CS294 Partition Functions **Fall 2023**

Lecture Note 7

Instructor: Alistair Sinclair

7.1 The Fundamental Theorem and mixing time

In this section we will give an elementary probabilistic proof of the Fundamental Theorem of Markov Chains that is more intuitive than the linear algebra proof we gave for the reversible case in the last lecture. This will also lead into the idea of *coupling* for Markov chains, which we'll explore extensively in the next few lectures.

Theorem 7.1 (Fundamental Theorem of Markov Chains)**.** *If a Markov chain P is irreducible and aperiodic then it has a unique stationary distribution π. This is the unique left eigenvector of P (normalized so that its entries sum to 1) with eigenvalue* 1*. Moreover, $P^t(x,y) \to \pi(y)$ as $t \to \infty$ for all $x, y \in \Omega$.*

We will start from the assumption that some stationary distribution $\pi > 0$ exists; we can either prove this from first principles (see exercise below) or appeal to (part of) the Perron-Frobenius Theorem from the last lecture.

We recall the following standard notion of distance between probability distributions.

Definition 7.2. *For two probability distributions μ and ν on Ω, the* total variation distance *is*

$$\|\mu - \eta\|_{\text{TV}} := \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \eta(x)| = \max_{A\subseteq\Omega} |\mu(A) - \eta(A)|\,.$$

Note also that the total variation distance is quite unforgiving. E.g., let μ be the uniform distribution over all permutations of a deck of n cards, and η the same distribution conditional on the bottom card being the ace of spades. Then $\|\mu - \eta\|_{\text{TV}} = 1 - \frac{1}{n}$.

We will prove the Fundamental Theorem by showing that the total variation distance between π and $p_x^{(t)}$ decreases to zero. (In fact, we'll show that it decreases exponentially with t/τ_{mix}, where τ_{mix} is a parameter of the Markov chain. This will also give us an estimate on the time required to get within a given distance of π.)

Definition 7.3. (Coupling). *Let μ and η be any two probability distributions over Ω. A probability distribution ω over $\Omega \times \Omega$ is said to be a* coupling *of μ and η if its marginals are μ and η:*

$$\mu(x) = \sum_{y\in\Omega} \omega(x,y); \qquad \eta(y) = \sum_{x\in\Omega} \omega(x,y).$$

Lemma 7.4. (Coupling Lemma). *Let μ and η be probability distributions on Ω, and let X and Y be rvs with distributions μ and η, respectively. Then*

1. $\Pr[X \neq Y] \geq \|\mu - \eta\|_{\text{TV}}$.
2. *There exists a coupling of (μ, η) such that* $\Pr[X \neq Y] = \|\mu - \eta\|_{\text{TV}}$.

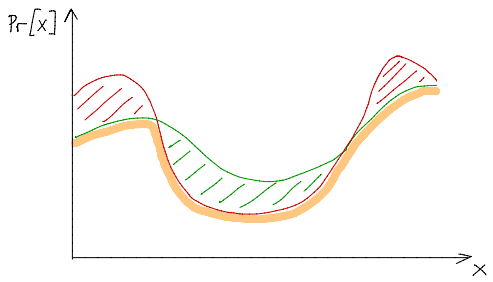

Figure 7.1: Two probability distributions. The thick line indicates the "lower envelope" of the two distributions.

Proof (informal sketch). Consider the two probability distributions shown in Figure 7.1. Suppose we try to construct a joint distribution for the two of them that maximizes the probability that they are equal. Clearly, the best we can do is to make $X = Y = z$ with probability $\min\{\Pr(X = z), \Pr(Y = z)\}$ for each value $z \in \Omega$. This is indicated by the thick line in the figure (the "lower envelope" of the two distributions). In this case, the probability that $X \neq Y$ is given by half the area of the shaded region. We then have

$$\frac{1}{2}(\text{Area of Shaded Region}) = \Pr[X \neq Y] = \frac{1}{2}\sum_{x \in \Omega} |\mu(x) - \eta(x)| = \|\mu - \eta\|_{\text{TV}}.$$

□

In order to discuss the convergence of $p_x^{(t)}$ to π, we will need the following definitions:

Definition 7.5.

