Massachusetts Institute of Technology
Department of Electrical Engineering and Computer Science
6.438 Algorithms for Inference
Fall 2020

24 Markov Chain Monte Carlo Methods

In our exploration of approximate inference, we now turn from deterministic methods to stochastic methods, which are also referred to as *Monte Carlo* methods. Our objective will be to produce samples from the distributions. As seen in the lecture on computational equivalence of inference tasks, if we can produce samples then we can approximate virtually any quantity of interest.

Our focus will be on a technique for sampling from a distribution without knowing its partition function. (Of course, we saw that if one can compute partition functions then one can sample, and vice versa...) The approach we will develop is rooted in the concept of rejection sampling, which we describe next.

24.1 Rejection Sampling

Suppose we want to sample from a distribution

$$p(x) = \frac{\tilde{p}(x)}{Z}, \qquad x \in \mathcal{X}$$

where $\mathcal{X}$ is sufficently large that Z is computationally infeasible to compute. But suppose have some other distribution $q(x)$ over the same alphabet that we can sample from; q is referred to as the *proposal distribution*. Associated with this distribution is a constant c such that

$$\tilde{p}(x) \le cq(x), \qquad \text{all } x \in \mathcal{X}. \tag{1}$$

Then we can obtain samples from p as follows:

1. Generate a sample x' from $q(\cdot)$
2. Generate a sample u of a Bernoulli random variable u on $\mathcal{U} = \{\mathrm{A}, \mathrm{R}\}$ from x' according to $p_{\mathsf{u}|\mathsf{x}'}(\cdot|x')$, where

$$p_{\mathsf{u}|\mathsf{x}'}(\mathrm{A}|x') = \frac{\tilde{p}(x')}{c\, q(x')}. \tag{2}$$

3. If $u = \mathrm{A}$, the let $x = x'$. Otherwise discard x'.
4. Repeat the above steps until the desired number of samples are obtained.

This method of sampling from p is referred to as *rejection sampling*. To verify that it produces samples from the correct distribution, we note that

$$\begin{aligned}\mathbb{P}\left(\mathsf{x} = x\right) &= \mathbb{P}\left(\mathsf{x}' = x \mid \mathsf{u} = \mathrm{A}\right)\\ &= p_{\mathsf{x}'|\mathsf{u}}(x|\mathrm{A})\\ &\propto p_{\mathsf{u}|\mathsf{x}'}(\mathrm{A}|x')\, q(x')\\ &= \frac{\tilde{p}(x')}{c\, q(x')} q(x')\\ &\propto p(x).\end{aligned}$$

The acceptance rate, i.e., the fraction of samples from q that are retained as samples from p, is, in turn,

$$\begin{aligned}\mathbb{P}\left(\mathsf{u} = \mathrm{A}\right) &= p_{\mathsf{u}}(\mathrm{A})\\ &= \sum_{x' \in \mathcal{X}} p_{\mathsf{u}|\mathsf{x}'}(\mathrm{A}|x')\, q(x')\\ &= \sum_{x' \in \mathcal{X}} \frac{\tilde{p}(x')}{c\, q(x')} q(x')\\ &= \sum_{x' \in \mathcal{X}} \frac{Z p(x')}{c}\\ &= Z/c,\end{aligned}$$

from which we see that when the value of c for which (1) is satisfied is large, rejection sampling is inefficient. This is the typical behavior when $\mathcal{X}$ is large, i.e., precisely when we wanted to use rejection sampling since it wasn't feasible to compute Z.

Nevertheless, while this form of rejection sampling is not useful, by applying the concept in a different way, so that (1) is imposed locally instead of globally, it is possible to avoid the inefficiency. The key is to create a Markov chain whose stationary distribution is p via a rejection sampling approach. Our treatment is necessarily very cursory; for those who are interested, there is a large literature on the topic worth exploring after this introduction.

24.2 Markov Chain Monte Carlo

Suppose, we are interested in sampling from $p_{\mathsf{x}}(\mathbf{x})$, but we only know p_{x} up to a multiplicative constant (i.e. $p_{\mathsf{x}}(\mathbf{x}) = \tilde{p}_{\mathsf{x}}(\mathbf{x})/Z$ and we can calculate $\tilde{p}_{\mathsf{x}}(\mathbf{x})$).

Our approach will be to construct a Markov chain $\mathbf{P}$ whose stationary distribution π is equal to p_{x} while only using $\tilde{p}_{\mathsf{x}}$ in our construction. Once we have created the Markov chain, we can start from an arbitrary $\mathbf{x}$, run the Markov chain until it converges to π and we will have a sample from p_{x}. Such an approach is called a *Markov Chain Monte Carlo (MCMC)* method. It is necessary to answer:

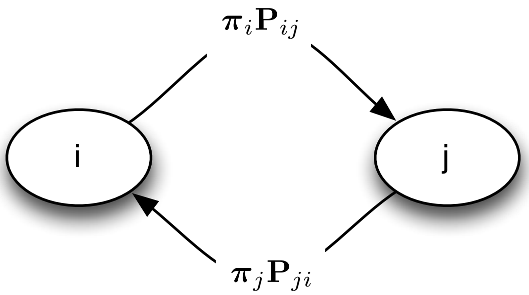
$\boldsymbol{\pi}_i\mathbf{P}_{ij}$
i
j
$\boldsymbol{\pi}_j\mathbf{P}_{ji}$

Figure 1: Detailed balance says the $\pi_i P_{ij} = \pi_j P_{ji}$, in other words the probability flowing from i to j is the same as the flow from j to i.

1. How to construct such a Markov chain $\mathbf{P}$?
2. How long it takes for the Markov chain to converge to its stationary distribution?

The Metropolis-Hastings (MH) algorithm provides one answer the first question. We will then make some remarks about the second question, and define the notion of *mixing time* that makes the answer precise.

24.2.1 Metropolis-Hastings

First, let us introduce some notation. Let $\mathbf{x} \in \mathcal{X}$, and we will assume that $|\mathcal{X}| = L < \infty$. For example, if $\mathbf{x}$ were a binary vector of length 10, then $\mathcal{X} = \{0,1\}^{10}$. We will construct a Markov chain which we will represent as a matrix $\mathbf{P} = [P_{ij}]$ where the (i,j) entry corresponds to the probability of transitioning from state $i \in \mathcal{X}$ to state $j \in \mathcal{X}$. We want the stationary distribution of $\mathbf{P}$, denoted by a vector $\boldsymbol{\pi} = [\pi_i]$, to be equal to $p_{\mathbf{x}}$ (i.e. $\pi_i = p_{\mathbf{x}}(i)$). (Recall that π is stationary for $\mathbf{P}$ if $\sum_j \pi_j P_{ji} = \pi_i$.)

