Cornell University, Spring 2022 **CS 4850/5850**
Lecture notes: Markov chains **29 April – 6 May 2022**

1 Random Sampling and Markov Chain Monte Carlo

Markov chains model discrete-time random processes whose future state evolution depends only on the present state, not on the entire sequence of states leading up to the present. As such, they represent an important class of probabilistic models. However, in algorithm design they serve an important additional role: the most popular algorithmic procedure for sampling from complicated probability distributions is to design an appropriate Markov chain and simulate its state evolution. This method is known as *Markov Chain Monte Carlo (MCMC)*. In these notes we will present some aspects of the fundamental theory of Markov chains and of the MCMC paradigm for designing sampling algorithms.

Before delving into definitions, let us give some examples to illustrate what we mean by "sampling from complicated probability distributions."

Example 1.1. If G is a q-colorable graph then the uniform distribution on proper q-colorings of G is easy to define but potentially hard to sample. For example if $q \geq 3$ and G is allowed to be an arbitrary graph, it is NP-hard to decide if *any* q-coloring of G exists, let alone sample a uniformly random one.

Example 1.2. Generalizing the preceding example, given a graph G and two parameters β, γ, we may want to sample a random labeling of its vertices using labels in some set Σ, i.e. a random function $L : V(G) \to \Sigma$, with probability proportional to

$$w(L) = \prod_{(u,v)\in E(G)} \begin{cases} \beta & \text{if } L(u) = L(v) \\ \gamma & \text{if } L(u) \neq L(v). \end{cases}$$

The first example (sampling a random q-coloring) specializes this one by setting $|\Sigma| = q$, $\beta = 0$, $\gamma = 1$.

Example 1.3. Given a tuple of non-negative integers $(d_1, d_2, \ldots, d_n)$, consider the set of graphs with vertex set $[n] = \{1, 2, \ldots, n\}$ such that for all $i \in [n]$ the degree of vertex i is d_i. When this set is non-empty, one may wish to draw random samples from it. For example, sampling graphs from this distribution may be useful for simulating the performance of algorithms or distributed protocols on networks that resemble (in terms of their size and degree distribution) observed real-world network topologies. Alternatively, the ability to draw samples from this distribution may aid a statistician in testing the hypothesis that a network topology observed in the real world has some structure that is statistically distinguishable from random graphs with the same size and degree distribution.

Example 1.4. Suppose we are given:

1. a deep neural network (DNN) that generates random images by transforming an input layer of independent (Gaussian) random numbers into an output layer of pixels;

2. an image I with some missing pixels.

The DNN defines a probability distribution over output images (i.e., the distribution that results from feed-forward propagation of Gaussian random numbers at the input layer), and one may wish to draw samples from the conditional distribution over output images, conditioned on the pixel values matching the data present in I. For example, this sampling task may form part of the pipeline in an image completion algorithm: given a DNN that models natural scenes, and an image with a natural scene in the background and an object in the foreground that occludes part of the scene, the sampling algorithm could be used to generate hypothetical completions of the background image.

One can define the following class of algorithmic random sampling problems that includes all of the examples above, along with many other important and practical random sampling problems.

Definition 1.1. An *unnormalized distribution* on a finite set $\mathbf{X}$ is a function $w : \mathbf{X} \to \mathbb{R}_{\geq 0}$ such that

$$Z_w \triangleq \sum_{x \in \mathbf{X}} w(x) > 0.$$

The corresponding probability distribution is $p(x) = w(x)/Z_w$. Sampling from w refers to the process of drawing a random sample $x \in \mathbf{X}$ with probability $p(x) = w(x)/Z_w$. Approximately sampling from w refers to any process that draws a random sample x from $\mathbf{X}$ such that for all $A \subseteq \mathbf{X}$,

$$|\Pr(x \in A) - \sum_{y \in A} p(y)| \leq \varepsilon$$

for some specified approximation parameter $\varepsilon > 0$.

One can often specify an unnormalized distribution w by specifying an efficient algorithm to calculate $w(x)$ for every $x \in \mathbf{X}$. This brings us to the main question we address below.

> *Given an efficient algorithm for evaluating an unnormalized distribution $w(x)$, when is it possible to efficiently draw random samples from the probability distribution $p = w/Z_w$?*

Before continuing, let us pause to illustrate how the first and last examples above can be cast as special cases of this problem.

For the example of sampling a random q-coloring of a graph $G = (V, E)$, we can take $\mathbf{X}$ to be the set of all functions from V to $[q]$ (called "labelings" henceforth), and we can take w to be a function that assigns the value 1 to labelings that are proper colorings of G and 0 to all other labelings. Then the probability distribution p is the uniform distribution on proper colorings of G.

