CS 761 : Randomized Algorithms , Spring 2023, Waterloo

Lecture 16 : Coupling

Coupling is an important technique in bounding the mixing time of Markov chains.

We will see the basic method and some examples.

Coupling

This is the most commonly used method in bounding the mixing time of Markov chains.

In L12, we mentioned that if two Markov chains reach the same state, then one cannot distinguish the two distributions afterwards because Markov chains don't remember the history.

Coupling is a method to make this argument formal.

Before stating the coupling method, let us first recall the definition of mixing time.

Variation distance and mixing time

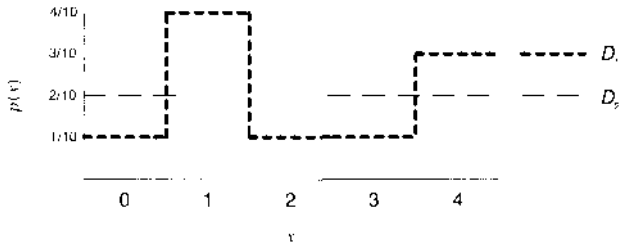
The total variation distance is to measure how close is two probability distributions.

Definition The total variation distance between two probability distributions $\vec{p}$ and $\vec{q}$ is defined as $\|\vec{p}-\vec{q}\| = \frac{1}{2}\sum_{x\in S}|p(x)-q(x)|$ where S is the state space.

The following is an equivalent characterization of the total variation distance.

Lemma For any $A \subseteq S$, let $p(A) = \sum_{x\in A} p(x)$. Then $\|\vec{p}-\vec{q}\| = \max_{A\subseteq S} |p(A)-q(A)|$.

Proof (by picture from MU)



The area where $\vec{p}$ is "above" $\vec{q}$ is equal to the area where $\vec{p}$ is "below" $\vec{q}$.

The sum of these two areas is equal to $\sum_{x\in S}|p(x)-q(x)|$.

Let A be the set of those values corresponding to the area where $\vec{p}$ is "above" $\vec{q}$.

Then $|p(A)-q(A)| = \frac{1}{2}\sum_{x\in S}|p(x)-q(x)| = \|\vec{p}-\vec{q}\|$. □

Mixing time is the time when the probability distribution is close to the stationary distribution, regardless of the initial distribution.

Definition (Mixing time) Let $\vec{\pi}$ be the stationary distribution.

Let $\vec{p}_x^{\,t}$ be the probability distribution after t steps starting at state x.

Define $\Delta_x(t) = \|\vec{p}_x^{\,t} - \vec{\pi}\|$ and $\Delta(t) = \max_{x\in S}\Delta_x(t)$.

Also define $\tau_x(\varepsilon) = \min\{t : o_x(t) \le \varepsilon\}$ and $\tau(\varepsilon) = \max_{x \in S} \tau_x(\varepsilon)$.

A Markov chain is rapidly mixing if $\tau(\varepsilon)$ is polynomial in $\ln(1/\varepsilon)$ and the number of states.

Now comes the important definition of coupling.

Definition 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:

① $\Pr(X_{t+1} = x' \mid Z_t = (x, y)) = \Pr(M_{t+1} = x' \mid M_t = x)$, and

② $\Pr(Y_{t+1} = y' \mid Z_t = (x, y)) = \Pr(M_{t+1} = y' \mid M_t = y)$.

In words, coupling is a joint random process of two Markov chains such that each Markov chain behaves exactly as the original Markov chain, even though the moves of the two chains could be dependent.

The power of the coupling method is to allow us to design a joint random process to:

① bring the two copies to the same state quickly, and

② to keep them in the same state by having the two chains make identical moves once they are in the same state.

Since both chains behave the same as the original Markov chain, we can then argue that the distributions of X_t and Y_t are the same after they merged.

An upper bound of the time of merging is an upper bound on the mixing time.

Lemma (Coupling lemma) Let $Z_t = (X_t, Y_t)$ be a coupling for a Markov chain M on state space S. Suppose that there exists a time T such that

$$\Pr(X_T \neq Y_t \mid X_0 = x, Y_0 = y) \le \varepsilon \quad \text{for every } x, y \in S.$$

Then $\tau(\varepsilon) \le T$.

Proof Consider the coupling where Y_0 is chosen according to the stationary distribution and X_0 takes on arbitrary value.

For the given T and ε and for any $A \subseteq S$, we have

$$\begin{aligned}
\Pr(X_T \in A) &\ge \Pr\big((X_T = Y_T) \cap (Y_T \in A)\big) \\
&= 1 - \Pr\big((X_T = Y_T) \cup (Y_T \notin A)\big) \\
&\ge 1 - \Pr(X_T = Y_T) - \Pr(Y_T \notin A) \qquad \text{// union bound} \\
&= \Pr(Y_T \in A) - \Pr(X_T \neq Y_T) \\
&\ge \Pr(Y_T \in A) - \varepsilon \qquad \text{// by assumption} \\
&= \pi(A) - \varepsilon.
\end{aligned}$$

Similarly, we can argue $\Pr(X_T \& A) \geq \pi(S-A) - \varepsilon$, which implies that $\Pr(X_T \in A) \leq \pi(A) + \varepsilon$.