If $\mathbf{P}$ is irreducible and aperiodic, then it has a unique stationary distribution to which the chain converges from any starting point. Such a chain is often called *ergodic*. There is a situation in which it is very easy to check that a given π is indeed stationary, and that is what is known as a *reversible* Markov chain. We will design our Markov chain $\mathbf{P}$ to be reversible.

Definition 1 (Reversible Markov chain)**.** *A Markov Chain* $\mathbf{P}$ *is called reversible with respect to* $\boldsymbol{\pi}$ *if it satisfies*

$$\pi_i P_{ij} = \pi_j P_{ji} \quad \textit{for all } i \neq j.$$

This equation is also referred to as detailed balance.

Intuitively, detailed balance says that the probability “flowing” from i to j is the same amount of probability “flowing” from j to i, where by probability “flow” from i to j we mean $\pi_i P_{ij}$ (i.e. the probability of being in i and transitioning to j).

Importantly, if $\mathbf{P}$ is reversible with respect to $\boldsymbol{\pi}$ and we did not assume $\boldsymbol{\pi}$ was the stationary distribution, detailed balance implies that $\boldsymbol{\pi}$ is a stationary distribution because

$$\sum_j \pi_j P_{ji} = \sum_j \pi_i P_{ij} = \pi_i \left(\sum_j P_{ij} \right) = \pi_i.$$

So showing that $\mathbf{P}$ satisfies the detailed balance equation with $\boldsymbol{\pi}$ is one way of showing the $\boldsymbol{\pi}$ is a stationary distribution of $\mathbf{P}$.

To obtain such a $\mathbf{P}$, we will start with a "proposed" Markov chain $\mathbf{Q}$ which we will modify to create $\mathbf{P}$. As we will see, the conditions that $\mathbf{Q}$ must satisfy are very mild and $\mathbf{Q}$ may have little or no relation to p_{x}. Again, we will represent $\mathbf{Q}$ as a matrix $[Q_{ij}]$ and we require that

$Q_{ii} > 0$ for all $i \in \mathcal{X}$ and

$\mathcal{G}(\mathbf{Q}) = (\mathcal{X}, \mathcal{E}(\mathbf{Q}))$ is connected where $\mathcal{E}(\mathbf{Q}) \triangleq \{(i,j) : Q_{ij}Q_{ji} > 0\}$.

In other words, all self-transitions must be possible and it must be possible to move from any state to another state in some number of transitions. The chain $\mathbf{Q}$ is readily seen to be aperiodic (due to the self-transitions) and irreducible.

Now define[1]

$$A_{ij} \triangleq \min\left(1, \frac{\tilde{p}_{\mathsf{x}}(j)Q_{ji}}{\tilde{p}_{\mathsf{x}}(i)Q_{ij}}\right) = \min\left(1, \frac{p_{\mathsf{x}}(j)Q_{ji}}{p_{\mathsf{x}}(i)Q_{ij}}\right)$$

and

$$P_{ij} \triangleq \begin{cases} Q_{ij}A_{ij} & j \neq i \\ 1 - \sum_{j \neq i} P_{ij} & j = i \end{cases}.$$

This is the *Metropolis-Hastings* Markov chain $\mathbf{P}$ that we were after. You should convince yourself that $\mathbf{P}$ inherits from $\mathbf{Q}$ that it is aperiodic and irreducible[2]. It remains to show that p_{x} is the stationary distribution of $\mathbf{P}$, and as we commented above it suffices to show that $\mathbf{P}$ satisfies the detailed balance equation with p_{x}.

Lemma 1. *For all $i, j \in \mathcal{X}$, detailed balance $p_{\mathsf{x}}(i)P_{ij} = p_{\mathsf{x}}(j)P_{ji}$ holds and hence p_{x} is the stationary distribution of* $\mathbf{P}$.

Proof. For $(i,j) \notin \mathcal{E}(\mathbf{Q})$ this is trivially true since $P_{ij} = P_{ji} = 0$. For $(i,j) \in \mathcal{E}(\mathbf{Q})$, without loss of generality assume $\tilde{p}_{\mathsf{x}}(j)Q_{ji} \geq \tilde{p}_{\mathsf{x}}(i)Q_{ij}$. This implies that

$$A_{ij} = 1 \quad \text{and} \quad A_{ji} = \frac{\tilde{p}_{\mathsf{x}}(i)Q_{ij}}{\tilde{p}_{\mathsf{x}}(j)Q_{ji}}.$$

[1] The equality holds because Z cancels out. Also, the astute reader will notice that the ratio in A_{ij} is directly related to the detailed balance equation.

[2] If $p(i) = 0$ for some state i, then there are no transitions to i and hence i is a transient state; one restricts attention to recurrent states, which can be seen to form a single class.

Then

$$\begin{aligned}
p_{\mathsf{x}}(i)P_{ij} &= p_{\mathsf{x}}(i)Q_{ij} \\
&= p_{\mathsf{x}}(i)Q_{ij}\frac{p_{\mathsf{x}}(j)Q_{ji}}{p_{\mathsf{x}}(j)Q_{ji}} \\
&= \left(\frac{p_{\mathsf{x}}(i)Q_{ij}}{p_{\mathsf{x}}(j)Q_{ji}}\right)Q_{ji}p_{\mathsf{x}}(j) \\
&= A_{ji}Q_{ji}p_{\mathsf{x}}(j) = P_{ji}p_{\mathsf{x}}(j)\,,
\end{aligned}$$

which shows that **P** is reversible with respect to p_{x} and completes the proof. □

The Markov transition matrix **P** describes a process as follows: starting from state i, to generate state j:

> Generate *proposed state* j' according to **Q** with current state i.
>
> Flip a biased coin with $\mathbb{P}\left(\text{heads}\right) = A_{ij'}$.
>
> > If heads, then *accept* the new state $j = j'$.
> >
> > If tails, then *reject* and set the new state $j = i$, the old state.

This gives a convenient description of **P**, which can easily be implemented in code. The number A_{ij} is commonly referred to as the *acceptance probability* because it describes the probability of accepting the proposed new state j'.