For the example of image completion, we can take $\mathbf{X}$ to be the set of all functions that label each node of the DNN with a number called the node's *activation*.[1] We can then define $w(x)$ to be zero if the node activations in x don't obey the DNN's weights and activation functions, or if the values in the output layer don't match the pixel values given in the input, I. However, when x does obey the DNN's weights and activation functions and matches the given pixel values in the output layer, we define $w(x)$ to be the product of the (Gaussian) probabilities of the input node activations. Then the distribution $p(x)$ is the conditional distribution defined in Example 1.4.

2 Markov chains and their stationary distributions

In this section we formally define Markov chains, introduce the notion of a stationary distribution, and identify conditions under which a Markov chain has a unique stationary distribution such that the marginal distribution of the time-t state is guaranteed to converge to the stationary distribution as $t \to \infty$.

Definition 2.1. A Markov chain with (finite) state set $\mathbf{X}$ is a probability distribution on infinite sequences $X_0, X_1, \dots$ of elements of $\mathbf{X}$, satisfying the Markov property:

$$\forall t > 0 \, \forall (x_0, x_1, \dots, x_t) \in \mathbf{X}^{t+1} \quad \Pr(X_t = x_t \mid X_0 = x_0, \dots, X_{t-1} = x_{t-1}) = \Pr(X_t = x_t \mid X_{t-1} = x_{t-1}).$$

In other words, the conditional distribution of X_t depends only on the value of X_{t-1} and not on any of the values that came before time $t-1$.

A Markov chain is *time-homogeneous* if for all pairs $(x, y) \in S^2$, and all $t > 0$,

$$\Pr(X_t = x | X_{t-1} = y) = \Pr(X_{t+1} = x | X_t = y).$$

For a time-homogeneous Markov chain, the matrix P defined by $P_{xy} = \Pr(X_t = y | X_{t-1} = x)$ is called the *transition matrix*.

For the remainder of these lecture notes, all the Markov chains we consider will be time-homogeneous. Accordingly, when we use the term *Markov chain* below it always implicitly refers to a time-homogeneous Markov chain.

The probability distribution of a Markov chain's state at time t can be represented by a row vector $\pi_t \in \mathbb{R}^{\mathbf{X}}$, whose x^{th} coordinate is the probability that $X_t = x$:

$$(\pi_t)_x = \Pr(X_t = x).$$

For $t > 0$ we can then calculate that

$$\begin{aligned} (\pi_t)_x = \Pr(X_t = x) &= \sum_{y \in \mathbf{X}} \Pr(X_t = x \wedge X_{t-1} = y) \\ &= \sum_{y \in \mathbf{X}} \Pr(X_t = x \mid X_{t-1} = y) \cdot \Pr(X_{t-1} = y) = \sum_{y \in \mathbf{X}} (\pi_{t-1})_y P_{yx} \end{aligned}$$

[1] Since our formalism requires $\mathbf{X}$ to be finite, we must quantize the set of numbers that can be used as a node's label. For example, we could limit the label set to be the set of 32-bit floating point numbers, or we could quantize node activations even more aggressively. Such quantization schemes have been advocated in the neural network literature, for the sake of making the training and inference process more efficient in terms of storage space, running time, and energy consumption.

This can be summarized more succinctly as

$$\pi_t = \pi_{t-1} P$$

and, by induction, we obtain

$$\pi_t = \pi_0 P^t.$$

Definition 2.2. A probability distribution π is a stationary distribution for a Markov chain with transition matrix P if it satisfies

$$\pi P = \pi.$$

A stationary distribution is thus a fixed point of the Markov chain's transition dynamics: if the initial state distribution π_0 is equal to the stationary distribution π, then every future state π_t is also distributed according to π.

It turns out that every Markov chain with finite state set has a stationary distribution. This fact, as well as a sufficient condition for the stationary distribution to be unique, can be deduced from the Perron-Frobenius Theorem, a fundamental theorem from linear algebra that concerns the eigenvalues of square matrices with non-negative entries.

Definition 2.3. If A is an $n \times n$ square matrix with non-negative entries, let G_A be the directed graph (potentially with self-loops) having vertex set $[n]$ and edge set $\{(i,j) | A_{ij} > 0\}$. We say A is *irreducible* if G_A is strongly connected, and we say A is *aperiodic* if the cycle lengths in G_A have no common divisor greater than 1.

Irreducible matrices are characterized by the property that every entry of $A + A^2 + A^3 + \ldots + A^n$ is strictly positive. Among irreducible matrices, the aperiodic ones are characterized by the property that for some positive integer k, every entry of A^k is strictly positive.