1. *For any $x \in \Omega$, we define $\Delta_x(t) := \|p_x^{(t)} - \pi\|_{\text{TV}}$.*
2. *$\Delta(t) := \max_{x \in \Omega} \Delta_x(t)$ is the maximum possible distance from π after t steps.*

We now prove some basic facts about the time-dependent behavior of the Markov chain.

Claim 7.6. *$\Delta_x(t)$ is non-increasing in t.*

Proof. Let $X_0 = x$ and Y_0 have the stationary distribution π. We fix t and couple the distributions of the random variables X_t and Y_t such that $\Pr[X_t \neq Y_t] = \|p_x^{(t)} - \pi\|_{\text{TV}} = \Delta_x(t)$, which is possible by the Coupling Lemma. We now use this coupling to define a coupling of the distributions of X_{t+1} and Y_{t+1} as follows:

- If $X_t = Y_t$, then set $X_{t+1} = Y_{t+1}$.
- Otherwise, let $X_t \to X_{t+1}$ and $Y_t \to Y_{t+1}$ independently.

Then we have

$$\Delta_x(t+1) = \|p_x^{(t+1)} - \pi\|_{\text{TV}} \leq \Pr[X_{t+1} \neq Y_{t+1}] \leq \Pr[X_t \neq Y_t] = \Delta_x(t).$$

The first inequality holds because of the Coupling Lemma, and the second inequality comes from the construction of the coupling. □

We now define more specific quantities that capture the evolution of distance between corresponding distributions for arbitrary initial configurations.

Definition 7.7.

1. $D_{xy}(t) = \|p_x^{(t)} - p_y^{(t)}\|_{\text{TV}}$.
2. $D(t) = \max_{x,y\in\Omega} D_{xy}(t)$.

The following simple relationship between $D(t)$ and $\Delta(t)$ is left as an **exercise**:

Claim 7.8. $\Delta(t) \leq D(t) \leq 2\Delta(t)$.

Next we prove that the quantity $D(t)$ is submultiplicative.

Claim 7.9. *For all positive integers s, t, $D(s+t) \leq D(s)D(t)$.*

Proof. Let $X_0 = x$ and $Y_0 = y$. We use the Coupling Lemma to couple the distributions of X_t and Y_t so that

$$D_{xy}(t) = \|p_x^{(t)} - p_y^{(t)}\|_{\text{TV}} = \Pr[X_t \neq Y_t].$$

We then construct a coupling of X_{t+s} and Y_{t+s} as follows:

- If $X_t = Y_t$ then set $X_{t+i} = Y_{t+i}$ for $i = 1, 2, \ldots s$.
- Otherwise, let $X_t = x'$ and $Y_t = y' \neq x'$. Use the Coupling Lemma to couple the distributions of X_{t+s} and Y_{t+s}, conditioned on $X_t = x'$ and $Y_t = y'$, such that

$$\Pr[X_{t+s} \neq Y_{t+s} | X_t = x', Y_t = y'] = \|p_{x'}^{(s)} - p_{y'}^{(s)}\|_{\text{TV}} = D_{x'y'}(s) \leq D(s). \tag{7.1}$$

The last inequality holds by the definition of $D(s)$. We now have

$$\begin{aligned} D_{xy}(t+s) &= \|p_x^{(t+s)} - p_y^{(t+s)}\|_{\text{TV}} \\ &\leq \Pr[X_{t+s} \neq Y_{t+s}] \text{ by the Coupling Lemma} \\ &\leq D(s)D_{xy}(t) \text{ by the construction of the coupling} \\ &\leq D(s)D(t). \end{aligned}$$

Since this holds for all x, y, we get that $D(t+s) \leq D(s)D(t)$. □

Claim 7.10. *If P is irreducible and aperiodic, then $D(t) < 1$ for some finite t.*

Proof. Since P is irreducible and aperiodic, there exists some finite t_0 for which $P^t(x, y) > 0$ for all x, y and $t \geq t_0$. Then we certainly have $D(t_0) < 1$. □