Intuitively, we can think of Metropolis-Hastings as forcing **Q** to be reversible in a specific way. Given an arbitrary **Q**, there's no reason that $p_{\mathsf{x}}(i)Q_{ij}$ will be equal to $p_{\mathsf{x}}(j)Q_{ji}$, in other words, the probability flow from i to j will not necessarily be equal to the flow from j to i. To fix this, we could scale the flow on one side to be equal to the other and that is what Metropolis-Hastings does. To make this notion more precise, let $\mathcal{M}^*(p_{\mathsf{x}})$ be the space of all reversible Markov chains that have p_{x} as a stationary distribution. Then the Metropolis-Hastings algorithm takes **Q** and gives us $\mathbf{P} \in \mathcal{M}^*(p_{\mathsf{x}})$ that satisfies

Theorem 1.

$$\mathbf{P} = \operatorname*{arg\,min}_{\mathbf{R}\in\mathcal{M}^*(p_{\mathsf{x}})} \sum_i p_{\mathsf{x}}(i) \sum_{j\neq i} |Q_{ij} - R_{ij}|.$$

Hence, **P** *is the* ℓ_1*-projection of* **Q** *on* $\mathcal{M}^*(p_{\mathsf{x}})$*.*

24.3 Exploiting Structure

We will describe a simple example of using Metropolis-Hastings to sample from an MRF. Suppose we have $x_1, \dots, x_N$ binary with

$$p(\mathbf{x}) \propto \exp\underbrace{\left(\sum_{i\in\mathcal{V}} \phi_i(x_i) + \sum_{(i,j)\in\mathcal{E}} \psi_{ij}(x_i, x_j)\right)}_{\triangleq -E(\mathbf{x})},$$

for some graph $\mathcal{G}$. Here $\mathcal{X} = \{0,1\}^n$, so it has 2^n elements. Suppose we have $\mathbf{Q} = [\frac{1}{2^n}]$ the matrix with all entries equal to $\frac{1}{2^n}$, that is, the probability of transitioning from i to j is equally probable for all j. Then the Metropolis-Hastings algorithm would give

$$\begin{aligned} P_{ij} &= Q_{ij} \min\left(1, \frac{\exp(E(j))}{\exp(E(i))}\right) \\ &= \frac{1}{2^n} \min\left(1, \exp(E(j) - E(i))\right). \end{aligned}$$

Is there any downside to choosing such a simple $\mathbf{Q}$? If i has moderate probability, the chance of randomly choosing a j that has higher probability is very low, so we are very unlikely to transition away from i. Thus it may take a long time for the Markov chain to reach its stationary distribution.

24.3.1 Gibbs Sampling

Metropolis-Hastings is a general purpose sampling procedure. However, we would also like to be able to directly exploit factorization structure in the target distribution corresponding to an undirected graphical model $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. In this section, we develop *Gibbs sampling*, a special case of Metropolis-Hastings that can do this.

Let $|\mathcal{V}| = N$ so the states are of the form $x_\mathcal{V} = (x_1, \dots, x_N)$ with $x_i \in \mathcal{X}$, $i \in \mathcal{V}$, and $\mathbf{Q} = [Q_{a_\mathcal{V}, b_\mathcal{V}}]$ is defined as follows:

$$Q_{a_\mathcal{V}, b_\mathcal{V}} = \begin{cases} p(b_k | b_{\mathcal{V}\setminus\{k\}})/N & \text{if } a_\mathcal{V} \text{ and } b_\mathcal{V} \text{ differ only in the } k\text{th element, } k \in \mathcal{V} \\ 0 & \text{otherwise} \end{cases} \tag{3a}$$

where

$$p(b_k | b_{\mathcal{V}\setminus\{k\}}) = \frac{p(b_\mathcal{V})}{\sum\limits_{b_k} p(b_\mathcal{V})} = \frac{\tilde{p}(b_\mathcal{V})}{\sum\limits_{b_k} \tilde{p}(b_\mathcal{V})}. \tag{3b}$$

We can interpret the chain (3) in terms of the following implementation: given a current state $x_\mathcal{V}$, we obtain the next state $x'_\mathcal{V}$ via

1. Select $k \in \mathcal{V}$ from a uniform distribution.

2. Set $x'_{\mathcal{V}\setminus\{k\}} = x_{\mathcal{V}\setminus\{k\}}$ and generate a sample x'_k from $p(\cdot|x_{\mathcal{V}\setminus\{k\}})$.

Note that for sparse undirected graphical models, the conditional independencies generally simplify $p(\cdot|x_{\mathcal{V}\setminus\{k\}})$ so that conditioning on only a few variables is required. An example is a grid graph, where the variable at a node will be conditionally independent of the remaining variables in the graph given its immediate neighbors. In such cases, sampling from $p(\cdot|x_{\mathcal{V}\setminus\{k\}})$ can be quite easy. More specifically, given a factorization

$$p(x_{\mathcal{V}}) = \frac{1}{Z} \prod_{\mathcal{C} \in \mathrm{cl}^*(\mathcal{G})} \psi_{\mathcal{C}}(x_{\mathcal{C}}),$$

we can partition the maximal cliques $\mathrm{cl}^*(\mathcal{G})$ into those that involve node k, which we denote using $\mathrm{cl}^*{}_k(\mathcal{G})$, and those that do not, which we denote using $\mathrm{cl}^*{}_{\setminus k}(\mathcal{G})$, so

$$\mathrm{cl}^*(\mathcal{G}) = \mathrm{cl}^*{}_k(\mathcal{G}) \cup \mathrm{cl}^*{}_{\setminus k}(\mathcal{G}). \tag{4}$$

Then

$$\begin{aligned}
p(x_k|x_{\mathcal{V}\setminus\{k\}}) &= \frac{(1/Z) \displaystyle\prod_{\mathcal{C} \in \mathrm{cl}^*(\mathcal{G})} \psi_{\mathcal{C}}(x_{\mathcal{C}})}{\displaystyle\sum_{x_k} (1/Z) \prod_{\mathcal{C} \in \mathrm{cl}^*(\mathcal{G})} \psi_{\mathcal{C}}(x_{\mathcal{C}})} \\
&= \frac{\displaystyle\prod_{\mathcal{C} \in \mathrm{cl}^*{}_k(\mathcal{G})} \psi_{\mathcal{C}}(x_{\mathcal{C}}) \prod_{\mathcal{C} \in \mathrm{cl}^*{}_{\setminus k}(\mathcal{G})} \psi_{\mathcal{C}}(x_{\mathcal{C}})}{\displaystyle\prod_{\mathcal{C} \in \mathrm{cl}^*{}_{\setminus k}(\mathcal{G})} \psi_{\mathcal{C}}(x_{\mathcal{C}}) \sum_{x_k} \prod_{\mathcal{C} \in \mathrm{cl}^*{}_k(\mathcal{G})} \psi_{\mathcal{C}}(x_{\mathcal{C}})} \\
&= \frac{\displaystyle\prod_{\mathcal{C} \in \mathrm{cl}^*{}_k(\mathcal{G})} \psi_{\mathcal{C}}(x_{\mathcal{C}})}{\displaystyle\sum_{x_k} \prod_{\mathcal{C} \in \mathrm{cl}^*{}_k(\mathcal{G})} \psi_{\mathcal{C}}(x_{\mathcal{C}})},
\end{aligned}$$

and since $\mathrm{cl}^*{}_k(\mathcal{G})$ is typically quite small, this computation is typically simple. We emphasize that the partition function does not enter into the computation.

If $p > 0$ then it follows that the graph for $\mathbf{Q}$ is connected and all self-transitions are possible. We will now show that $\mathbf{Q}$ satisfies detailed balance, so the acceptance probability will always be 1, hence the Metropolis-Hastings transition matrix $\mathbf{P}$ will be equal to $\mathbf{Q}$.