Theorem 2.1 (Perron-Frobenius). *If A is an irreducible $n \times n$ square matrix with non-negative entries, then A has a unique right eigenvector $v \in \mathbb{R}^n$ whose components are strictly positive. The eigenvalue associated to v, called the* Perron-Frobenius eigenvalue*, has multiplicity one, and every other (complex) eigenvalue λ' satisfies $|\lambda'| \leq \lambda$. This inequality is strict if A is aperiodic.*

The proof of the Perron-Frobenius Theorem can be found in many linear algebra textbooks, for example Felix Gantmacher's *The Theory of Matrices* (AMS Chelsea Publishing, 2000). For the sake of making these lecture notes self-contained, we will prove an easier result that pertains to Markov chain transition matrices.

Theorem 2.2. *If P is the transition matrix of an irreducible, aperiodic Markov chain with finite state set, then there is a unique stationary distribution π such that $\pi P = \pi$. For any starting distribution π_0, the time-t state distribution $\pi_t = \pi_0 P^t$ converges to π as $t \to \infty$. In fact, the convergence is exponentially fast: there are constants $C < \infty$ and $\delta > 0$ such that*

$$\|\pi_t - \pi\|_1 \leq C(1-\delta)^t$$

for all $t \in \mathbb{N}$.

Proof. Since P is irreducible and aperiodic, there exists some k such that all entries of P^k are positive. Let $N = |\mathbf{X}|$ denote the number of states of the Markov chain, and choose $\varepsilon < 0$ such that all entries of P^k are greater than or equal to ε/N. Let $Q = (\mathbf{1}\mathbf{1}^\top)/N$. Then

$$P^k = \varepsilon Q + (1-\varepsilon)R$$

where R is a non-negative matrix.

A *row-stochastic matrix* is a non-negative matrix whose row sums are all equal to 1. Equivalently, the non-negative matrix A is called row-stochastic if $A\mathbf{1} = \mathbf{1}$; from this characterization it is evident that the set of row-stochastic matrices is closed under multiplication. Note that Q is row-stochastic since $\mathbf{1}^\top\mathbf{1} = N$. Furthermore, P is row-stochastic since for every $x \in \mathbf{X}$ we have $\sum_y P_{xy} = \sum_{y \in \mathbf{X}} \Pr(X_t = y \mid X_{t-1} = x) = 1$. Hence P^k is row-stochastic, and we may conclude that R is also row-stochastic using the equation

$$(1-\varepsilon)R\mathbf{1} = P^k\mathbf{1} - \varepsilon Q\mathbf{1} = \mathbf{1} - \varepsilon\mathbf{1} = (1-\varepsilon)\mathbf{1}.$$

For $t \geq 0$ let $\Delta_t = \pi_{t+1} - \pi_t = \pi_0(P^{t+1} - P^t)$. We have

$$\Delta_t Q = \frac{1}{N}\pi_0(P^{t+1} - P^t)\mathbf{1}\mathbf{1}^\top = 0,$$

since $(P^{t+1} - P^t)ones = \mathbf{1} - \mathbf{1} = 0$. Therefore,

$$\Delta_{t+k} = \Delta_t P^k = (1-\varepsilon)\Delta_t R.$$

The inequality $\|vR\|_1 \leq \|v\|_1$ holds for any vector v. To prove this, it suffices to verify it when $\|v\|_1 \leq 1$. A vector whose 1-norm is less than or equal to 1 is a convex combination of the standard basis vectors and their negations, hence we only need to check that $\|vR\|_1 \leq 1$ when v is one of the standard basis vectors. In that case vR is a row of R, i.e. a non-negative vector whose components sum up to 1, so $\|vR\| = 1$. Now, using the inequality $\|vR\|_1 \leq \|v\|_1$, we find that

$$\|\Delta_{t+k}\|_1 \leq (1-\varepsilon)\|\Delta_t\|.$$

For any $t \in \mathbb{N}$, if $q = \lfloor t/k \rfloor$, then

$$\begin{aligned}
\sum_{s=t}^{\infty} \|\Delta_s\| &\leq \sum_{r=q}^{\infty}\sum_{i=0}^{k-1} \|\Delta_{kr+i} \\
&\leq \sum_{r=q}^{\infty}\sum_{i=0}^{k-1} (1-\varepsilon)^r \|\Delta_i\| \\
&= \frac{(1-\varepsilon)^q}{\varepsilon}\left(\|\Delta_0\|_1 + \cdots + \|\Delta_{k-1}\|_1\right)
\end{aligned}$$

This confirms that the sequence $\pi_t = \pi_0 + \sum_{s=0}^{t-1} \Delta_s$ converges absolutely as $t \to \infty$ and that the rate of convergence is exponential. Denote the limit point by π. To conclude the proof we must show that π is a stationary distribution of P. The equation $\pi P = \pi$ follows by observing that

$$\pi P = \lim_{t\to\infty}(\pi_t P) = \lim_{t\to\infty} \pi_{t+1} = \pi.$$

The fact that π is a probability distribution follows from the fact that π_t is a probability distribution for each t, and that the set of probability distributions on $\mathbb{R}^{\mathbf{X}}$ is topologically closed. □

3 Reversible Markov chains and the Metropolis-Hastings algorithm

In general, computing the stationary distribution of a Markov chain requires solving a linear system, but there is one case in which the stationary distribution has a simple closed-form formula. This is the case of a reversible Markov chain.

