Lecture 2: Introduction to Markov Chains and Mixing Times

September 12, 2023

A significant portion of the course will focus on studying *Markov chains*. Needless to say, they are one of the most ubiquitous approaches to sampling in practice. As an example, Markov chains have recently been deployed to detect gerrymandering in redistricting plans [DDS21]. The high-level idea is one can compare a proposed plan with an ensemble of "typical" plans drawn from some probability distribution. Markov chains are also famously used to model various card-shuffling strategies [**???**]. Persi Diaconis also describes in his survey a very cool application of Markov chains to decrypting secret messages in prisons [Dia09].

At a high level, the main idea is to run a stochastic process where simple, easy-to-compute, random updates are repeatedly applied to some initial starting point. The hope is after enough iterations of this process, we've "added enough randomness in the right ways" so that the final iterate is (approximately) distributed according to some target probability measure. Markov chains are typically easy to design and implement in software, and demonstrate strong empirical performance in downstream applications. However, despite these appealing features, one major difficulty is they do not automatically come equipped with a testable criterion for termination. One of the major goals of this course is to provide a mathematically rigorous toolbox for bounding the "mixing time" of a Markov chain, which quantifies how long the chain should be run.

1 Basics of Markov Chains

Let Ω be a finite state space.

Definition 1 (Markov Chain)**.** *A* (discrete-time) Markov chain *on* Ω *is a sequence of random variables* $\{X_t\}_{t=0}^{\infty}$ *taking values in* Ω *satisfying the* Markov property*:*

$$\Pr[X_t = x_t \mid X_0 = x_0, \dots, X_{t-1} = x_t] = \Pr[X_t = x_t \mid X_{t-1} = x_{t-1}], \qquad \forall t \geq 0, \forall x_0, \dots, x_t \in \Omega. \tag{1}$$

In other words, the distribution of the next state X_t is independent of the history $X_0, \dots, X_{t-2}$ given X_{t-1}. Throughout the course, our Markov chains will be *time-homogeneous*, meaning we can describe the Markov chain by two parameters:

- First, we have a *transition probability matrix* (or *Markov kernel*) $\mathsf{P} \in \mathbb{R}_{\geq 0}^{\Omega \times \Omega}$. The entries $\mathsf{P}(x, y)$ (or $\mathsf{P}(x \to y)$) specify the transition probabilities $\Pr[X_t = y \mid X_{t-1} = x]$ for all $x, y \in \Omega$. Thus, each row $\mathsf{P}(x \to \cdot)$ is a probability distribution over Ω.
- Second, we have an initial distribution $\mu_0 \in \mathbb{R}_{\geq 0}^{\Omega}$ specifying the law of X_0.

Linear algebraically, the distribution $\mu_t = \mathsf{Law}(X_t)$ of X_t over Ω is then given by

$$\mu_t = \mu_0 \mathsf{P}^t$$

when viewed as row vectors. We'll often refer to P as the Markov chain itself (even if μ_0 is unspecified) since P is more important. One should conceptually imagine the Markov chain as a *random walk* on Ω, where we make Ω into a directed graph with a directed edge $(x \to y)$ having weight $\mathsf{P}(x \to y)$ for every $x, y \in \Omega$.

We'll use the following as a running example.

Example 1 (Random Walks on Graphs). Let $G = (V, E)$ be an undirected graph. The *simple random walk* on G is a Markov chain with state space $\Omega = V$ described by the following process: If the current vertex is $u \in V$, then we transition to a uniformly random neighbor of u. We can write down the transition probability matrix P_G of this Markov chain as

$$\mathsf{P}_G = D_G^{-1} A_G,$$

where A_G denotes the $\{0,1\}$-adjacency matrix of G, and $D_G = \mathrm{diag}(\deg_G(v) : v \in V)$ is a diagonal matrix which normalizes everything so that the rows of P_G sum to 1.

1.1 Stationarity and Convergence

Having defined what a Markov chain is, let us now build up the connection to sampling.

