Approximate Counting and Mixing Time of Markov chains **Fall 2024**

Lecture 3: Coupling

Lecturer: Shayan Oveis Gharan *October 1st*

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

this lecture we will discuss coupling as an important tool in studying Markov chains. First, we will discuss the Heat-Bath chain as a general rule in designing Markov chains with a given stationary distribution, then we introduce coupling and see its applications.

3.1 The Heat-Bath Chain

Let us recall the Ising model. Suppose we have a spin system where the spin of each particle i is $\sigma_i \in \{+1, -1\}$ and the weight of a configuration is

$$\pi(\sigma) \propto e^{\beta \sum_{i \sim j} \sigma_i \sigma_j}.$$

where β corresponds to the temperature.

The idea in the Heat-Bath chain is that each time we "re-randomize" an element of the current configuration with respect to π conditioned on the rest of the configuration. For example, in a spin-system we re-randomize the spin of a particle conditioned on the spin of the rest of the particles. This chain is typically used to generate random samples from spin systems. It is also straightforward to study an implementation of these chains in distributed systems, see [SRO16] for a recent best paper award of ICML studying the behavior of the Heat-Bath chain in a distributed system.

Let us discuss the details of the Heat-Bath chain in the Ising model. Consider a configuration σ. Let us write

$$\begin{aligned} a(\sigma) &:= |\{i \sim j : \sigma_i = \sigma_j\}|, \\ d(\sigma) &:= |\{i \sim j : \sigma_i \neq \sigma_j\}|. \end{aligned}$$

We can write

$$\pi(\sigma) = e^{\beta(a(\sigma) - d(\sigma))} = \frac{e^{2\beta a(\sigma)}}{e^{\beta |E|}}.$$

So, we can just assume that the weight of each configuration σ is $e^{2\beta a(\sigma)}$.

As discussed above in the Heat-Bath chain, we choose a site i uniformly at random and we replace σ_i by a spin chosen from π conditioned on the spins of all neighbors of i.

Suppose choose a site i. Say $d^+_\sigma(i)$ is the number of neighbors of i which are $+$ and $d^-_\sigma(i)$ is the number of neighbors which are $-$. It follows that for any sign $s \in \{+, -\}$

$$\mathbb{P}_\pi\left[\sigma_i = s | d^+_\sigma(i), d^-_\sigma(i)\right] = \frac{e^{2\beta d^s_\sigma(i)}}{e^{2\beta d^+_\sigma(i)} + e^{2\beta d^-_\sigma(i)}}$$

It is not hard to see that this is an aperiodic, and irreducible Markov chain and it is reversible with stationary distribution π.

As an exercise write down the Metropolis rule for the same Ising model example.

3.2 Coupling

Definition 3.1 (Coupling)**.** *Let μ, ν be probability distributions over Ω. A coupling between μ, ν is a probability distribution π on $\Omega \times \Omega$ that preserves the marginals of μ, ν respectively. In particular, for all $x \in \Omega$,*

$$\sum_y \pi(x, y) = \mu(x) \textit{ and } \sum_y \pi(y, x) = \nu(x).$$

Let us give an example: Consider the following two distributions over $\{1, 2, 3\}$. Consider the following

	1	2	3
$\mu(.)$	0.2	0.5	0.3
$\nu(.)$	0.3	0.4	0.3

coupling $\pi(1, 1) = 0.2, \pi(2, 2) = 0.4, \pi(3, 3) = 0.3, \pi(2, 1) = 0.1$. Observe that all marginal probabilities are satisfied; for example $\pi(2, 2) + \pi(2, 1) = 0.5 = \mu(2)$. Furthermore, if (X, Y) is a sample of π, we have that $\mathbb{P}_\pi [X = Y] = 0.9$. You can compare this coupling with an independent coupling in which $\widetilde{pi}(i, j) = \mu(i)\nu(j)$. In that case we would have $\mathbb{P}_{\tilde{\pi}} [X = y] = \pi(1)\nu(1) + \pi(2)\nu(2) + \pi(3)\nu(3) = 0.35$.

Lemma 3.2 (Coupling Lemma)**.** *Let μ and ν be probability distributions on Ω, and let X and Y be random variables with distributions μ and ν, respectively. Then*

1. $\mathbb{P}[X \neq Y] \geq \|\mu - \nu\|_{TV}$.
2. *There exists a coupling between μ and ν such that* $\mathbb{P}[X \neq Y] = \|\mu - \nu\|_{TV}$.