Proof of Theorem 7.1. Let t be such that $D(t) = 1 - \delta$ for some $\delta > 0$. By Claim 7.8, for any positive integer k we have

$$\Delta(kt) \leq D(kt) \leq D(t)^k \leq (1-\delta)^k. \tag{7.2}$$

Since $\Delta(t)$ is non-increasing, we see that $\Delta(t) \to 0$ as $t \to \infty$, as required. □

Exercise: The above argument almost constitutes a complete elementary probabilistic proof of the fundamental theorem. The only hole is that we assumed that a stationary distribution π exists. Follow these simple steps to prove that π exists whenever P is irreducible and aperiodic:

(i) For arbitrary $x \neq y \in \Omega$, let $q_x(y)$ be the expected number of times that the Markov chain, started in state x, visits y before returning to x. Also, let $q_x(x) = 1$. Show that, for any x, the vector $q_x(\cdot)$ is finite and strictly positive.

(ii) Show that, for any x, $q_x P = q_x$. [Hint: expand the expectation as $q_x(y) = \sum_{t \geq 1} r_x^{(t)}(y)$, where $r_x^{(t)}(y)$ is the probability that the chain is at y after t steps assuming it hasn't returned to x.]

(iii) Deduce that a strictly positive stationary distribution π always exists.

We now give a more quantitative version of the above theorem. The following is a key definition for this part of the course.

Definition 7.11.

1. $\tau_x(\varepsilon) := \min\{t : \Delta_x(t) \leq \varepsilon\}$ *is the first time* t *at which the distance* $\|p_x^{(t)} - \pi\|_{\mathrm{TV}}$ *drops to* ε.
2. $\tau(\varepsilon) := \max_{x \in \Omega} \tau_x(\varepsilon)$.
3. $\tau_{\mathrm{mix}} := \tau(1/2e)$ *is called the* mixing time *of the chain.*

In other words, the mixing time is the time until the variation distance, starting from the worst possible initial state $x \in \Omega$, reaches $1/2e$. This value is chosen for algebraic convenience only, as we shall see below.

Claim 7.12. $\Delta(t) \leq \exp\left(-\left\lfloor \frac{t}{\tau_{\mathrm{mix}}} \right\rfloor\right)$.

Proof. Following the same reasoning as in (7.2), we have

$$\Delta(k\tau_{\mathrm{mix}}) \leq D(k\tau_{\mathrm{mix}}) \leq D(\tau_{\mathrm{mix}})^k \leq (2\Delta(\tau_{\mathrm{mix}}))^k \leq e^{-k}.$$

The last inequality follows from the definition of τ_{mix}, and proves the claim. (It is in the last step that we need $\Delta(\tau_{\mathrm{mix}})$ to be *strictly* less than $\frac{1}{2}$; our choice of $\Delta(\tau_{\mathrm{mix}}) = \frac{1}{2e}$ satisfies this and leads to a particularly simple expression for $\tau(\varepsilon)$.) □

Corollary 7.13. $\tau(\varepsilon) \leq \tau_{\mathrm{mix}} \left\lceil \ln\left(\varepsilon^{-1}\right) \right\rceil$.

The corollary follows immediately from Claim 7.12 and justifies our definition of mixing time: the cost of obtaining any desired variation distance ε is only a modest factor times τ_{mix}.

Note: Claim 7.12 shows that the variation distance to stationarity, $\Delta(t)$, decays (at worst) exponentially with time constant τ_{mix}, so that the variation distance goes from close to 1 to close to 0 over a time interval of length $O(\tau_{\mathrm{mix}})$. As was first observed in [Dia96, SC97], many Markov chains (in particular, highly symmetric ones) exhibit a so-called "cutoff" phenomenon, whereby the decay of $\Delta(t)$ actually happens much more sharply, over a time interval $\tau_{\mathrm{mix}}(1 + o(1))$ (very roughly speaking). See [LPW09] for further examples.

7.2 Coupling for mixing times

We now discuss how to use coupling as a tool for bounding the mixing time. This method, in addition to being beautifully simple, often gives tight bounds (up to constant factors) when it works.