In this section, for an unnormalized distribution w, we will use the notations $w(x)$ and w_x interchangeably.

Definition 3.1. A Markov chain with transition matrix P is reversible with respect to (unnormalized) distribution w if it satisfies

$$w_x P_{xy} = w_y P_{yx}$$

for all $x, y \in \mathbf{X}$.

Lemma 3.1. *If P is reversible with respect to w, then $\pi = w/Z_w$, is a stationary distribution for P.*

Proof. Multiplying both sides of the reversibility equation $w_x P_{xy} = w_y P_{yx}$ by the normalizing constant $Z_w^{-1} = \left(\sum_x w_x\right)^{-1}$, we find that $\pi_x P_{xy} = \pi_y P_{yx}$ for all $x, y \in \mathbf{X}$. Hence,

$$(\pi P)_x = \sum_{y \in S} \pi_y P_{yx} = \sum_{y \in S} \pi_x P_{xy} = \pi_x \left(\sum_{y \in S} P_{xy} \right) = \pi_x.$$

□

The reversibility condition can be interpreted as a type of "detailed balance" condition: at stationarity, the rate of state transitions from x to y equals the rate of state transitions from y to x, for all state pairs x and y.

The Metropolis-Hastings algorithm is a procedure that takes an unnormalized distribution w and creates a Markov chain P whose state transitions are computationally easy to simulate, and whose stationary distribution is $\bar{w}$. Actually the procedure makes use of an auxiliary Markov chain K, called the *proposal distribution*, whose stationary distribution is simple and often unrelated to w. In many applications the stationary distribution of K is simply the uniform distribution on $\mathbf{X}$. To define the Metropolis-Hastings algorithm we assume we have:

1. An unnormalized probability distribution specified by a function $\kappa : \mathbf{X} \to [0, 1]$.

2. A Markov chain K that is reversible with respect to κ.

3. Algorithms for sampling state transitions of K and for computing the function κ.

The Markov chain K is called the *proposal distribution* for the Metropolis-Hastings procedure. As stated earlier, in many applications $\kappa(x) = 1$ for all x (i.e., the normalization of κ is the uniform distribution on $\mathbf{X}$) and the reversibility condition $\kappa_x K_{xy} = \kappa_y K_{yx}$ simply states that the Markov transition matrix K is a symmetric matrix.

Now for $x \neq y$ define

$$P_{xy} = K_{xy} \cdot \kappa_x \cdot \frac{\min\{w_x, w_y\}}{w_x}, \tag{1}$$

and define $P_{xx} = 1 - \sum_{y \neq x} P_{xy}$. Note that

$$\sum_{y \neq x} P_{xy} = \kappa_x \cdot \left(\sum_{y \neq x} K_{xy} \frac{\min\{w_x, w_y\}}{w_x} \right) \leq \kappa_x \cdot \left(\sum_{y \neq x} K_{xy} \right) \leq \kappa_x \leq 1,$$

so $P_{xx} \geq 0$. Thus, P is indeed a Markov transition matrix.

Lemma 3.2. *The Markov chain P defined by Equation (1) is reversible with respect to w.*

Proof. Consider any $x, y \in \mathbf{X}$. If $x = y$ then the equation $w_x P_{xy} = w_y P_{yx}$ holds trivially. Otherwise,

$$\begin{aligned} w_x P_{xy} &= K_{xy} \cdot \kappa_x \cdot \min\{w_x, w_y\} \\ w_y P_{yx} &= K_{yx} \cdot \kappa_y \cdot \min\{w_y, w_x\}. \end{aligned}$$

The lemma follows because $\min\{w_x, w_y\} = \min\{w_y, w_x\}$ and because our assumption that K is reversible with respect to κ implies $K_{xy}\kappa_x = K_{yx}\kappa_y$. □

An algorithm to simulate state transitions of the Markov chain P can be described as follows. Suppose the current state of the Markov chain is $x \in \mathbf{X}$.

1. Using the sampling oracle for Markov chain K, sample "proposed state" $y \in \mathbf{X}$ with probability K_{xy}.
2. Compute w_x, w_y, and κ_x.
3. With probability $\frac{\min w_x, w_y}{w_x} \cdot \kappa_x$, transition to state y.
4. Otherwise, remain at state x.