Lemma 2. *$\mathbf{Q}$ satisfies detailed balance with respect to p.*

Proof. For $x_{\mathcal{V}}$ and $x'_{\mathcal{V}}$, we must show detailed balance with respect to p, i.e., that

$$p(x_{\mathcal{V}})\, Q_{x_{\mathcal{V}} x'_{\mathcal{V}}} = p(x'_{\mathcal{V}})\, Q_{x'_{\mathcal{V}} x_{\mathcal{V}}}.$$

Suppose $x_{\mathcal{V}} \neq x'_{\mathcal{V}}$, and suppose they do not differ by exactly one position. Then according to (3), $Q_{x_{\mathcal{V}} x'_{\mathcal{V}}} = Q_{x'_{\mathcal{V}} x_{\mathcal{V}}} = 0$, so the equation is satisfied trivially. Now suppose $x_{\mathcal{V}} \neq x'_{\mathcal{V}}$ and they differ in exactly one position, which we denote by k. Then,

$$\begin{aligned} p(x_{\mathcal{V}})\, Q_{x_{\mathcal{V}} x'_{\mathcal{V}}} &= \frac{1}{N}\, p(x_{\mathcal{V}})\, p(x'_k | x_{\mathcal{V}\setminus\{k\}}) \\ &= \frac{1}{N} \big[p(x_k | x_{\mathcal{V}\setminus\{k\}})\, p(x_{\mathcal{V}\setminus\{k\}})\big] p(x'_k | x_{\mathcal{V}\setminus\{k\}}) \\ &= \frac{1}{N} \big[p(x_k | x'_{\mathcal{V}\setminus\{k\}})\, p(x'_{\mathcal{V}\setminus\{k\}})\big] p(x'_k | x'_{\mathcal{V}\setminus\{k\}}) \qquad (5) \\ &= \frac{1}{N}\, p(x_k | x'_{\mathcal{V}\setminus\{k\}}) \big[p(x'_{\mathcal{V}\setminus\{k\}})\, p(x'_k | x'_{\mathcal{V}\setminus\{k\}})\big] \\ &= \frac{1}{N}\, p(x_k | x'_{\mathcal{V}\setminus\{k\}})\, p(x'_{\mathcal{V}}) \\ &= p(x'_{\mathcal{V}})\, Q_{x'_{\mathcal{V}} x_{\mathcal{V}}}, \end{aligned}$$

where to obtain (5) we have used that $x_{\mathcal{V}\setminus\{k\}} = x'_{\mathcal{V}\setminus\{k\}}$. □

In practice, Gibbs sampling works well and in many cases it is simple to implement. Explicitly, we start from an arbitrary initial state $x^0_{\mathcal{V}}$ and generate putative samples $x^1_{\mathcal{V}}, x^2_{\mathcal{V}}, \ldots$ according to the following process: for $t = 0, 1, \ldots$

1. Select $i \in \mathcal{V}$ uniformly.
2. Set $x^{t+1}_{\mathcal{V}\setminus\{i\}} = x^t_{\mathcal{V}\setminus\{i\}}$ and sample x^{t+1}_i from $p(\cdot | x^t_{\mathcal{V}\setminus\{i\}})$.

All of the caveats regarding the Markov chain Monte Carlo apply: most importantly, we need to run the Markov chain until it has reached its stationary distribution, so we need to toss out a number of initial samples in a process called "burn-in". In the continuation of these notes, we will describe a way to analyze the time it takes the Markov chain to reach its stationary distribution, but in practice people rely on heuristics. Other forms of Gibbs sampling exist, such as *block* Gibbs sampling as explored in the homework, but their full development is beyond our scope.

24.4 Mixing Time

Now that we have constructed the Markov chain, we can sample from it, but we want samples from the stationary distribution. We turn to the second question: how long does it take for the Markov chain to converge to its stationary distribution?

For simplicity of exposition, we will focus on a generic reversible Markov chain $\mathbf{P}$ with state space $\mathcal{X}$ and unique stationary distribution $\boldsymbol{\pi}$. We will also assume that $\mathbf{P}$ *regular*, that is $\mathbf{P}^k > 0$ for some $k > 0$ (i.e., each entry of $\mathbf{P}^k$ is positive). This means that for some $k > 0$, it is possible to transition from any $i \in \mathcal{X}$ to any $j \in \mathcal{X}$ in exactly k steps. Additionally, we will assume that $\mathbf{P}$ is a *lazy* Markov chain, which means that $P_{ii} > 0$ for all $i \in \mathcal{X}$. This is a mild condition because we can take any Markov chain $\mathbf{Q}$ and turn it into a lazy Markov chain without changing its stationary distribution by considering $\frac{1}{2}(\mathbf{Q}+\mathbf{I})$. The lazy condition ensures that all of the eigenvalues of $\mathbf{P}$ are positive and it does not substantially increase the mixing time.

24.4.1 Total Variation Distance

We are interested in measuring the time it takes $\mathbf{P}$ to go from any initial state to its stationary distribution. The first question to address is how we will measure the distance to stationarity. The appropriate notion of distance is known as the total variation distance.

Definition 2 (Total Variation)**.** *Given two probability measures μ, ν defined on the same space Ω, we define*

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

The total variation distance is one half of the ℓ_1 norm of the difference of the measures. It is therefore a metric: it is nonnegative, satisfies the triangle inequality, and is equal to zero if and only if $\mu = \nu$.

Importantly, if $f : \Omega \to \mathbf{R}$, then one can check that

$$|\mathbf{E}_\mu f(X) - \mathbf{E}_\nu f(X)| \le \|\mu - \nu\|_{\mathrm{TV}} \sup_x |f(x)|\,.$$

What this means is that if we are ϵ-close to stationarity in total variation for a small ϵ, then virtually any quantity we might like to approximate (such as marginal probabilities) will have good accuracy.

Lemma 3. *The total variation distance has the alternative characterization*

$$\|\mu - \nu\|_{\mathrm{TV}} = \sup_{A\subseteq\Omega} \mu(A) - \nu(A)\,.$$

Proof. Let $A = \{x : \mu(x) \ge \nu(x)\}$. Then

$$\mu(A) - \nu(A) = \sum_{x\in A} \big(\mu(x) - \nu(x)\big) = \sum_{x\in A} |\mu(x) - \nu(x)|\,.$$

But

$$0 = \sum_{x\in\Omega} \big(\mu(x) - \nu(x)\big) = \sum_{x\in A} |\mu(x) - \nu(x)| - \sum_{x\in A^c} |\mu(x) - \nu(x)|$$

and hence in the prior displayed equation we may replace the sum over A to over Ω and include a factor two. Since this is a specific set A, we get $\sup_{A\subseteq\Omega} \mu(A) - \nu(A) \ge \frac{1}{2}\sum_{x\in\Omega} |\mu(x) - \nu(x)|$.