It follows that $\max_{x,A} |p_x^t(A) - \pi(A)| \leq \varepsilon$. This implies that $\tau(\varepsilon) \leq T$. □

Stationary distribution

Now we can explain why any finite, irreducible, aperiodic Markov chain will converge to the same distribution.

The coupling is easy: before they met they run independently, and after they met they always make the same moves. It is easy to see that both chains behave as the original one.

Recall that for any finite, irreducible and aperiodic Markov chain, there exists T such that $P_{i,j}^t > 0$ for all i,j and for all $t \geq T$. Let $\delta = \min_{i,j} \{P_{i,j}^T\}$.

Then, with probability at least δ, the two Markov chains will merge after T steps.

So, after $t = kT$ steps, the two Markov chains do not merge with probability $\leq (1-\delta)^k$.

This tends to zero when t tends to infinity, and so any initial distribution will converge to the stationary distribution.

Shuffling cards

Consider the following method for shuffling n cards.

In each step, we pick a random card and put it on the top of the deck.

How good is this shuffling process?

This is a Markov chain whose state space is the set of all permutations of the n cards.

It is not difficult to check that this chain is irreducible and aperiodic, and also the stationary distribution is the uniform distribution.

To bound the mixing time, we consider the following coupling:

- Choose a position j uniformly at random from 1 to n, and then move the j-th card to the top in the first chain. Denote the card value by C.
- Move the card with value C to the top in the second chain.

This is a valid coupling, since the probability that a card is moved to top is $1/n$.

With this coupling, once a card C is moved to the top, then it will be in the same position in both chains.

So, the two Markov chains will be coupled if every card has been moved to top at least once.

This is just the coupon collector problem.

So, after $n \ln n + n \ln(\frac{1}{\varepsilon})$ steps, the probability that the two chains haven't coupled is at most ε.

Random walks on the hypercube

Recall that an n-dimensional hypercube is a graph with 2^n vertices, where each vertex corresponds to an n-bit string, and two vertices have an edge iff the two bit strings differ in one bit.

Starting from an arbitrary vertex, we do a random walk by choosing a random position and set its value to one with prob $1/2$ and zero with prob $1/2$.

It is not difficult to check that this chain is irreducible and aperiodic, and also the stationary distribution is the uniform distribution.

To bound the mixing time, we consider a simple coupling: both chains choose the same position and set the same value. It is clear that it is a valid coupling.

With this coupling, once the i-th coordinate has been chosen, it will always be the same in two chains.

Once every coordinate has been chosen at least once, the two chains coupled.

Again, this is the coupon collector problem, and $\tau(\varepsilon) \le n \ln(n/\varepsilon)$.

Random spanning trees

For simplicity, let G be a d-regular undirected graph.

Define $\overleftrightarrow{G}$ be the directed graph such that each edge uv in G is replaced by two arcs uv and vu.

An arborescence rooted at r is a subset of $|V|-1$ edges where every vertex $v \neq r$ has exactly one outgoing edge, and there is a directed path from any vertex to r.

Note that there is a bijection between the set of spanning trees and the set of arborescences rooted at r.

The following Markov chain is defined on the set of arborescences.

① Start on an arborescence rooted at r. Start at vertex r.

② Let the current vertex be u. Choose a random neighbor v of u.

Set $X_{t+1} = X_t + uv - vw$, where vw is the unique outgoing edge of v in X_t.

Note that vertex v becomes the root, and the current vertex is always the root of the arborescence.

Set $X_{t+1} = X_t$ with probability $1/2$, i.e. lazy random walk.

It is not difficult to check that this chain is irreducible and aperiodic, and also the stationary distribution is the uniform distribution.

To bound the mixing time, we consider the coupling that the two chains always make the same move.

In particular, the current vertex of the two chains will always be the same.

Once two chains reach a vertex x, then the outgoing edge of x will always be the same.

Therefore, once every vertex is visited and returned to r, then the two arborescences agree completely and will be the same.

The expected time to visit every vertex at least once is the cover time.

By Markov's inequality, $\Pr(X_t \neq Y_t \mid X_0, Y_0) \leq \frac{1}{4}$ in $4T_{cover}$ steps.

Forget about the directions gives us an almost uniform random spanning tree.

Independent sets of fixed size

This example is more interesting as the distance between the two chains may increase or decrease.

Consider a Markov chain whose states are all independent sets of size exactly k in a graph:

① Choose a vertex v in X_t uniformly at random and a vertex $w \in V$ uniformly at random.