Example 3.1 (Glauber dynamics for sampling q-colorings)**.** To illustrate the Metropolis-Hastings procedure, we show how to use it to define a simple Markov chain whose unique stationary distribution is the uniform distribution over proper q-colorings of a graph $G = (V, E)$. For two labelings $x, y : V \to [q]$ define their Hamming distance as

$$d(x, y) = \#\{v \in V \mid x(v) \neq y(v)\}.$$

Assume that q is large enough that the graph whose vertices are proper q-colorings of G, and whose edges are pairs of colorings whose Hamming distance is 1, constitutes a non-empty connected graph. (If this graph is not connected, the Markov chain defined here will be reducible and it will have multiple stationary distributions.)

We will take $\kappa(x) = 1$ for all $x \in \mathbf{X}$, and for our proposal distribution we will define $n = |V|$ and

$$K_{xy} = \begin{cases} \frac{1}{nq} & \text{if } d(x,y) = 1 \\ \frac{1}{q} & \text{if } x = y \\ 0 & \text{otherwise.} \end{cases}$$

A state transition of K can be simulated by the following algorithm: starting from state x, sample vertex $v \in V$ and color $c \in [q]$ independently and uniformly at random, and let y be the state obtained from x by recoloring v with color c and leaving all other colors the same. From the definition of K it follows easily that $K_{xy} = K_{yx}$, i.e. K is reversible with respect to κ.

Recall that our goal is to design a Markov chain whose stationary distribution is the uniform distribution on proper q-colorings of G. In other words, we want to draw samples from the distribution given by the unnormalized density function w such that $w(x) = 1$ when x is a proper coloring and $w(x) = 0$ otherwise. To simulate a state transition of the Markov chain P defined by the Metropolis-Hastings procedure we do the following steps, starting from state x. Assume that x is a proper coloring.

1. *Sample "proposed state" $y \in \mathbf{X}$ with probability K_{xy}.*
   In other words, sample vertex $v \in V$ and color $c \in [q]$ independently and uniformly at random, and let y be the state obtained from x by recoloring v with color c and leaving all other colors the same.

2. *Compute w_x, w_y, and κ_x.*
   By assumption, x is a proper coloring, so $w_x = \kappa_x = 1$. Recall from above that $w_y = 1$ if and only if y is a proper coloring. Since x is a proper coloring and y is obtained from x by recoloring v, we only need to check whether every edge incident to v remains properly colored. In other words, to execute this step we merely need to test whether vertex v has any neighbor whose color is already c. If so, $w_y = 0$; otherwise, $w_y = 1$.

3. *With probability $\frac{\min\{w_x, w_y\}}{w_x} \cdot \kappa_x$, transition to state y.*
   The probability in question is 1 if the color of every neighbor of v is different from c, and 0 otherwise.

4. *Otherwise, remain at state x.*

Hence, the Metropolis-Hastings Algorithm in this case corresponds to the following very simple procedure. The starting state of the Markov chain is any proper coloring of G. To simulate one state transition, we sample a uniformly random vertex v and uniformly random color c, and we change the color of v to c if and only if the color of every neighbor of v is different from c. This Markov chain on the set of proper colorings of G is called *Glauber dynamics.*

4 Mixing time

The ability to efficiently simulate state transitions of a Markov chain whose stationary distribution is π doesn't necessarily imply the ability to efficiently draw samples from π, or from a distribution close to π. The reason is that the Markov chain might be *slowly mixing*: for small — or even moderately large — values of t, the state distribution after t steps, π_t, might be quite far from the eventual stationary distribution, π. Distance between distributions is often measured using the *total variation distance* (also known as statistical distance):

$$\|\pi - \pi'\|_{TV} = \max_{S \subset \mathbf{X}} \{|\pi(S) - \pi'(S)\} = \tfrac{1}{2}\|\pi - \pi'\|_1.$$

(The second equation can be confirmed by observing that the maximum of $|\pi(S) - \pi'(S)|$ is attained when $S = \{x \mid \pi(x) > \pi'(x)\}$.)

Theorem 4.1. *If the state transition matrix of a Markov chain is irreducible and aperiodic, with stationary distribution π, then for any initial distribution π_0 the total variation distance $\delta_t = \|\pi_t - \pi\|_{\mathsf{TV}}$ converges to zero exponentially fast: there exists $c < 1$ such that $\delta_t < c^t$ for all sufficiently large t.*

Proof. We have $\pi_t - \pi = (\pi_0 - \pi)P^t$. Note that $\pi_0 - \pi$ is orthogonal to $\mathbf{1}$. The linear subspace $W \subset \mathbb{R}^{\mathbf{X}}$ consisting of all row vectors orthogonal to $\mathbf{1}$ is preserved under right multiplication by P, since $w\mathbf{1} = 0$ implies

$$(wP)\mathbf{1} = w(P\mathbf{1}) = w\mathbf{1} = 0.$$

Since W is complementary to the eigenspace spanned by the Perron-Frobenius eigenvector of P, the eigenvalues of P acting on W are all the other eigenvalues of P other than 1. Since P is assumed to be irreducible and aperiodic, the Perron-Frobenius Theorem ensures that the absolute value of every such eigenvalue is strictly less than 1. Hence, for any $w \in W$, $\|wP^t\|_2$ decreases exponentially as $t \to \infty$. The ratio $\frac{\|wP^t\|_1}{\|wP^t\|_2}$ is bounded by the square-root of the dimension of $\mathbb{R}^{\mathbf{X}}$, so $\|wP^t\|_1$ also decreases exponentially as $t \to \infty$. □