Consider as usual an ergodic (i.e., irreducible, aperiodic) Markov chain on some state space Ω. Consider two particles started at positions x and y in Ω, each individually moving through the state space according to the Markov transition matrix P, but whose evolutions may be coupled in some way. We will show below that the time until the two particles meet gives a bound on the mixing time; more precisely, we will show that

$$\Delta(t) \leq \max_{x,y} \Pr[\text{two particles started at positions } x,\, y \text{ have not met by time } t],$$

where we recall that $\Delta(t) := \max_x \|p_x^{(t)} - \pi\|_{\text{TV}}$.

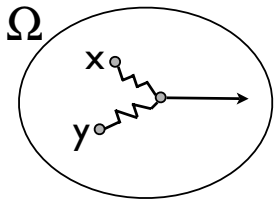

Figure 7.2: Coupling

To describe this approach, we first specialize the definition of coupling slightly from the previous section.

Definition 7.14. *A* coupling *of a Markov chain P is a pair process (X_t, Y_t) such that:*

1. *each of $(X_t, \cdot)$ and $(\cdot, Y_t)$, viewed in isolation, is a faithful copy of the Markov chain; that is,*

$$\Pr[X_{t+1} = b \,|\, X_t = a] = P(a,b) = \Pr[Y_{t+1} = b \,|\, Y_t = a];$$

   and

2. *if $X_t = Y_t$ then $X_{t+1} = Y_{t+1}$.*

Figure 7.2 gives a pictorial illustration of this definition.

Now define the random variable $T_{xy} = \min\{t : X_t = Y_t \,|\, X_0 = x, Y_0 = y\}$ to be the time until the two processes meet. The following claim gives the desired upper bound on the mixing time:

Claim 7.15. $\Delta(t) \leq \max_{x,y} \Pr[T_{xy} > t]$.

Proof. Recall the following definition from the previous section: $D(t) := \max_{x,y} \|p_x^{(t)} - p_y^{(t)}\|_{\text{TV}}$. Then

$$\begin{aligned}
\Delta(t) &\leq D(t) \\
&= \max_{x,y} \|p_x^{(t)} - p_y^{(t)}\|_{\text{TV}} \\
&\leq \max_{x,y} \Pr[X_t \neq Y_t \,|\, X_0 = x, Y_0 = y] \\
&= \max_{x,y} \Pr[T_{xy} > t \,|\, X_0 = x, Y_0 = y].
\end{aligned}$$

The only real content in this proof is the third line, where we use the Coupling Lemma. □

Corollary 7.16. *For any coupling, the mixing time is bounded by* $\tau_{\text{mix}} \leq 2e \max_{x,y} \mathrm{E}(T_{xy})$.

Proof. Use the definition of τ_{mix} and apply Markov's inequality. □

Coupling ideas for analyzing the time-dependent behavior of Markov chains can be traced back to Doeblin in the 1930s. However, the modern development of the topic was initiated by David Aldous [Ald83].

7.3 Examples

7.3.1 Simple random walk on the hypercube $\{0,1\}^n$

The n-dimensional cube is a graph with 2^n vertices, each of which can be encoded as an n-bit binary string $x_1 x_2 \cdots x_n$, whose neighbors are the strings which differ from it by Hamming distance exactly 1. We define a lazy random walk on the cube as follows:

1. With probability 1/2, do nothing.
2. Else, pick a coordinate $i \in \{1, \ldots, n\}$ uniformly at random and flip coordinate x_i (i.e. $x_i \to 1 - x_i$).

Note that the cube is bipartite, so we need to introduce some laziness to make the random walk ergodic. A self-loop probability of $\frac{1}{2}$ is a natural choice.

The above process is clearly equivalent to the following:

1. Pick a coordinate $i \in \{1, \ldots, n\}$ uniformly at random *and* a bit $b \in \{0, 1\}$ uniformly at random.
2. Set $x_i = b$.

This second description of the random walk suggests the following coupling: make X_t and Y_t choose the *same* i and b at every step. Clearly this is a valid coupling: obviously each of X_t and Y_t is performing exactly the above random walk.