Definition 2 (Stationary/Equilibrium Distribution)**.** *A probability measure μ on Ω is* stationary *w.r.t. a Markov chain P if $\mu\mathsf{P} = \mu$.*

For instance, if the initial distribution μ_0 is stationary w.r.t. P, then every state X_t in the stochastic process $\{X_t\}_{t=0}^{\infty}$ is distributed according to μ_0. Continuing Example 1, if we run the simple random walk on $G = (V, E)$ for a long time, intuitively we expect vertices with high degree will be visited more often. If we define the distribution μ on V by $\mu(v) \propto \deg_G(v)$, then μ is stationary w.r.t. P_G.

It turns out that in general, a stationary distribution always exists.

Lemma 1.1. *Every Markov chain P has at least one stationary distribution.*

We prove this in Appendix A. Ideally, we'd like our Markov chains to have a *unique* stationary distribution. The following gives a criterion for uniqueness.

Definition 3 (Ergodicity)**.** *Fix a Markov chain P on a finite state space Ω. We say P is* ergodic *if P satisfies both of the following properties:*

- ***Irreducibility:*** *P is* irreducible *if for all $x, y \in \Omega$, there exists $t \geq 0$ such that $\mathsf{P}^t(x \to y) > 0$. In other words, the underlying weighted directed graph of P is strongly connected.*
- ***Aperiodicity:*** *The* period *of a state $x \in \Omega$ under P is defined as the greatest common divisor of $\{t \geq 1 : \mathsf{P}^t(x, x) > 0\}$. We say P is* aperiodic *if all states have period 1.*

Remark 1. Note that ergodicity is actually a very weak and easy-to-satisfy property. For irreducibility, we just need connectivity of Ω under the transitions of P. One way to ensure aperiodicity is to ensure that $\mathsf{P}(x, x) > 0$ for all $x \in \Omega$. In particular, any Markov chain P, aperiodic or not, can be made into an "equivalent" aperiodic Markov chain by replacing P with $\frac{\mathsf{Id}+\mathsf{P}}{2}$, where Id is the $\Omega \times \Omega$ identity matrix. This is sometimes called the *lazification* of P, since in each step, there is a $\frac{1}{2}$-probability of staying in the same state. Essentially all of the usual properties of P are preserved when looking at $\frac{\mathsf{Id}+\mathsf{P}}{2}$.

Lemma 1.2. *Let P be a Markov chain on a finite state space Ω. If P is irreducible and aperiodic, then there exists t^* such that for all $x, y \in \Omega$, $\mathsf{P}^{t^*}(x \to y) > 0$.*[1]

We prove this in Appendix A.

Theorem 1.3 (Fundamental Theorem of Markov Chains; see e.g. [LPW17])**.** *Let P be an ergodic Markov chain on a state space Ω. Then P has a unique stationary distribution μ on Ω. Furthermore, for every initial distribution μ_0, the distribution $\mu_t = \mu_0\mathsf{P}^t$ of X_t converges (pointwise) to μ as $t \to \infty$.*

We will prove this theorem in the next lecture.

[1] The latter condition is the more "complete" definition of ergodicity.

The Markov Chain Monte Carlo Paradigm Theorem 1.3 highlights the relevance of Markov chains to sampling. To sample from some complicated probability distribution μ on some complicated state space Ω, it suffices to design an ergodic Markov chain P on Ω such that:

- μ is its equilibrium distribution, and
- the transitions of P are efficiently implementable.

The algorithm is then to select an arbitrary initial state $X_0 \in \Omega$, simulate P started at X_0 for T steps, and output the final state X_T as your sample. Theorem 1.3 guarantees that if T is sufficiently large, then $\mathsf{Law}(X_T)$ is "close" to μ. Of course, the challenge then becomes how to choose T. This will be the focus of a nontrivial fraction of the course.

1.2 Reversible Markov Chains

It turns out, if you're given a Markov chain "in the wild", it is a highly nontrivial task to determine its stationary distribution. However, there is a large class of Markov chains for which this turns out to be easy.