To see the other direction, observe that for any set B,

$$\begin{aligned}\mu(B)-\nu(B) &= \frac{1}{2}\big(\mu(B)-\nu(B)+(\nu(B^c)-\mu(B^c))\big)\\ &\leq \frac{1}{2}\big(|\mu(B)-\nu(B)|+|\nu(B^c)-\mu(B^c)|\big)\\ &\leq \frac{1}{2}\sum_{x\in\Omega}|\mu(x)-\nu(x)|\end{aligned}$$

by the triangle inequality. □

24.4.2 Mixing Time

Definition 3 (ϵ-mixing time of $\mathbf{P}$)**.** *Given $\epsilon > 0$, $T_{\mathrm{mix}}(\epsilon)$ is the smallest time such that for $t \geq T_{mix}(\epsilon)$*

$$\|\boldsymbol{\mu}\mathbf{P}^t - \boldsymbol{\pi}\|_{\mathrm{TV}} \leq \epsilon,$$

for any initial distribution $\boldsymbol{\mu}$.

We will consider two different methods for obtaining bounds on $T_{mix}(\epsilon)$: (1) coupling, and (2) spectral. The first method is entirely probabilistic, and involves constructing copies of the chain defined on the same space that can be glued together. In the second method, we will focus our attention how what the operation of multiplying with $\mathbf{P}$ does. Recall that as we apply $\mathbf{P}$ to any vector, the eigenvector with the largest eigenvalue dominates and how quickly it dominates is determined by the second largest eigenvalue. We will exploit this intuition to get a bound on $T_{\mathrm{mix}}(\epsilon)$ that depends on the difference between the largest and second largest eigenvalues.

24.5 Coupling

A coupling is a joint distribution over random variables and provides a bound on total variation distance.

Definition 4 (Coupling)**.** *Suppose $X \sim \mu$ and $Y \sim \nu$. A coupling is a random variable (X', Y') such that the marginals agree with the original distributions, i.e., $X' \overset{d}{=} X$ and $Y' \overset{d}{=} Y$.*

Lemma 4. *Given $X \sim \mu$, $Y \sim \nu$,*

$$\|\mu - \nu\|_{\mathrm{TV}} = \inf_{(X,Y)} \mathbb{P}(X \neq Y)\,.$$

Here the infimum is over all couplings (X, Y).

Proof. We will only show that $\|\mu-\nu\|_{\text{TV}} \leq \mathbb{P}(X \neq Y)$ for any coupling, which suffices for our purposes. We write

$$\begin{aligned}\mathbb{P}(X \neq Y) &= \sum_x \Big(\mathbb{P}(X=x) - \mathbb{P}(X=x, Y=x)\Big)\\ &\geq \sum_x \Big(\mathbb{P}(X=x) - \min\{\mathbb{P}(X=x), \mathbb{P}(Y=x)\}\Big)\\ &= \sum_x \max\{\mathbb{P}(X=x) - \mathbb{P}(Y=x), 0\}\\ &= \frac{1}{2}\sum_x |\mathbb{P}(X=x) - \mathbb{P}(Y=x)|\,.\end{aligned}$$

To see the last step, suppose that f is a function with $\sum_x f(x) = 0$. Then

$$\sum_{x:f(x)>0} f(x) + \sum_{x:f(x)<0} f(x) = 0$$

which implies that

$$\sum_{x:f(x)>0} |f(x)| = \sum_{x:f(x)<0} |f(x)|\,. \qquad \square$$

The total variation distance is a minimum over all couplings. The following example shows an instance of a coupling that achieves this minimum.

Example 1. Suppose $X \sim \text{Ber}(p)$ and $Y \sim \text{Ber}(q)$. We will construct a coupling on the pair of random variables (X, Y) as follows. Define

$$\begin{aligned}U &\sim \text{Unif}[0,1]\\ X &= \mathbb{1}_{U\leq p}\\ Y &= \mathbb{1}_{U\leq q}\,.\end{aligned}$$

Use the following rule to determine the values of X and Y: Sample U. If $U \geq p$, both X and Y are 0. If $U \leq q$, both X and Y are 1. X and Y only differ when $q \leq U \leq p$. This is shown graphically in Figure 2. Thus, we see that the probability that $X \neq Y$ is equal to $|p-q|$. This is precisely the total variation distance between p and q:

$$\|\mu-\nu\|_{\text{TV}} = \mathbb{P}(X \neq Y) = |p-q|\,.$$

Coupling is a way to relate two random variables by describing a joint distribution over them. The simplest coupling is assuming independence of the random variables. This is a valid upper bound on the total variation distance. For instance, in the example above, we could define the joint distribution of X and Y as the product of two independent distributions:

$$(X, Y) \sim \text{Ber}(p) \times \text{Ber}(q)$$

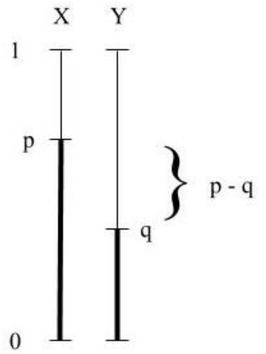

Figure 2: The shaded bar with height p represents the probability of X being 1. Similarly, the shaded bar with height q represents the probability of Y being 1. Observe that both X and Y are 1 if our uniform random variable $U > p$ and both X and Y are 0 if $U < q$. X and Y only differ if U falls within a range of size $|p - q|$.

In this case

$$\mathbb{P}(X \neq Y) = p(1-q) + q(1-p) > |p - q| \,.$$

Corollary 1. *Let X^t, Y^t be copies of the same Markov Chain $\mathbf{P}$ with stationary distribution μ. Let $Y^0 \sim \mu$ and $X^0 = x$. Then*

$$\|\mu_x - \nu\|_{\mathrm{TV}} \leq \mathbb{P}(X^t \neq Y^t) \,.$$

Building upon the above corollary, suppose we have $X^1, X^2, \ldots$ evolving according to $\mathbf{P}$ and $Y^1, Y^2, \ldots$ also evolving according to $\mathbf{P}$. When constructing a coupling, the goal is to evolve the Markov Chain in such a way that $X^t = Y^t$ as soon as possible (equivalently, minimize $\mathbb{P}(X^t \neq Y^t)$). If at some time $\tau, X^\tau = Y^\tau$, then we can just make the same transitions for all future times so that $X^t = Y^t$ for all $t \geq \tau$. At the same time, we must make sure that the marginal processes X and Y are each valid copies of the Markov chain (i.e., evolve according to $\mathbf{P}$).

