Markov Chains and Coupling

In this class we will consider the problem of bounding the time taken by a Markov chain to reach the stationary distribution. We will do so using the *coupling technique*, which helps bound the distance between two distribution by reasoning about coupled random variables.

1 Distance to Stationary Distribution

Let P be an ergodic transition matrix, and let π be the stationary distribution. Let $x_0 \in \Omega$ be some starting point. In order to test convergence we would like to bound the following *total variation distance*:

$$d(t) \ := \ \max_{x \in \Omega} ||P^t(x, \cdot) - \pi||_{TV} \tag{1}$$

where the total variation distance between two distributions μ and ν is given by:

$$||\mu - \nu||_{TV} \ := \ \frac{1}{2} \sum_{x \in \Omega} |\mu(x) - \nu(x)| \tag{2}$$

Exercise: Prove that the total variation distance can be equivalently written as:

$$||\mu - \nu||_{TV} \ := \ \max_{A \subseteq \Omega} (\mu(A) - \nu(A)) \tag{3}$$

Let $\bar{d}(t)$ denote the variation distance between two Markov chain random variables $X_t \sim P^t(x, \cdot)$ and $Y_t \sim P^t(y, \cdot)$. That is:

$$\bar{d}(t) \ := \ \max_{x, y \in \Omega} ||P^t(x, \cdot) - P^t(y, \cdot)||_{TV} \tag{4}$$

We can show the following important claim:

Claim 1. $d(t) \leq \bar{d}(t) \leq 2d(t)$

PROOF: $\bar{d}(t) \leq 2d(t)$ is immediate from the triangle inequality for the total variation distance.

Proof of $d(t) \leq \bar{d}(t)$: Since π is the stationary distribution, for any set $A \subseteq \Omega$, we have $\pi(A) = \sum_{y \in \Omega} \pi(y) P^t(y, A)$. Therefore, we get

$$\begin{aligned}
||P^t(x, \cdot) - \pi||_{TV} &= \max_{A \subseteq \Omega} (P^t(x, A) - \pi(A)) \\
&= \max_{A \subseteq \Omega} \left[P^t(x, A) - \sum_{y \in \Omega} (\pi(y) P^t(y, A)) \right] \\
&= \max_{A \subseteq \Omega} \left[\sum_{y \in \Omega} \pi(y) (P^t(x, A) - P^t(y, A)) \right] \\
&\leq \sum_{y \in \Omega} \pi(y) \max_{A \subseteq \Omega} (P^t(x, A) - P^t(y, A)) \\
&\leq \max_{y \in \Omega} \max_{A \subseteq \Omega} (P^t(x, A) - P^t(y, A))
\end{aligned}$$

□

The above claim is important since it allows us to quantify the variation distance to the stationary distribution ($d(t)$) using the distance between two Markov chains ($\bar{d}(t)$) from the same transition matrix (within a factor of 2). Moreover, it allows us to do so without knowing what the stationary distribution. We will see how to bound $\bar{d}(t)$ in the rest of the class using coupling techniques.

2 Coupling

Coupling is a powerful technique that will help us bound the convergence rates of a Markov chain.

Definition 1. *Let X and Y be random variables with probability distributions μ and ν on Ω. A distribution ω on $\Omega \times \Omega$ is a coupling if*

$$\forall x \in \Omega, \qquad \sum_{y \in Omega} w(x,y) = \mu(x)$$
$$\forall x \in \Omega, \qquad \sum_{x \in Omega} w(x,y) = \nu(y)$$

2.1 Coupling Lemma

Lemma 1. *Consider a pair of distributions μ and ν over Ω.*

(a) For any coupling w of μ and ν, let (X, Y) w,

$$||\mu - \nu||_{TV} \leq P(X \neq Y)$$

(b) There always exists a coupling w s.t.,

$$||\mu - \nu||_{TV} = P(X \neq Y)$$

Proof of (a): For any valid coupling w,

$$\forall z, w(z,z) \leq \min(\mu(z), \nu(z)) \tag{5}$$

Therefore,

$$\begin{aligned} P(X \neq Y) &= 1 - P(X = Y) = 1 - \sum_z w(z,z) \\ &\geq \sum_z \mu(z) - \sum_z \min(\mu(z), \nu(z)) \\ &\geq \sum_{z:\mu(z)>\nu(z)} (\mu(z) - \nu(z)) \\ &= ||\mu - \nu||_{TV} \end{aligned}$$

Proof of (b): We are now going to construct a coupling w s.t. $P(X \neq Y) = ||\mu - \nu||_{TV}$.

First we fix the diagonal entries:

$$\forall z, w(z,z) = \min(\mu(z), \nu(z)$$

This ensures that $P(X \neq Y)$ indeed equals the total variation distance between the two distributions. We set the off diagonal entries as follow:

$$w(y,z) = \frac{(\mu(y) - w(y,y))(\nu(z) - w(z,z))}{1 - \sum_x w(x,x)}$$

We leave it as an exercise to verify that w is indeed a coupling. □

3 Coupling and Markov Chains

The key insight from the coupling lemma is that the total variation distance between two distributions μ and ν is bounded above by $P(X \neq Y)$ for any two random variables that are coupled with respect to μ and ν. This turns out to be very useful in the context of Markov chains. First, we know from Claim 1 that the variation distance to the stationary distribution at time t is bounded (within a factor of 2) by the variation distance between any two Markov chains with the same transition matrix at time t. Moreover, by choosing an appropriately couple pair of Markov chains, we can bound $||P^t(x,\cdot) - P^t(y,\cdot)||_{TV}$ by the probability $P(X^t \neq Y^t)$.