To analyze the time T_{xy}, notice that once every $i \in \{1, \ldots, n\}$ has been chosen at least once, X_t must equal Y_t. (This is because, once a coordinate i has been chosen, X_t and Y_t agree on that coordinate at all future times.) Thus for any x and y, T_{xy} is stochastically dominated by the time for a coupon collector to collect all n coupons. Hence $\Pr[T_{xy} > n \ln n + cn] \leq e^{-c}$, and so by Claim 7.15 we have $\Delta(n \ln n + cn) \leq e^{-c}$; therefore in particular

$$\tau_{\text{mix}} \leq n \ln n + O(n),$$

and more precisely

$$\tau(\varepsilon) \leq n \ln n + \lceil n \ln(\varepsilon^{-1}) \rceil.$$

An exact analysis of this very simple random walk reveals that in fact $\tau_{\text{mix}} \sim (1/2) n \ln n$, so our analysis is tight up to a factor of 2.

Exercise: Show that $\tau_{\text{mix}} \geq \frac{1}{2} n \ln n (1 - o(1))$ using the fact that, for any constant C, the time needed by the collector to collect all but $C\sqrt{n}$ coupons is a.s. $\frac{1}{2} n \ln n - \omega(n)$, where $\omega(n)$ is any function s.t. $\frac{\omega(n)}{n} \to \infty$. [Hint: What is the standard deviation of the number of 1's in the stationary distribution?]

Note: This random walk exhibits the "cutoff" phenomenon mentioned in the previous section, in the sense that $\Delta(\frac{1}{2}n\ln n - \omega(n)) = 1 - o(1)$ and $\Delta(\frac{1}{2}n\ln n + \omega(n)) = o(1)$, for any function $\omega(n)$ as above. In particular, we can take $\omega(n) = o(n\ln n)$.

Exercise: Consider the same random walk but with self-loop probability $\frac{1}{n+1}$ rather than $\frac{1}{2}$ at every state. Design a coupling that gives a bound of $\frac{1}{2}n\ln n + O(n)$ on the mixing time for this walk. (The true answer here is $\tau_{\text{mix}} = \frac{1}{4}n\ln n + O(n)$ [Ald83], so again we are off only by a factor of 2.)

Exercise: For every self-loop probability δ with $\delta > \text{const}/n$ and $\delta < 1 - \text{const}/n$, show that $\tau_{\text{mix}} \leq c_\delta n\ln n + O(n)$, where c_δ is a constant. (The exact value of c_δ is not important.)

7.3.2 Top-in-at-Random Shuffle

Recall that the top-in-at-random shuffle involves repeatedly taking the top card from a deck of n cards and inserting it at a position chosen uniformly at random in the deck. Analyzing this shuffle by coupling is best done via its inverse, defined as follows:

- Pick a card c from the deck uniformly at random.
- Move card c to the top of the deck.

Exercise: Show that the variation distances at time t for the inverse shuffle and the original shuffle are the same. [Hint: The original shuffle is a random walk on the symmetric group, with generators $\{g_i\}$, where g_i moves the top card to position i in the permutation. The inverse shuffle is random walk with generators g_i^{-1}. Show that there is a 1-1 correspondence between paths of length t starting from state x in the two shuffles.]

To construct a coupling for the inverse shuffle, let X_t, Y_t be two copies of the deck of cards at time t, and make X_t and Y_t choose the *same* card c (which of course is not necessarily in the same position in both decks) and move it to the top. Now the key observation is the following: once a card has been chosen in the coupling, this card will be in the same position in both decks for the rest of time. [**Exercise:** Check this.] T_{xy} is therefore once again dominated by the coupon collector random variable for n coupons. This leads to $\tau_{\text{mix}} \leq n\ln n + O(n)$ and $\tau(\varepsilon) \leq n\ln n + \lceil n\ln\varepsilon^{-1}\rceil$. This value for τ_{mix} can be shown to be tight, even up to the constant.

7.3.3 Random Transposition Shuffle

This shuffle is defined as follows:

- Pick positions i and j uniformly at random.
- Switch the cards at positions i and j.

An equivalent, more convenient description is the following:

- Pick card c and position i uniformly at random.
- Exchange card c with the card at position i.