24.5.1 Bound on Mixing Time

Theorem 2. *Consider the Glauber dynamics for an Ising Model on a graph with n vertices and a maximum degree $\leq \Delta$. Let $C(\beta) = 1 - \Delta \tanh(\beta)$. If $\Delta \tanh(\beta) < 1$ then*

$$T_{\mathrm{mix}}(\epsilon) \leq \frac{n(\log n + \log \frac{1}{\epsilon})}{C(\beta)}$$

If $\Delta \tanh(\beta) > 1$, then mixing can be exponentially slow.

This theorem states that in time $O(n \log n)$, the distribution will be close to stationary. We can see that the $n \log n$ term comes from the coupon collector problem.

It is a classical result that $O(n \log n)$ draws are needed in order to obtain each of n coupons when they are drawn with replacement. Additionally, because $\tanh \beta < \beta$, note that the condition $\beta < \frac{1}{\Delta}$ suffices in place of $\Delta \tanh(\beta) < 1$.

Our goal when defining a coupling is to obtain the highest probability that two random variables are equal at some time t when both are evolving according to Markov Chain updates. Recall that the mixing time can be defined as

$$T_{\text{mix}}(\epsilon) = \max_{x_0} \min\{t : \|\mu_x^t - \pi\|_{\text{TV}} \leq \epsilon\}$$

for stationary distribution π. Note that once the total variation drops below ϵ it stays due to the monotonicity

$$\|\mu^{(t+1)} - \pi\|_{\text{TV}} \leq \|\mu^{(t)} - \pi\|_{\text{TV}} \,.$$

This holds due to the so-called "data processing inequality" for f-divergences (total variation is an f-divergence). It is a worthwhile exercise to go ahead and prove this from first principles, noting that $\|\mu^{(t+1)} - \pi\|_{\text{TV}} = \|(\mu^{(t)} - \pi)\mathbf{P}\|_{\text{TV}}$ and then using properties of $\mathbf{P}$.

24.5.2 Glauber Dynamics and Mixing Time

Suppose we have a distribution p_{x} on our sample space $\Omega = \mathcal{X}^n$. We define our Markov Chain updates as follows.

At each time step $t \geq 1$:

1. Choose $i \in [n]$ uniformly at random.
2. Update $X_i^{t+1} \sim p_{\mathsf{x}_i}(\cdot | x_{\mathcal{V}\setminus\{i\}})$. For all nodes $j \neq i, X_j^{t+1} = X_j^t$.

Note that instead of choosing our nodes uniformly at random, we could iterate through the nodes sequentially. This is often used in ML applications. As well, we could modify this Markov Chain to perform block updates. Rather than updating a single node at a time, we update an entire block of nodes conditional on the rest of the graph. During each block update, we run updates for multiple time-steps over the entire block of nodes.

We now illustrate the process of choosing a coupling and bounding the mixing time.

Example 2. Suppose we want to sample from the uniform distribution on $\{0,1\}^n$, $\pi \sim \text{Unif}(\{0,1\}^n)$. The Glauber Dynamics chain is as follows.

At each time step $t \geq 1$:

1. Choose $i \in [n]$ u.a.r.
2. Update $X_i^{t+1} \sim \text{Ber}(\frac{1}{2})$

Here, we update a random coordinate at each time-step. We can think of this as having a vector of n variables and randomizing one at a time. Now, we want to find the mixing time.

Let us define a coupling. Let $Y^0 \sim \pi$. Let $X^0 = x = (0, 0, ..., 0)$ (by symmetry we might as well consider an arbitrary fixed x). After running a few updates of our Markov chain, we see that X will still be mostly zeroes. Intuitively, this seems far from a uniform distribution – we would expect approximately half 1 and half 0.

We want to come up with a coupling for our two Markov Chains X^t and Y^t such that each chain evolves according to our Markov Chain update outlined above. However, we have freedom in defining the relationship between the updates for X and Y. For instance, rather than randomly choosing a coordinate to update for X and then again randomly choosing a coordinate to update for Y, we could first choose a random coordinate and then update both X and Y at that same coordinate. Note that this gives the correct transition probabilities for each of the individual chains.

Define a coupling for (X^t, Y^t) as follows. At each time step $t \geq 1$:

1. Choose the same $i \in [n]$ u.a.r.

2. Update $X_i^{t+1} = Y_i^{t+1} \sim \text{Ber}(\frac{1}{2})$. Leave rest unchanged.

Intuitively, we expect that for large enough t, we will hopefully have updated all of the coordinates and thus $X^t = Y^t$.

In order to analyze the mixing time, we first recognize that this problem is an instance of the coupon collector problem. Repeating the standard bound for the coupon collector problem, we can develop an upper bound on the mixing time. Suppose that $t = \alpha n \log n$. Then

$$\begin{aligned}
\mathbb{P}(X^t \neq Y^t) &\leq \mathbb{P}(\text{haven't updated all coordinates by time } t) \\
&\leq n \cdot \mathbb{P}(\text{don't update coordinate 1})^t \\
&= n \cdot \left(1 - \frac{1}{n}\right)^t \\
&= n \cdot \left(1 - \frac{1}{n}\right)^{\alpha n \log n} \\
&\leq n \cdot e^{-\alpha \log n} \\
&= \frac{1}{n^{\alpha - 1}} \,.
\end{aligned}$$

The second inequality is due to a union bound over the n coordinates. If we take $\alpha = 2$, for instance, then we obtain $T_{\text{mix}}(1/n) \leq 2n \log n$.

24.5.3 Path Coupling

The idea of path coupling is to only define a coupling for pairs of states that differ in a single coordinate. Here we develop the idea of path coupling for the Ising model, but the framework holds for other models as well.

Consider a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$. To begin, recall that the Ising model (with uniform edge potentials and no external fields) is defined as

$$\pi(x) = \frac{1}{Z} \exp(\beta \sum_{ij \in \mathcal{E}} x_i x_j)$$

for $x \in \{-1, +1\}^n$. We omit the nodewise terms for simplicity here. Recall that the conditional update probabilities are

$$\pi(x_i = +1 | x_{\mathcal{V} \setminus \{i\}}) = \frac{e^{\beta \sum_{j \in \partial i} x_j}}{e^{\beta \sum_{j \in \partial i} x_j} + e^{-\beta \sum_{j \in \partial i} x_j}}$$

The Hamming Distance is defined as

$$\rho(\sigma, \tau) = \frac{1}{2} \sum_{u \in \mathcal{V}} |\sigma(u) - \tau(u)|$$

where σ and τ are two configurations: ρ is the number of coordinates in which they differ.

Proposition 1 (Path Coupling)**.** *Suppose that for all x, y such that $\rho(x, y) = 1$ there exists a coupling with $X^0 = x$, $Y^0 = y$, $X^1 \sim P(x, \cdot)$, $Y^1 \sim P(y, \cdot)$, such that $\mathbb{E}[(\rho(X^1, Y^1)] \leq e^{-\alpha} \rho(x, y)$. Then*

$$t_{\text{mix}}(\epsilon) \leq \alpha^{-1} \log \frac{n}{\epsilon}$$

where $\log n$ *comes from the log of the diameter of the state space.*

Note that we only define a coupling for the single next step of the Markov chain for states that differ by a single coordinate. This is often a lot simpler than attempting to devise a coupling for any two states of the Markov chain.

