\mathbb{E}f)^2 = \langle f, f\rangle_\pi - (\mathbb{E}f)^2$$

where as usual we used $\|\mathbf{1}\| = 1$ and $\mathbb{E}f = \langle f, \mathbf{1}\rangle_\pi$. Note that by definition $\mathrm{Var}(f) = \mathrm{Var}(f + \alpha 1)$ for any $\alpha \in \mathbb{R}$. Furthermore,

$$\mathrm{Var}(K(f + \alpha 1)) = \mathrm{Var}(Kf + \alpha 1) = \mathrm{Var}(Kf).$$

Lemma 17.8. *For any function $f : V \to \mathbb{R}$, $\mathrm{Var}(f) - \mathrm{Var}(Pf) = \langle (I - K^2)f, f\rangle_\pi$.*

Proof. First, by definition of variance we can shift f and assume without loss of generality that $\mathbb{E}f = \langle f, \mathbf{1}\rangle_\pi = 0$. Also note that $\mathbb{E}Kf = \langle \mathbf{1}, Kf\rangle_\pi = 0$. Therefore,

$$\mathrm{Var}(f) - \mathrm{Var}(Kf) = \langle f, f\rangle_\pi - \langle Kf, Kf\rangle_\pi = \langle (I - K^2)f, f\rangle_\pi$$

The RHS is also known as the Dirichlet form of f with respect to K. The operator $I - K$ is called the normalized Laplacian of f. □

Equivalently, the Dirichlet form can be written as

$$\langle (I - K)f, f\rangle_\pi = \sum_{x,y} \pi(x)K(x, y)(f(x) - f(y))^2 \geq 0$$

The above lemma in particular implies that $\mathrm{Var}(Kf) \leq \mathrm{Var}(f)$. Let

$$\lambda_2(I - K^2) = \min_f \frac{\langle (I - K^2)f, f\rangle_\pi}{\mathrm{Var}(f)}$$

This quantity is the second smallest eigenvalue of $I - K^2$ which can also be seen as difference of 1 and the square of the second largest eigenvalue of K in absolute value. Then, we have So, equivalently,

$$\lambda_2(K^2) \geq \frac{\mathrm{Var}(Kf)}{\mathrm{Var}(f)}.$$

Applying this repeatedly, we get

Corollary 17.9. *For any function $f : V \to \mathbb{R}$,*

$$\mathrm{Var}(K^t f) \leq \lambda_2(K^2)^t\, \mathrm{Var}(f).$$

Lemma 17.10. *For any $x \in Omega$, and any $\epsilon > 0$ the walk started at x satisfies:*

$$\sum_y |K^t(x, y) - \pi(y)| \leq \epsilon$$

as long as $t \geq \frac{\log \epsilon^{-1} \ln \pi(x)^{-1}}{1 - \lambda_2(K^2)}$. In fact the same proof also bounds the L-2 mixing.

Proof. Let $f = \mathbf{1}_x/\pi(x)$, i.e., $f(y) = 0$ for $y \neq x$ and $1/\pi(x)$ otherwise. Note that $\mathbb{E}f = 1$.

For some t that we choose later we write

$$\begin{aligned}\mathrm{Var}(K^{2t}f) &= \langle K^t f, K^t f\rangle_\pi - (\mathbb{E}K^t f)^2 \\ &= \mathbb{E}_{y\sim\pi} K^t f(y) \cdot K^t f(y) - \langle K^t f, \mathbf{1}\rangle^2 \\ &\underset{\text{using } f=\mathbf{1}_x/\pi(x)}{=} \mathbb{E}_{y\sim\pi}\left(\frac{K^t(y,x)}{\pi x(i)}\right)^2 - (\mathbb{E}f)^2 \\ &\underset{\text{using reversibility}}{=} \mathbb{E}_{y\sim\pi}\left(\frac{K^t(x,y)}{\pi(y)}\right)^2 - 1 = \mathbb{E}_{y\sim\pi}\left(\frac{K^t(x,y)}{\pi(x)} - 1\right)^2\end{aligned}$$

The RHS is called the L-2 mixing for the random walk started at x.

To get to the L-1 mixing we can just Cauchy-Schwarz inequality. In particular,

$$\begin{aligned}\sum_y |K^t(x,y) - \pi(y)| = \mathbb{E}_{y\sim\pi}\left|\frac{K^t(x,y)}{\pi(y)} - 1\right| \leq \mathbb{E}_{y\sim\pi}\left(\frac{K^t(x,y)}{\pi(y)} - 1\right)^2 \leq \\ \leq \mathrm{Var}(K^t f) \underset{\textit{Corollary 17.9}}{\leq} \lambda_2(K^2)^t\,\mathrm{Var}(f) \leq \frac{\lambda_2(K^2)^t}{\pi(i)} \leq \epsilon\end{aligned}$$

Where to get the last inequality it is enough to let $t = \frac{\ln \epsilon^{-1} \ln \pi(i)^{-1}}{1-\lambda_2(K^2)}$. □

References