Definition 4 (Reversibility)**.** *We say a Markov chain* P *is* reversible *w.r.t. a distribution* μ *if together they satisfy the* detailed balance condition*:*

$$\mu(x) \cdot \mathsf{P}(x \to y) = \mu(y) \cdot \mathsf{P}(y \to x), \qquad \forall x, y \in \Omega.$$

Lemma 1.4. *Let* P *be a Markov chain reversible w.r.t. a distribution* μ *on* Ω*. Then* μ *is stationary w.r.t.* P*.*

We prove this in Appendix A.

For instance, if P is *symmetric*, i.e. $\mathsf{P}(x \to y) = \mathsf{P}(y \to x)$ for all $x, y \in \Omega$, then the uniform measure over Ω is stationary w.r.t. P. Again, returning to Example 1, the simple random walk on an undirected graph $G = (V, E)$ is reversible w.r.t. the distribution $\mu(v) \propto \deg_G(v)$. This is, in some sense, the "only" example of a reversible Markov chain: If P is a reversible w.r.t. μ, then we can define a weighted graph on Ω, where we have an edge $\{x, y\}$ connecting $x, y \in \Omega$ if $\mathsf{P}(x \to y) > 0$. Furthermore, we assign this edge weight $\mu(x) \cdot \mathsf{P}(x \to y)$. The analogous simple random walk on this weighted graph recovers the Markov chain P.

For reversible chains (e.g. simple random walk on a graph G), irreducibility is equivalent to connectivity of the underlying graph. Similarly, aperiodicity is equivalent to the underlying graph being *not* bipartite. Throughout this course, all of our Markov chains will be reversible. This reversibility condition can be interpreted linear algebraically as saying the matrix P is *self-adjoint* w.r.t. the inner product $\langle f, g \rangle_\mu \stackrel{\text{def}}{=} \sum_{x \in \Omega} \mu(x) f(x) g(x) = \mathbb{E}_\mu[fg]$ induced by μ on $\{f : \Omega \to \mathbb{R}\} \cong \mathbb{R}^\Omega$. In particular, P has all real eigenvalues. We'll say more about this later.

2 Markov Chain Examples

We now give some examples of useful Markov chains in various settings.

2.1 Glauber Dynamics

Let μ be a probability measure on $\{\pm 1\}^n$. We define a local Markov chain which changes only a single coordinate in each step called *Glauber dynamics*. For each $\sigma \in \{\pm 1\}^n$, we write $\sigma^{\oplus i}$ for the unique configuration where we replace σ_i by $-\sigma_i$, while keeping all other coordinates fixed. The *(heat-bath) Glauber dynamics* (sometimes called the *Gibbs sampler*, especially in machine learning circles) is described by the following two-step process:

1. Select a coordinate $i \in [n]$ uniformly at random.
2. Flip coordinate i with probability $\frac{\mu(\sigma^{\oplus i})}{\mu(\sigma) + \mu(\sigma^{\oplus i})}$.

Thus, the transition probabilities are given by

$$\mathsf{P}^{\mathsf{GD}}(\sigma \to \sigma^{\oplus i}) = \frac{1}{n} \cdot \frac{\mu(\sigma^{\oplus i})}{\mu(\sigma) + \mu(\sigma^{\oplus i})}, \qquad \forall i \in [n].$$

With all of the remaining probability, we stay at the current state. Since

$$\mu(\sigma) \cdot \mathsf{P}^{\mathsf{GD}}(\sigma \to \sigma^{\oplus i}) = \frac{1}{n} \cdot \frac{\mu(\sigma) \cdot \mu(\sigma^{\oplus i})}{\mu(\sigma) + \mu(\sigma^{\oplus i})} = \mu(\sigma^{\oplus i}) \cdot \mathsf{P}^{\mathsf{GD}}(\sigma^{\oplus i} \to \sigma),$$

P^{GD} is reversible w.r.t. μ by construction.