Proof. We want to show that, with as high probability as possible, the Hamming distance between X^t and Y^t is 0. Using iterated expectation,

$$\mathbb{E}[\rho(X^t, Y^t)] = \mathbb{E}\big[\mathbb{E}[\rho(X^t, Y^t) | X^{t-1}, Y^{t-1}\big] \, .$$

As illustrated in Figure 3, define

$$Z_0^{t-1} = X^{t-1} \quad \text{and} \quad Z_\rho^{t-1} = Y^{t-1} \, ,$$

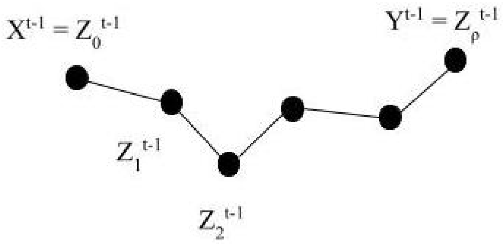

Figure 3: We define $Z_0^{t-1} := X^{t-1}$ and $Z_\rho^{t-1} := Y^{t-1}$. The Hamming distance between X and Y is the sum of Hamming distances along a shortest path between X and Y, and Z_i^{t-1} for $i = 1, \ldots, \rho - 1$ are states along this path.

as well as all the intermediate Z_i^{t-1} for $i = 1, \ldots, \rho - 1$ where Z_{i-1}^{t-1} and Z_i^{t-1} differ in just one coordinate. We assumed in the statement of the proposition that there is a coupling between each Z_{i-1}^{t-1} and Z_i^{t-1}, and suppose that it can be extended jointly to all of them. (In our finite state space this is merely the statement that if we have a distribution $p(u, v)$ and $p(v, w)$ then we can extend this to $p(u, v, w) = p(u, v)p(w|v)$ with the same pairwise marginals.) To define the coupling between X^t and Y^t we *define* $X^t = Z_0^t$ and $Y^t = Z_\rho^t$. The triangle inequality now implies that

$$\rho(X^t, Y^t) = \rho(Z_0^t, Z_\rho^t) \leq \sum_{i=1}^{\rho} \rho(Z_{i-1}^t, Z_i^t)\,,$$

hence

$$\begin{aligned}\mathbb{E}\rho(X^t, Y^t) &= \mathbb{E}\big[\mathbb{E}[\rho(X^t, Y^t)|X^{t-1}, Y^{t-1}]\big] \\ &\leq \mathbb{E}\left[\mathbb{E}\left[\sum_{i=1}^{\rho(X^{t-1}, Y^{t-1})} \rho(Z_{i-1}^t, Z_i^t)|X^{t-1}, Y^{t-1}\right]\right].\end{aligned}$$

By the assumed contraction on $\rho(Z_{i-1}^{t-1}, Z_i^{t-1})$ under one step of the coupled dynamics, we have that

$$\mathbb{E}[\rho(X^t, Y^t)] \leq \mathbb{E}[\rho(X^{t-1}, Y^{t-1})e^{-\alpha}]$$

(This last step requires a bit of care to justify formally.) If we continue iterating,

$$\begin{aligned}\mathbb{E}[\rho(X^t, Y^t)] &\leq e^{-\alpha t}\rho(X^0, Y^0) \\ &\leq e^{-\alpha t} n\,.\end{aligned}$$

Putting this together, we obtain

$$\begin{aligned}\mathbb{P}(X^t \neq Y^t) &\leq \mathbb{E}[\rho(X^t, Y^t)] \\ &\leq e^{-\alpha t} n \\ &\leq \varepsilon, \text{ for } t \geq \alpha^{-1} \log \frac{n}{\varepsilon}\end{aligned} \qquad \square$$

Here we illustrate path coupling for the Ising model Glauber dynamics.

Example 3. Because we only need to couple two states differing in a single coordinate, we consider states σ, τ where $-1 = \sigma(v) \neq \tau(v) = +1$ and $\sigma(u) = \tau(u)$ for $u \neq v$.

We define our coupling as follows. Pick a node $w \in \mathcal{V}$ u.a.r. Define

$$\begin{aligned}\pi(\sigma, w) &= \mathbb{P}(X_w = +1 | X_{\mathcal{V}\backslash w} = \sigma_{\mathcal{V}\backslash w}) \\ \pi(\tau, w) &= \mathbb{P}(Y_w = +1 | Y_{\mathcal{V}\backslash w} = \tau_{\mathcal{V}\backslash w})\end{aligned}$$

Let $\mathtt{U} \sim \text{Unif}[0, 1]$.
Set

$$X_w = \begin{cases} +1 & \text{if } \mathtt{U} \leq \pi(\sigma, w) \\ -1 & \text{if } \mathtt{U} > \pi(\sigma, w) \end{cases}$$

and

$$Y_w = \begin{cases} +1 & \text{if } \mathtt{U} \leq \pi(\tau, w) \\ -1 & \text{if } \mathtt{U} > \pi(\tau, w) \end{cases}$$

Now we have three cases:

1. If $w = v$, then $\rho(X, Y) = 0$.

2. If $w \notin (\partial v \cup \{v\})$, then all neighbors agree and $\rho(X, Y) = 1$.

3. If $w \in \partial v$ and $\pi(\sigma, w) < \mathtt{U} \leq \pi(\tau, , w)$, then $\rho(X, Y) = 2$.

Case 1 holds because X and Y only differ at node v, so the neighborhood of node v is identical for both distributions. Case 3 holds because the probability of updating w is different in the σ state and in the τ state and if $\mathtt{U}$ lands in this interval the updates differ.

Now we can write the expected value of the Hamming distance between X and Y obtained by one step update from states σ and τ:

$$\mathbb{E}_{\sigma,\tau}[\rho(X, Y)] \leq 1 - \frac{1}{n} + \frac{1}{n} \sum_{w \in \partial v} (\pi(\tau, w) - \pi(\sigma, w))$$

Let $S = \sum_{u \in \partial w} \sigma(u)$ and $\sum_{u \in \partial w} \tau(u) = 2 + S$. Now we plug these values into the expected value above. First, we can upper bound the term in the sum.

$$\begin{aligned}
\pi(\tau, w) - \pi(\sigma, w) &= \frac{1}{2}[\tanh(\beta(S+2)) - \tanh(\beta S)] \\
&\leq \tanh \beta
\end{aligned}$$

All this allows us to determine the amount of contraction (i.e., α in the statement of the path coupling proposition):

$$\begin{aligned}
\mathbb{E}_{\sigma,\tau}[\rho(X,Y)] &\leq 1 - \frac{1}{n} + \frac{1}{n} \sum_{w \in \partial v} \pi(\tau, w) - \pi(\sigma, w) \\
&\leq 1 - \frac{1}{n} + \frac{\Delta \tanh \beta}{n} \\
&= 1 - \frac{1}{n}(1 - \Delta \tanh \beta) \\
&\leq e^{-\frac{1}{n}(1 - \Delta \tanh \beta)}
\end{aligned}$$

From the last step, we can plug this into our mixing time bound in the Path Coupling proposition in order to get an upper bound on the mixing time for this model and prove the claim.