Definition 4.1. For any $\varepsilon > 0$ and any irreducible Markov chain P, the ε-*mixing time* $\tau_P(\varepsilon)$ is defined to be the smallest t_0 such that for all initial state distributions π_0 and all $t \geq t_0$, the time-t state distribution $\pi_t = \pi_0 P^t$ satisfies $\|\pi_t - \pi\|_{TV} \leq \varepsilon$, where π denotes the stationary distribution of P.

Theorem 4.1 shows that when P is irreducible and aperiodic, the mixing time $\tau_P(\varepsilon)$ depends logarithmically on $1/\varepsilon$ as $\varepsilon \to 0$. On the other hand, since we are primarily interested in Markov chains whose state space $|\mathbf{X}|$ is exponentially large (i.e., exponential in the size of the problem description) it is usually very important to understand how $\tau_P(\varepsilon)$ depends on $|\mathbf{X}|$.

Definition 4.2. A Markov chain P is called *rapidly mixing* if its mixing time $\tau_P(\varepsilon)$ is bounded above by a polynomial function of $\log |\mathbf{X}/\varepsilon|$.

Determining which Markov chains are rapidly mixing and which ones aren't is a very active research area. In the following section we will present a very useful technique for proving rapid mixing of Markov chains.

5 Coupling

This section presents a method for bounding the mixing time of a Markov chain by "coupling" two parallel executions of the Markov chain that start from different states but converge toward occupying the same state as time progresses.

Definition 5.1. If π, π' are two probability distributions on a sample set $\mathbf{X}$, a *coupling* of π and π' is a probability measure ν on ordered pairs $(x, x') \in \mathbf{X} \times \mathbf{X}$ such that the marginal distribution of x is π and the marginal distribution of x' is π'. In other words, for every set $S \subseteq \mathbf{X}$,

$$\nu(S \times \mathbf{X}) = \pi(S), \qquad \nu(\mathbf{X} \times S) = \pi'(S).$$

The total variation distance has an important characterization in terms of coupling.

Lemma 5.1. $\|\pi - \pi'\|_{TV} = \inf\{\nu(x \neq x') \mid \nu$ *a coupling of* $\pi, \pi'\}$.

Proof. Let $\Delta = \{(x, x) \mid x \in \mathbf{X}\} \subseteq \mathbf{X} \times \mathbf{X}$, and let Δ^c denote the complementary set,

$$\Delta^c = \{(x, x') \mid x \neq x'\} \subset \mathbf{X} \times \mathbf{X}.$$

The probability denoted by $\nu(x \neq x')$ in the lemma statement can also (more accurately) be written as $\nu(\Delta^c)$. If ν is a coupling of π and π', then for every set $S \subseteq \mathbf{X}$,

$$\pi(S) - \pi'(S) = \nu(S \times \mathbf{X}) - \nu(\mathbf{X} \times S) \leq \nu(S \times (\mathbf{X} \setminus S)) \leq \nu(\Delta^c).$$

Since the inequality holds for every $S \subseteq \mathbf{X}$ and every coupling ν, it follows that

$$\sup_{S \subseteq \mathbf{X}} \|\pi(S) - \pi'(S)\| \leq \inf\{\nu(\Delta^c) \mid \nu \text{ a coupling of } \pi, \pi'\}.$$

The left side is $\|\pi - \pi'\|_{TV}$, so we have proven an inequality between the two sides of the equation asserted by the lemma. To prove the opposite inequality, we directly construct a coupling ν such that $\|\pi - \pi'\|_{TV} = \nu(\Delta^c)$. For this purpose, let $\delta = \|\pi - \pi'\|_{TV}$. If $\delta = 0$ then $\pi = \pi'$ and the coupling can simply be defined by setting $\nu(x, x) = \pi(x) = \pi'(x)$ for all $x \in \mathbf{X}$ and $\nu(x, x') = 0$ for $x \neq x'$. If $\delta > 0$ then for each $x \in \mathbf{X}$ let

$$\begin{aligned} \delta(x) &= (\pi(x) - \pi'(x))^+ = \max\{\pi(x) - \pi'(x), 0\} \\ \delta'(x) &= (\pi'(x) - \pi(x))^+ = \max\{\pi'(x) - \pi(x), 0\} \end{aligned}$$

and define

$$\nu(x, x') = \begin{cases} \min\{\pi(x), \pi'(x)\} & \text{if } x = x' \\ \delta^{-1} \cdot \delta(x) \cdot \delta'(x') & \text{if } x \neq x'. \end{cases}$$