We note that Glauber dynamics generalizes to any probability distribution over a product space $[q]^n$, where $[q] = \{1, \ldots, q\}$ and $q \geq 2$. If $\sigma \in [q]^n$ and $\mathfrak{c} \in [q]$, write $\sigma^{i,\mathfrak{c}} \in [q]^n$ for the unique configuration obtained by replacing σ_i with $\mathfrak{c}$ and keeping other coordinates fixed. Then

$$\mathsf{P}^{\mathsf{GD}}(\sigma \to \sigma^{i,\mathfrak{c}}) = \frac{\mu(\sigma^{i,\mathfrak{c}})}{\sum_{\mathfrak{b} \in [q]} \mu(\sigma^{i,\mathfrak{b}})}, \qquad \forall \mathfrak{c} \in [q].$$

In other words, in each step, we select a uniformly random coordinate $i \in [n]$ and resample σ_i conditioned on the current assignments for all other coordinates.

2.2 Perfect Matchings via Random Transpositions

Let $G = (V, E)$ be a bipartite graph with bipartition $L \sqcup R$ and $|L| = |R|$. Let Ω be the set of perfect matchings on G, i.e. subsets of edges $M \subseteq E$ such that every vertex in V is incident to *exactly* one edge in M. We define a Markov chain P with uniform stationary distribution as follows:

- Select two distinct edges $u_1v_1, u_2v_2 \in M$ uniformly at random. Note that all four vertices $u_1, u_2 \in L$ and $v_1, v_2 \in R$ must all be distinct by the matching constraint.
- If $M \cup \{u_1v_2, u_2v_1\} \setminus \{u_1v_1, u_2v_2\}$ is also a perfect matching, then we perform the swap, i.e. transition to this new matching. Otherwise, we stay at M.

It is straightforward to check that this Markov chain is reversible w.r.t. the uniform distribution over perfect matchings.

Unfortunately, for general bipartite graphs, the Markov chain is not even connected. One example of this is a long even-length cycle C_{2n}. In this case, there are just two perfect matchings which are completely disjoint from each other. Hence, you cannot possibly reach one matching from the other via local moves like those in the swap Markov chain we just defined.

2.3 The Metropolis Filter

Let Ω be a giant state space, and let $w : \Omega \to \mathbb{R}_{\geq 0}$ be a nonnegative weight function. Our goal is to design an ergodic Markov chain with unique stationary distribution $\mu(x) \propto w(x)$. Reversibiliity gives us a simple recipe to do this.

We start by constructing a Markov chain Q on Ω which is symmetric, i.e. $Q(x \to y) = Q(y \to x)$ for all $x, y \in \Omega$. Note that the uniform distribution over Ω is stationary w.r.t. Q, not μ. Effectively, this Q endows Ω with a *neighborhood structure*, and all we ask is that Ω is connected under Q. For each $x \in \Omega$, the transition distribution $Q(x \to \cdot)$ is called the *proposal distribution*.

With this in hand, we can then add a *Metropolis filter* on top of Q to correct its stationary distribution. More specifically, to take a step from the current state x, we first draw a sample $y \sim Q(x \to \cdot)$ ("the proposal"). We then either transition to y with probability $\min\left\{1, \frac{w(y)}{w(x)}\right\}$, or stay at x with the remaining probability. In other words, our new Markov chain P will have transitions

$$\mathsf{P}(x \to y) = \begin{cases} Q(x \to y) \cdot \min\left\{1, \frac{w(y)}{w(x)}\right\}, & \text{if } y \neq x \\ 1 - \sum_{z \neq x} \mathsf{P}(x \to z), & \text{if } y = x. \end{cases}$$

Lemma 2.1. *Let Q be any symmetric Markov chain. Then the "Metropolized" chain P is reversible w.r.t. $\mu(x) \propto w(x)$.*