It is easy to define a coupling using this second definition: namely, make X_t and Y_t choose the same c and i at each step. This coupling ensures that the distance between X and Y is non-increasing. More explicitly, writing $d_t = d(X_t, Y_t)$ for the number of positions at which the two decks differ, we have:

1. If card c is in the same position in both decks, then $d_{t+1} = d_t$.
2. If card c is in different positions in the two decks, there are two possible subcases:
   (a) If the card at position i in both decks is the same, then $d_{t+1} = d_t$.
   (b) Otherwise, $d_{t+1} \leq d_t - 1$.

Thus we get a decrease in distance only in case 2(b), and this occurs with probability

$$\Pr[d_{t+1} < d_t] = \left(\frac{d_t}{n}\right)^2.$$

Therefore, the time for d_t to decrease from value d is stochastically dominated by a geometric random variable with mean $\left(\frac{n}{d}\right)^2$. This implies that $E[T_{xy}] \leq \sum_{d=1}^{n} \left(\frac{n}{d}\right)^2$, which is $O(n^2)$.

Invoking Markov's inequality, we get that $\Pr[T_{xy} > cn^2] < c' = \frac{1}{2e}$ for a suitable constant c, which leads to the bound

$$\tau_{\text{mix}} \leq cn^2.$$

Actually, for this shuffle it is known [DS81] that

$$\tau_{\text{mix}} \sim \frac{1}{2} n \ln n,$$

so our analysis in this case is off by quite a bit.

Challenge: Design a coupling that gives $\tau_{\text{mix}} = O(n \log n)$. [Note: This is a long-standing open problem (though not terribly pressing as we already know the mixing time exactly by other methods); unpublished solutions can be found in [Bor11, BK11] but I haven't read these in detail.]

Remark: It turns out that for any ergodic Markov chain there is *always* a coupling that is optimal, in the sense that the coupling time satisfies

$$\Pr[T_{xy} > t] = D_{xy}(t).$$

This is a very general theorem of Griffeath [Gri78]. However, the couplings that achieve the mixing time may involve looking arbitrarily far into the future in the two Markov chains, and thus are not useful in practice. The couplings we have used so far—and almost all couplings used in algorithmic applications—are *Markovian* couplings: the evolution of X_t and Y_t depends only on the current values of X_t and Y_t. For some Markov chains, there is (provably) no Markovian coupling that achieves the mixing time [KR99]. Occasionally, it has been possible in concrete examples to make use of non-Markovian couplings; see, e.g., [HV03].

References

[Ald83] D. Aldous. Random walks on finite groups and rapidly mixing Markov chains. *Séminaire de Probabilites XVII*, pages 243–297, 1983.

[BK11] R.M. Burton and Y. Kovchegov. Mixing times via super-fast coupling. Technical Report, Oregon State Univ. Math Dept. 2011.

[Bor11] O. Bormashenko. A coupling argument for the random transposition shuffle. ArXiv preprint 1109.3915. 2011.

[Dia96] P. Diaconis. The cutoff phenomenon in finite Markov chains. *Proceedings of the National Academy of Sciences*, 93:1659–1654, 1996.

[DS81] P. Diaconis and M. Shahshahani. Generating a random permutation with random transpositions. *Zeitschrift für Wahrscheinlichketistheorie und verwandte Gebiete*, 57:159–179, 1981.

[Gri78] D. Griffeath. Coupling methods for Markov processes. In G.-C. Rota, editor, *Studies in Probability and Ergodic Theory*, pages 1–43. Academic Press, 1978.

[HV03] T. Hayes and E. Vigoda. A non-Markovian coupling for randomly sampling colorings. *Proceedings of the 44th Annual IEEE Symposium on Foundations of Computer Science*, pages 618–627, 2003.

[KR99] V.S. Anil Kumar and H. Ramesh. Markovian coupling vs conductance for the Jerrum-Sinclair chain. *Proceedings of the 40th IEEE FOCS*, pages 241–250, 1999.

[LPW09] D. Levin, Y. Peres, and E. Wilmer. *Markov Chains and Mixing Times*. American Mathematical Society, Providence, RI, 2009.

[SC97] L. Saloff-Coste. Lectures on finite Markov chains. *Lectures on Probability Theory and Statistics*, pages 301–413, 1997.