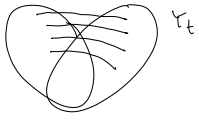
② if $w \notin X_t$ and $X_t - v + w$ is independent, then $X_{t+1} = X_t - v + w$; otherwise $X_{t+1} = X_t$.

Let n be the number of vertices and Δ be the maximum degree of any vertex.

We will show that this Markov chain is rapidly mixing if $k \leq n/(3\Delta+3)$.

It is not difficult to check that this chain is irreducible and aperiodic, and also the stationary distribution is the uniform distribution.

The coupling will require an arbitrary bijection M between the vertices of $X_t - Y_t$ and the vertices in $Y_t - X_t$.



For the first chain, choose a random $v \in X_t$ and a random $w \in V$ and apply the move.

So, the first Markov chain behaves as the original one.

For the second chain, if $v \in Y_t$ then use the same pair of vertices v and w and make the same move.

Otherwise, if $v \notin Y_t$, make the move with $M(v)$ and w.

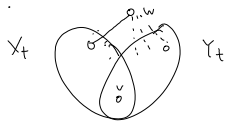
This also behaves the same as the original Markov chain, since each pair of vertices with $v \in Y_t$ and $w \in V$ is chosen with probability $1/kn$.

Let $d_t = |X_t - Y_t|$ measures the difference between X_t and Y_t. Clearly d_t can change by at most one.

We will show that d_t is more likely to decrease than to increase.

In order for $d_{t+1} = d_t + 1$. It must be the case that $v \in X_t \cap Y_t$, and that w is adjacent to some vertex in $X_t - Y_t$ but not adjacent to any vertex in $Y_t - X_t$, or vice versa.

In this case, a move is made in Y but not in X.



Thus, w must be a vertex or a neighbor of a vertex in the set $(X_t - Y_t) \cup (Y_t - X_t)$.

It follows that $\Pr(d_{t+1} = d_t + 1 \mid d_t > 0) \leq \frac{k - d_t}{} \cdot \frac{2d_t(\Delta+1)}{}$

Thus, w must be a vertex or a neighbor of a vertex in the set $(X_t - Y_t) \cup (Y_t - X_t)$.

It follows that $\Pr(d_{t+1} = d_t + 1 \mid d_t > 0) \le \underbrace{\frac{k - d_t}{k}}_{\text{choosing } v \text{ in } X_t \cap Y_t} \cdot \underbrace{\frac{2d_t(\Delta+1)}{n}}_{\text{choosing } w \text{ in } N(X_t \cup Y_t)}$.

Similarly, for $d_{t+1} = d_t - 1$, it is sufficient if $v \notin Y_t$ and w is neither a vertex nor a neighbor of a vertex in $X_t \cup Y_t - \{v, M(v)\}$.

Note that $|X_t \cup Y_t| = k + d_t$, and thus

$$\Pr(d_{t+1} = d_t - 1 \mid d_t > 0) \ge \frac{d_t}{k} \cdot \frac{n - (k + d_t - 2)(\Delta+1)}{n}$$

$$\begin{aligned} \text{So, } E[d_{t+1} \mid d_t] &= \Pr(d_{t+1} = d_t + 1) \cdot (d_t + 1) + (d_t - 1) \cdot \Pr(d_{t+1} = d_t - 1) + d_t \cdot \Pr(d_{t+1} = d_t) \\ &\le d_t + \left(\frac{k - d_t}{k}\right)\left(\frac{2d_t(\Delta+1)}{n}\right) - \left(\frac{d_t}{k}\right)\left(\frac{n - (k + d - 2)(\Delta+1)}{n}\right) \\ &= d_t\left(1 - \frac{n - (3k - d_t - 2)(\Delta+1)}{kn}\right) \\ &\le d_t\left(1 - \frac{n - (3k-3)(\Delta+1)}{kn}\right) \end{aligned}$$

By induction, $E[d_t] \le d_0\left(1 - \frac{n - (3k-3)(\Delta+1)}{kn}\right)^t$

Since $d_0 \le k$ and d_t is a non-negative integer, we have

$$\Pr(d_t \ge 1) \le E[d_t] \le k\left(1 - \frac{n - (3k-3)(\Delta+1)}{kn}\right)^t \le k e^{-t(n - (3k-3)(\Delta+1))/kn}$$

Since $k \le n/3(\Delta+1)$, $n - (3k-3)(\Delta+1) = n - (k-1)3(\Delta+1) \ge 1$ and thus $\Pr(d_t \ge 1) \to 0$.

It is easy to verify that $\tau(\varepsilon) \le \frac{kn \ln(k\varepsilon^{-1})}{n - (3k-3)(\Delta+1)}$ which is polynomial in n and $\ln(\varepsilon^{-1})$.

So, the Markov chain is rapidly mixing.

References

Most of the materials are from chapter 11 of Mitzenmacher and Upfal.

The spanning tree example is from the lecture notes of Eric Vigoda.