If $S = \{x \mid \pi(x) > \pi'(x)\}$ then $\pi(S) - \pi'(S) = \|\pi - \pi'\|_{TV} = \delta$. This justifies the following identities.

$$\sum_{x \in \mathbf{X}} \delta(x) = \sum_{x : \pi(x) > \pi'(x)} (\pi(x) - \pi'(x)) = \pi(S) - \pi'(S) = \delta \tag{2}$$

$$\sum_{x' \in \mathbf{X}} \delta'(x') = \sum_{x : \pi'(x) \geq \pi(x)} (\pi'(x) - \pi(x)) = \pi'(\mathbf{X} \setminus S) - \pi(\mathbf{X} \setminus S) = \delta. \tag{3}$$

Using these identities we can see that ν is a coupling of π and π'.

$$\sum_{x' \in \mathbf{X}} \nu(x,x') = \min\{\pi(x), \pi'(x)\} + \sum_{x' \neq x} \delta^{-1} \cdot \delta(x) \cdot \delta'(x') = \min\{\pi(x), \pi'(x)\} + \delta^{-1} \cdot \delta(x) \cdot \sum_{x' \neq x} \delta'(x').$$

If $\pi(x) \leq \pi'(x)$ then $\min\{\pi(x), \pi'(x)\} = \pi(x)$ and $\delta(x) = 0$, so the right side equals $\pi(x)$ as required by the definition of a coupling. If $\pi(x) > \pi'(x)$ then $\delta'(x) = 0$, so the right side is equal to $\min\{\pi(x), \pi'(x)\} + \delta^{-1} \cdot \delta(x) \cdot \sum_{x' \in \mathbf{X}} \delta'(x')$. According to equation (2) the sum equals δ, so the entire right side is equal to $\min\{\pi(x), \pi'(x)\} + \delta(x)$, which equals $\pi(x)$. Thus, in either case, $\sum_{x' \in \mathbf{X}} \nu(x,x') = \pi(x)$ as required by the definition of coupling. The proof that $\sum_{x \in \mathbf{X}} \nu(x,x') = \pi'(x')$ follows similarly. Finally, to prove that $\nu(\Delta^c) = \delta$, we calculate

$$\nu(\Delta) = \sum_{x \in \mathbf{X}} \min\{\pi(x), \pi'(x)\} = \sum_{x \in \mathbf{X}} (\pi(x) - \delta(x)) = \sum_{x \in \mathbf{X}} \pi(x) - \sum_{x \in \mathbf{X}} \delta(x) = 1 - \delta$$

and subtract both sides of this equation from 1. □

A special case of coupling two probability distributions occurs when both of the probability distributions are Markov chains with the same transition matrix.

Definition 5.2. A *Markov coupling* with transition matrix P and initial state distributions π_0, π'_0 is a probability distribution over sequences of pairs $\{(X_t, X'_t) \mid t = 0, 1, \ldots\}$ such that:

1. The distributions of the random sequences $X_0, X_1, X_2, \ldots$ and $X'_0, X'_1, X'_2, \ldots$ are both Markov chains with transition matrix P.
2. The distribution of X_0 is π_0, and the distribution of X'_0 is π'_0.

Although each of the random state sequences $X_0, X_1, \ldots$ and $X'_0, X'_1, \ldots$ in a Markov coupling must evolve according to the transition matrix P, they may use shared randomness to evolve in a correlated way. In particular, by constructing Markov couplings in which X_t and X'_t tend to become more similar over time, we can bound mixing times of Markov chains.

Lemma 5.2 (Markov Coupling Lemma). *Let P be a Markov transition matrix with stationary distribution π. For any $t_0 \in \mathbb{N}$ and $\varepsilon > 0$, the mixing time bound $\tau_P(\varepsilon) \leq t_0$ is implied by the following sufficient condition: every initial state distribution π_0 has a Markov coupling with transition matrix P and initial state distributions π_0, π, satisfying* $\Pr(X_t \neq X'_t) \leq \varepsilon$ *for all $t \geq t_0$.*

Proof. Let $\pi = \pi'_0$ be the stationary distribution of P. Since X'_0 is distributed according to π and π is stationary for P, the distribution of X'_t must be equal to π for every $t > 0$ as well. Letting π_t denote the distribution of X_t, we find that the joint distribution of the pair (X_t, X'_t) is a coupling of π_t with π. Lemma 5.1 now implies that $\|\pi_t - \pi\|_{TV} \leq \varepsilon$ for all $t \geq t_0$, hence $\tau_P(\varepsilon) \leq t_0$. □

5.1 Analyzing Glauber Dynamics via Coupling

Recall the Glauber dynamics for sampling a uniformly random q-coloring of an undirected graph G. This is the Markov chain whose states are proper colorings of G, and whose transition dynamics are described by the following sampling process: in state $x : V(G) \to [q]$, sample a uniformly random vertex v and color c, and let $y : V(G) \to [q]$ be the function defined by setting

$$y(u) = \begin{cases} c & \text{if } u = v \\ x(u)\text{if } u \neq v. \end{cases}$$

If y is a proper coloring then transition from x to y, otherwise remain in state x.