Proof. We prove the 2nd part, that is we construct the optimal coupling between μ and ν. The coupling will be very similar to the above example. For any i in the support of these distributions, we let $\pi(i, i) = \min\{\mu(i), \nu(i)\}$. For all other pairs $i \neq j$ we sequentially choose $\pi(i, j)$. Obviously, there is a way to match the remaining mass in the distributions such that all marginals are preserved. For example, when we process $i \neq j$ we define $\pi(i, j) = \min\{\mu(i) - \pi(i, .), \nu(j) - \pi(., j)\}$.

Note that obviously, this coupling has the highest probability that $X = Y$ because for any i, we must have

$$\mathbb{P}_{(X,Y)\sim\pi} [X = i, Y = i] \leq \min\{\mu(i), \nu(i)\}$$

in order to satisfy the marginal of i. Therefore,

$$\mathbb{P}[X \neq Y] = 1 - \sum_i \pi(i, i) = \sum_i \mu(i) - \min\{\mu(i), \nu(i)\} = \|\mu - \nu\|_{TV}.$$

□

3.3 Mixing Time

For a state x and time $t > 0$ let

$$\Delta_x(t) := \|K^t(x, .) - \pi\|_{TV}.$$

For a state x, let

$$\tau_x(\epsilon) = \min\{t : \Delta_x(t) \leq \epsilon\}$$

be the first time that the total variation distance of the walk started at x from the stationary distribution drops below ϵ. Now, define

$$\tau(\epsilon) = \max_x \tau_x(\epsilon).$$

Observe that if start the walk from a probability distribution p the time to reach total variation distance ϵ is at most $\tau(\epsilon)$. This is because for any t,

$$\begin{aligned} \|pK^t - \pi\|_{TV} &= \|\sum_i p(x)(K^t(x,.) - \pi)\|_{TV} \\ &\leq \sum_x p(x)\|K^t(x,.) - \pi\|_{TV} \leq \max_x \Delta_x(t). \end{aligned}$$

Definition 3.3 (Mixing Time)**.** *For any Markov chain the mixing time is defined as $\tau(1/2e)$. In other words, this is the time that the total variation distance of the walk started at the worst possible starting point drops be low $1/2e$.*

The choice of constant $1/2e$ is for algebraic convenience but as we will see this will only change the mixing time up to a constant.

We start by proving the following lemma:

Lemma 3.4. *For any state x and any time t,*

$$\Delta_x(t) \geq \Delta_x(t+1).$$

Proof. Let X_t be the location of the Markov chain started at $X_0 = x$ and Y_t be the location of the chain started at $Y_0 \sim \pi$. Note that since π is stationary Y_t is also distributed as π.

By the coupling lemma there is a coupling X_t, Y_t such that $\mathbb{P}[X \neq Y] = \Delta_x(t)$. Now we define a coupling of X_{t+1}, Y_{t+1} such that $\mathbb{P}[X_{t+1} \neq Y_{t+1}] \leq \Delta_x(t)$.

- If $X_t = Y_t$, then set $X_{t+1} = Y_{t+1}$.
- Otherwise, construct X_{t+1}, Y_{t+1} running the two chains independently for one step.

Obviously,

$$\Delta_x(t+1) = \|K^{t+1}(x,.) - \pi\|_{TV} \leq \mathbb{P}[X_{t+1} \neq Y_{t+1}] \leq \mathbb{P}[X_t \leq Y_t] = \Delta_x(t).$$

The first inequality follows by the coupling lemma and the second one by the above construction. □

The above lemma shows that once the total variation distance is small it will never gets big. To prove the fundamental theorem of Markov chains we need to show that for any chain there is a large enough t such that th1e total variation distance gets small.

Lemma 3.5. *Let $K \in \mathbb{R}_+^{\Omega \times \Omega}$ be such that for all x, y, $K(x,y) > 0$. Then,*

$$\tau_{\text{mix}} \leq \frac{1}{||\Omega| \min_{x,y} K(x,y)^2}$$

Proof. Let $k_{min} := \min_{x,y} K(x,y)$. We show that for any state x and any $t \geq 0$,

$$\Delta_x(t+1)) \leq \Delta_x(t)(1 - |\Omega| \cdot k_{min}^2).$$

This will prove the lemma because for $t = \frac{1}{|\Omega| \cdot k_{min}^2}$ we get

$$\Delta_x(t)) \leq (1 - |\Omega \cdot k_{min}^2)^t \leq 1/e.$$

Similar to Lemma 3.4 let X_t be the location of the chain started at x and Y_t be the location of the chain started at $Y_0 \sim \pi$. Also, suppose we have a coupling of X_t, Y_t such that $\mathbb{P}[X_t \neq Y_t] = \Delta_x(t)$. Consider the same coupling for X_{t+1}, Y_{t+1}: If $X_t = Y_t$ then $X_{t+1} = Y_{t+1}$ otherwise, we run each of these chains independently.