Proof. We verify detailed balanced. If $x \neq y$, then

$$\begin{aligned} w(x) \cdot \mathsf{P}(x \to y) &= w(x) \cdot \min\left\{1, \frac{w(y)}{w(x)}\right\} \cdot Q(x \to y) \\ &= \min\{w(x), w(y)\} \cdot Q(x \to y) \\ &= \min\{w(x), w(y)\} \cdot Q(y \to x) && \text{(Symmetry of } Q\text{)} \\ &= w(y) \cdot \min\left\{1, \frac{w(x)}{w(y)}\right\} \cdot Q(y \to x) \\ &= w(y) \cdot \mathsf{P}(y \to x). \end{aligned}$$

□

3 Quantifying Speed of Convergence: Mixing Times

Now that we can design ergodic Markov chains with the correct stationary distribution, the goal is quantify how quickly the chain equilibriates. This is crucial since it directly controls the efficiency of our Markov chain sampling algorithms, as well as the "accuracy" of the samples we get out.

Definition 5 ((Total Variation) Mixing Time)**.** *Fix an ergodic Markov chain* P *with stationary distribution* μ *on a state space* Ω*. Let* $\epsilon > 0$ *be an error parameter. We define the* ***ϵ-mixing time of*** P ***(with initial distribution*** μ_0***)*** *by*

$$\begin{aligned} \mathrm{T}_{\mathsf{mix}}(\epsilon; \mu_0, \mathsf{P}) &\stackrel{\mathsf{def}}{=} \min\left\{t \geq 0 : \left\|\mu_0 \mathsf{P}^t - \mu\right\|_{\mathsf{TV}} \leq \epsilon\right\} \\ \mathrm{T}_{\mathsf{mix}}(\epsilon; \mathsf{P}) &\stackrel{\mathsf{def}}{=} \sup_{\mu_0} \mathrm{T}_{\mathsf{mix}}(\epsilon; \mu_0, \mathsf{P}). \end{aligned}$$

When the Markov chain P *is clear from context, we drop the* P*. We define the total variation mixing time of* P *to be* $\mathrm{T}_{\mathsf{mix}} \stackrel{\mathsf{def}}{=} \mathrm{T}_{\mathsf{mix}}(1/4)$.

The constant $1/4$ is arbitrary, and can be chosen to be any constant less than $1/2$. We define it this way because $\mathrm{T}_{\mathsf{mix}}(\epsilon) \lesssim \mathrm{T}_{\mathsf{mix}} \cdot \log(1/\epsilon)$; see e.g. [LPW17]. A Markov chain whose mixing time is bounded by a polynomial in the size of the problem input is said to be *rapidly mixing* (or *fast mixing*); otherwise, it is *torpidly mixing* (or *slow mixing*). Of course, now the goal is to construct ergodic Markov chains which we can certify mixing time upper bounds. Note that one can define a version of mixing time with respect to any metric on probability measures (e.g. Wasserstein distance, KL-divergence, etc.) in the obvious way. To my knowledge, almost all analyses of Markov chain mixing times never directly study total variation distance; we typically replace it with a nicer, "smoother" distance measure. We'll see several examples throughout the course.

References

[DDS21] Daryl DeFord, Moon Duchin, and Justin Solomon. "Recombination: A Family of Markov Chains for Redistricting". In: *Harvard Data Science Review* 3.1 (2021) (cit. on p. 1).

[Dia09] Persi Diaconis. "The Markov Chain Monte Carlo Revolution". In: *Bull. Amer. Math. Soc.* 46.2 (2009) (cit. on p. 1).

[LPW17] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. 2nd ed. American Mathematical Society, 2017 (cit. on pp. 2, 5).

A Unfinished Proofs

Proof of Lemma 1.1. Since the fixed point equation $\mu\mathsf{P} = \mu$ is an eigenvalue problem, we first find some vector $v \in \mathbb{R}^\Omega$ such that $v\mathsf{P} = v$; this vector v might have negative entries, but we'll use this v to construct our distribution μ. Since P has rows summing to 1, $\mathsf{P}\mathbf{1} = \mathbf{1}$; in particular, P has eigenvalue 1. Since P and $\mathsf{P}^\top$ have the same eigenvalues (e.g. they have the same characteristic polynomials), it follows that $\mathsf{P}^\top$ has eigenvalue 1, i.e. there exists some $v \in \mathbb{R}^\Omega$ such that $v\mathsf{P} = v$.