Using this coupling argument, we will next prove that an ergodic Markov chain always converges to a unique stationary distribution, and then show a bound on the time taken to convergence (also known as *mixing time*) for the problem of randomly sampling graph colorings.

4 Ergodicity Theorem

Theorem 1. *If P is irreducible and aperiodic, then there is a unique stationary distribution π such that*

$$\forall x, \lim_{t \to \infty} P^t(x, \cdot) = \pi$$

Proof: Consider two copies of the Markov chain X_t and Y_t, both following P. We create a coupling distribution as follows:

- If $X_t \neq Y_t$, then choose X_{t+1} and Y_{t+1} independently according to P.
- If $X_t = Y_t$, then choose $X_{t+1} \sim P$, and set $Y_{t+1} = X_{t+1}$.

From the coupling lemma we know that

$$\forall t, ||X^t - Y^t||_{TV} \leq P(X^t \neq Y^t)$$

Due to ergodicity, there exist $t^\star$ such that $\forall x, y,\ P^{t^\star}(x,y) > 0$. Therefore, there is some $\epsilon > 0$, such that for all initial states X_0, Y_0,

$$P(X^{t^\star} \neq Y^{t^\star} | X_0, Y_0) \leq 1 - \epsilon \tag{6}$$

Similarly, due to the Markovian property, we can say

$$P(X^{2t^\star} \neq Y^{2t^\star} | X^{t^\star} \neq Y^{t^\star}) \leq 1 - \epsilon \tag{7}$$

Also, due to the coupling, $X^{2t^\star} = Y^{2t^\star}$ implies $X^{t^\star} = Y^{t^\star}$. Therefore,

$$\begin{aligned} P(X^{2t^\star} \neq Y^{2t^\star} | X_0, Y_0) &= P(X^{t^\star} \neq Y^{t^\star} \wedge X^{2t^\star} \neq Y^{2t^\star} | X_0, Y_0) \\ &= P(X^{2t^\star} \neq Y^{2t^\star} | X^{t^\star} \neq Y^{t^\star}) P(X^{t^\star} \neq Y^{t^\star} | X_0, Y_0) \\ &\leq (1-\epsilon)^2 \end{aligned}$$

Hence for any integer $k > 0$, we have

$$P(X^{kt^\star} \neq Y^{kt^\star} | X_0, Y_0) \leq (1-\epsilon)^k \tag{8}$$

As $k \to \infty$, $P(X^{kt^\star} \neq Y^{kt^\star} | X_0, Y_0) \to 0$. Since X^t and Y^t are coupled such that once they are the same at time t, they are the same for all $t' > t$, we have

$$\lim_{t \to \infty} P(X^t \neq Y^t | X_0, Y_0) \to 0$$

From the coupling lemma, we have

$$||P^t(x, \cdot) - P^t(y, \cdot)||_{TV} \leq P(X^t \neq Y^t) \to 0, \text{ when } t \to 0$$

To verify that, $\sigma = \lim_{t \to \infty} P^t(x, \cdot)$ is the required stationary distribution, note that

$$\begin{aligned} \sum_x \sigma(x) P(x, y) &= \sum_x \lim_{t \to \infty} P^t(z, x) P(x, y) \ \forall z \\ &= \lim_{t \to \infty} P^{t+1}(z, y) = \sigma(y) \end{aligned}$$

This shows that $\sigma P = \sigma$. Also, σ is unique since $|| \lim_{t \to \infty} P^t(x, \cdot) - \lim_{t \to \infty} P^t(y, \cdot)||_{TV} \to 0$. □

5 Mixing Time

Recall the definition of $d(t)$.

$$d(t) = \max_x d_x(t) = \max_x ||P^t(x, \cdot) - \pi||_{TV}$$

We can show that $d(t)$ is non-decreasing in t.

Claim 2. *$d_x(t)$ is non-decreasing in t.*

Proof: Let X_0 be some $x \in \Omega$, and let Y_0 have the stationary distribution. Fix t. By the coupling lemma, there is a coupling and random variables $X^t \sim P^t(x, \cdot)$ and $Y^t \sim \pi$ such that

$$d_x(t) = ||P^t(x, \cdot) - \pi||_{TV} = P(X^t \neq Y^t)$$

Using this coupling, we define a coupling of the distributions of X^{t+1}, Y^{t+1} as follows:

- If $X^t = Y^t$, set $X^{t+1} = Y^{t+1}$.
- Else, let $X^t \to X^{t+1}$ and $Y^t \to Y^{t+1}$ independently.

Then we have,

$$d_x(t+1) = ||P^{t+1}(x, \cdot) - \pi||_{TV} \leq P(X^{t+1} \neq Y^{t+1}) \leq P(X^t \neq Y^t) = d_x(t)$$

The first inequality holds due to the coupling lemma, and the second inequality holds by construction of the coupling. □

Since $d(t)$ never decreases, we can define the mixing time $\tau(\epsilon)$ of a Markov chain as:

$$\tau(\epsilon) = \min_t \{d(t) \leq \epsilon\} \tag{9}$$