$$\begin{aligned}
\mathbb{P}[X_{t+1} \neq Y_{t+1}] &= \mathbb{P}[X_{t+1} \neq Y_{t+1} | X_t \neq Y_t] \mathbb{P}[X_t \neq Y_t] \\
&= (1 - \mathbb{P}[X_{t+1} = Y_{t+1} | X_t \neq Y_t]) \Delta_x(t) \\
&\leq (1 - \min_{a \neq b} \sum_z K(a,z) K(b,z)) \Delta_x(t) \\
&\leq (1 - |\Omega| \cdot k_{min}^2) \Delta_x(t).
\end{aligned}$$

Therefore, by coupling lemma $\Delta_x(t+1) \leq \Delta_x(t)(1 - |\Omega| \cdot k_{min}^2)$ as desired. □

Recall that in the last lecture we discussed that for any aperiodic irreducible chain there exists $t = t_0$ such that $K^t(x,y) > 0$ for all x, y. Say $\tilde{K} = K^{t_0}$. Then, by the above lemma there exists $t > 0$ such that for all x, $\|\tilde{K}^t(x,.) - \pi\|_{TV} \leq 1/2e$. But this means that

$$\|K^{t_0 \cdot t}(x,.) - \pi\|_{TV} \leq 1/2e,$$

as desired. This completes the proof of the fundamental theorem of Markov chains.

The following theorem can be proven using the above techniques and we leave it as an exercise.

Theorem 3.6. *For any Markov chain, $\tau(\epsilon) \leq O(\tau_{mix} \log(1/\epsilon))$.*

3.4 Coupling for Bounding Mixing Time

In the last section we saw the coupling as a technique to prove upper bound on the mixing time. The coupling proof that we constructed was very general and used no structure about the underlying chain. Here, we make this more formal, and we will see many examples on how to use this idea to bound the mixing time of a chain.

Definition 3.7. *A coupling of a Markov chain is a pair process X_t, Y_t such that each process in isolation looks like an honest simulation of the chain, i.e., for all states x, y,*

$$\mathbb{P}[X_{t+1} = y | X_t = x] = K(x,y) = \mathbb{P}[Y_{t+1} = y | X_t = x],$$

and for all $t \geq 0$, if $X_t = Y_t$, then $X_{t+1} = Y_{t+1}$.

We need another definition before we prove a bound on mixing time using a Markov chain coupling.

Definition 3.8 (Stopping Time). *A stopping time with respect to a sequence of random variables $X_1, X_2, \ldots$ is an integer random variable T with the property that for each $t \in \{1, 2, \ldots\}$, the occurrence or non-occurrence of the event $T = t$ depends only on the values of $X_1, X_2, \ldots, X_t$.*

For example, consider a random walker on a line who starts from the origin. Consider the first time that he is distance n away from the origin. This is a stopping time.

Lemma 3.9. *Let X_t, Y_t be a coupling of a Markov chain where $X_0 = x$ and $Y_0 \sim \pi$. Let*

$$T_{X,Y} = \min\{t : X_t = Y_t\},$$

be the stopping time until the two process meet. Then,

$$\Delta_x(t) \leq \mathbb{P}\left[T_{X,Y} > t\right].$$

Proof. This is just by the coupling lemma.

$$\Delta_x(t) = \|K^t(x,.) - \pi\|_{TV} \leq \mathbb{P}\left[X_t \neq Y_t\right] = \mathbb{P}\left[T_{X,Y} > t\right].$$

The second to last inequality is by the the coupling lemma. The last inequality uses follows from the fact that once the two chains collide they remain the same. □

In the next section we prove a bound on the mixing time of a hypercube using this idea.

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

Consider the n dimensional hypercube with 2^n 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, $\mathbb{E}\left[T_{k+1}\right] = \frac{1}{1-k/n}$. 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

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

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

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

So, in this case we get a very tight upper bound on the mixing. For this Markov chain with a

References

[SRO16] Christopher De Sa, Christopher Ré, and Kunle Olukotun. Ensuring rapid mixing and low bias for asynchronous gibbs sampling. In *ICML*, pages 1567–1576, 2016. 3-1