Now define a distribution μ via $\mu(x) \propto |v(x)|$. We claim that $\mu\mathsf{P} = \mu$. To see this, observe that

$$\begin{aligned} |v(x)| &= \left| \sum_{y \in \Omega} v(y) \cdot \mathsf{P}(y \to x) \right| && \text{(Using } v\mathsf{P} = v\text{)} \\ &\le \sum_{y \in \Omega} |v(y)| \cdot \mathsf{P}(y \to x), \qquad \forall x \in \Omega. \end{aligned}$$

We claim that the above inequality must actually be an equality; if we can show this, then $\mu\mathsf{P} = \mu$ holds since we just scale both sides by the same constant $\sum_{x \in \Omega} |v(x)|$. To see this, observe that if the inequality is strict for any $x \in \Omega$, then summing over all $x \in \Omega$, we would obtain

$$\sum_{x \in \Omega} |v(x)| < \sum_{x \in \Omega} \sum_{y \in \Omega} |v(y)| \cdot \mathsf{P}(y \to x) = \sum_{y \in \Omega} |v(y)| \underbrace{\sum_{x \in \Omega} \mathsf{P}(y \to x)}_{=1} = \sum_{y \in \Omega} |v(y)|,$$

which is a contradiction. Hence, we must have $|v(x)| = \sum_{y \in \Omega} |v(y)| \cdot \mathsf{P}(y \to x)$ for all $x \in \Omega$, and we are done. □

Proof of Lemma 1.2. The key to the proof will be the simple inequality $\mathsf{P}^{s+\ell}(x, y) \ge \mathsf{P}^s(x, x) \cdot \mathsf{P}^\ell(x, y)$, which holds for all $x, y \in \Omega$ and all $s, \ell \in \mathbb{N}$. In particular, if we can guarantee that for every $x \in \Omega$, there exists some time $s^*(x)$ such that $\mathsf{P}^s(x, x) > 0$ for every $s \ge s^*(x)$, then letting $\ell^*(x, y)$ denote the first time ℓ such that $\mathsf{P}^\ell(x, y) > 0$, we have that for all $\ell \ge \ell^*(x, y)$,

$$\mathsf{P}^{s^*(x)+\ell}(x, y) = \mathsf{P}^{s^*(x)+\ell-\ell^*(x,y)}(x, x) \cdot \mathsf{P}^{\ell^*(x,y)}(x, y) > 0.$$

We could then take $t^* = \max_x s^*(x) + \max_{x,y} \ell^*(x, y)$.

To prove existence of this $s^*(x)$, the key is again that if $\mathsf{P}^j(x, x) > 0$ and $\mathsf{P}^k(x, x) > 0$, then $\mathsf{P}^{j+k}(x, x) > 0$ as well. Since P is aperiodic, there must exist $j, k \in \mathbb{N}$ such that $\mathsf{P}^j(x, x) > 0$, $\mathsf{P}^k(x, x) > 0$ and $\mathsf{gcd}(j, k) = 1$. The final number-theoretic claim is that $s^*(x) = \mathsf{lcm}(j, k) = jk$ suffices, which can be proved via Bézout's Lemma. □

Proof of Lemma 1.4. For every $x \in \Omega$,

$$\begin{aligned} (\mu\mathsf{P})(x) &= \sum_{y \in \Omega} \mu(y) \cdot \mathsf{P}(y \to x) \\ &= \sum_{y \in \Omega} \mu(x) \cdot \mathsf{P}(x \to y) && \text{(Reversibility)} \\ &= \mu(x) \sum_{y \in \Omega} \mathsf{P}(x \to y) \\ &= \mu(x). \end{aligned}$$

Since $x \in \Omega$ was arbitrary, $\mu\mathsf{P} = \mu$ as desired. □