In this section we will prove Glauber dynamics mixes rapidly when $q > 4\Delta$, where Δ is the maximum degree of a vertex of G. There is a long-standing conjecture that Glauber dynamics mixes rapidly whenever $q > \Delta + 1$; at present, however, the best known result in this direction asserts that Glauber dynamics mixed rapidly whenever $\frac{q-1}{\Delta} > \alpha$, where $\alpha \approx 1.763\ldots$ is the solution to the equation $e^{1/x} = x$.

To analyze Glauber dynamics we will use the Markov Coupling Lemma. The construction of the Markov coupling is very simple to describe. Starting from states X_0 and X'_0 sampled from some arbitrary initial distribution π_0 and from the stationary distribution, respectively, we repeatedly update the pair of states by choosing the same vertex v and color c in both Markov chains. To bound the probability of the event $X_t \neq X'_t$, we will analyze the Hamming distance

$$d(X_t, X'_t) = \#\{v \mid X_t(v) \neq X'_t(v)\}.$$

How does the Hamming distance change when both sides of the coupling undergo a Markov transition corresponding to choosing vertex v and color c?

1. If $X_t(v) \neq X'_t(v)$, and $X_{t+1}(v) = X'_{t+1}(v) = c$, then the Hamming distance increases by 1. Let us call this event a *color merge*.

2. If $X_t(v) = X'_t(v)$ but $X_{t+1}(v) \neq X'_{t+1}(v)$, then the Hamming distance decreases by 1. We will call this event a *color split*. A color split occurs when v is recolored with color c on one side of the coupling, but on the other side the recoloring doesn't take place because a neighbor of v is already colored with c.

3. In all other cases, the Hamming distance is unchanged.

Let $d_t = d(X_t, X'_t)$. To estimate the probability of a color merge, observe that the probability of sampling a vertex v such that $X_t(v) \neq X'_t(v)$ is d_t/n, and when such a vertex v is sampled, a color merge takes place unless we sample a color c which is among the colors of v's neighbors in X_t or X'_t. Since v has Δ or fewer neighbors, there are at least $q - 2\Delta$ colors that are not used by v's neighbors in either X_t or X'_t. Hence, the probability of a color merge is at least:

$$\Pr(\text{color merge}) \geq \frac{d_t}{n} \cdot \frac{q - 2\Delta}{q}.$$

Now let's estimate the probability of a color split. In order for such an event to take place, v must have a neighbor w such that $X_t(w) = c$ and $X'_t(w) \neq c$ or $X_t(w) \neq c$ and $X'_t(w) = c$.

When this happens, we will say that the color split is *blamed on* the directed edge (v, w). Every color split can be blamed on at least one directed edge, possibly more than one. Now, in order for a directed edge (v, w) to be blamed for a color split, w must be among the d_t vertices whose colors in X_t and X'_t differ, c must be one of the two elements of the set $\{X_t(w), X'_t(w)\}$, and v must be one of the (at most) Δ neighbors of w, so

$$\Pr(\text{color split}) \leq \mathbb{E}[\text{number of blamed edges}] \leq d_t \cdot \frac{2}{q} \cdot \frac{\Delta}{n} = \frac{d_t}{n} \cdot \frac{2\Delta}{q}.$$

Combining these two bounds, we find that

$$\begin{aligned}
\mathbb{E}[d_{t+1} \mid d_t] &= d_t - \Pr(\text{color merge}) + \Pr(\text{color split}) \\
&\leq d_t - \frac{d_t}{n} \cdot \frac{q - 2\Delta}{q} + \frac{d_t}{n} \cdot \frac{2\Delta}{q} \\
&= \left(1 - \frac{q - 4\Delta}{qn}\right) \cdot d_t.
\end{aligned}$$

By induction on t,

$$\mathbb{E}[d_t] \leq \left(1 - \frac{q - 4\Delta}{qn}\right)^t \cdot d_0 < \exp\left(-\frac{q - 4\Delta}{qn} \cdot t\right) \cdot n. \tag{4}$$

When $t \geq \frac{q}{q-4\Delta} \cdot n \ln(n/\varepsilon)$, the right side of (4) is less than or equal to ε. Hence, by Lemma 5.2, the ε-mixing time of Glauber dynamics is bounded above by $\frac{q}{q-4\Delta} \cdot n \ln(n/\varepsilon)$.