24.6 Mixing times, eigenvalues, and conductance (optional)

The following is technical and is included for completeness. First, we will bound the total variation by a term that does not depend on the initial distribution $\boldsymbol{\mu}$.

$$\begin{aligned}
\sum_i |(\boldsymbol{\mu}\mathbf{P}^t)_i - \pi_i| &= \sum_i \left| \sum_j \mu_j [\mathbf{P}^t]_{ji} - \pi_i \right| \\
&= \sum_i \left| \sum_j \mu_j ([\mathbf{P}^t]_{ji} - \pi_i) \right| \\
&\leq \sum_{ij} \boldsymbol{\mu}_j \left| [\mathbf{P}^t]_{ji} - \pi_i \right| \\
&= \sum_j \boldsymbol{\mu}_j \sum_i \left| [\mathbf{P}^t]_{ji} - \pi_i \right| \\
&\leq \|\boldsymbol{\mu}\|_1 \max_j \sum_i \left| [\mathbf{P}^t]_{ji} - \pi_i \right| \\
&= \max_j \sum_i \left| [\mathbf{P}^t]_{ji} - \pi_i \right| .
\end{aligned}$$

where used Hölder's inequality. Now we show how to bound $\sum_i |[\mathbf{P}^t]_{ji} - \pi_i|$ for every $j \in \mathcal{X}$. Using the Cauchy-Schwarz inequality we obtain

$$\begin{aligned}\sum_i \big|[\mathbf{P}^t]_{ji} - \pi_i\big| &= \sum_i \left|\frac{[\mathbf{P}^t]_{ij}}{\pi_i} - 1\right| \sqrt{\pi_i}\sqrt{\pi_i} \\ &\leq \left[\left(\sum_i \left|\frac{[\mathbf{P}^t]_{ij}}{\pi_i} - 1\right|^2 \pi_i\right)\left(\sum_i \pi_i\right)\right]^{1/2},\end{aligned}$$

which after some algebraic manipulation simplifies to

$$\begin{aligned}\left[\left(\sum_i \left|\frac{[\mathbf{P}^t]_{ij}}{\pi_i} - 1\right|^2 \pi_i\right)\left(\sum_i \pi_i\right)\right]^{1/2} &= \left(\sum_i \left(1 - \frac{2[\mathbf{P}^t]_{ji}}{\pi_i} + \frac{[\mathbf{P}^t]_{ji}^2}{\pi_i^2}\right)\pi_i\right)^{1/2} \\ &= \left(\sum_i \frac{[\mathbf{P}^t]_{ji}^2}{\pi_i} - 1\right)^{1/2}.\end{aligned}$$

Now we will use the reversibility of $\mathbf{P}$,

$$\begin{aligned}\left(\sum_i \frac{[\mathbf{P}^t]_{ji}^2}{\pi_i} - 1\right)^{1/2} = \left(\sum_i \frac{[\mathbf{P}^t]_{ji}[\mathbf{P}^t]_{ji}\pi_j}{\pi_i\pi_j} - 1\right)^{1/2} &= \left(\sum_i \frac{[\mathbf{P}^t]_{ji}[\mathbf{P}^t]_{ij}\pi_i}{\pi_i\pi_j} - 1\right)^{1/2} \\ &= \left(\sum_i \frac{[\mathbf{P}^t]_{ji}[\mathbf{P}^t]_{ij}}{\pi_j} - 1\right)^{1/2} \\ &= \left(\frac{[\mathbf{P}^{2t}]_{jj}}{\pi_j} - 1\right)^{1/2},\end{aligned}$$

where we used the reversibility of $\mathbf{P}$ to exchange $\pi_j[\mathbf{P}^t]_{ji}$ for $\pi_i[\mathbf{P}^t]_{ij}$. Putting this together, we conclude that

$$\sum_i \big|[\mathbf{P}^t]_{ji} - \pi_i\big| \leq \left(\frac{[\mathbf{P}^{2t}]_{jj}}{\pi_j} - 1\right)^{1/2}.$$

So we need to find a bound on the diagonal entries of $\mathbf{P}^{2t}$. Consider the following matrix

$$\mathbf{M} \triangleq \operatorname{diag}(\sqrt{\boldsymbol{\pi}})\mathbf{P}\operatorname{diag}(\sqrt{\boldsymbol{\pi}}^{-1})$$

where $\operatorname{diag}(\sqrt{\boldsymbol{\pi}})$ is the diagonal matrix with $\sqrt{\boldsymbol{\pi}}$ along the diagonal. Note that $\mathbf{M}$ is symmetric because

$$\begin{aligned}M_{ij} &= \sqrt{\frac{\pi_i}{\pi_j}}P_{ij} = \frac{\pi_i P_{ij}}{\sqrt{\pi_i\pi_j}} = \frac{\pi_j P_{ji}}{\sqrt{\pi_i\pi_j}} \\ &= \sqrt{\frac{\pi_j}{\pi_i}}P_{ji} = M_{ji},\end{aligned}$$

using the reversibility of $\mathbf{P}$. Recall the *spectral theorem* from linear algebra which says that for any real symmetric $L \times L$ matrix $\mathbf{A}$, there exist orthonormal eigenvectors $\mathbf{v}_1, \dots, \mathbf{v}_L$ with eigenvalues $\lambda_1 \geq \cdots \geq \lambda_L$ such that

$$\mathbf{A} = \sum_{k=1}^{L} \lambda_k \mathbf{v}_k \mathbf{v}_k^{\mathrm{T}} = \mathbf{V} \operatorname{diag}(\boldsymbol{\lambda}) \mathbf{V}^{\mathrm{T}},$$

where $\mathbf{V}$ is the matrix with $\mathbf{v}_1, \dots, \mathbf{v}_L$ as columns and $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_L)$.

Decomposing $\mathbf{M}$ in this way shows that

$$\mathbf{M}^t = (\mathbf{V} \operatorname{diag}(\boldsymbol{\lambda}) \mathbf{V}^{\mathrm{T}})^t = \mathbf{V} \operatorname{diag}(\boldsymbol{\lambda})^t \mathbf{V}^{\mathrm{T}}$$

because $\mathbf{V}$ is orthogonal. Hence

$$\begin{aligned}\mathbf{P}^t &= \operatorname{diag}(\sqrt{\boldsymbol{\pi}}^{-1}) \mathbf{M}^t \operatorname{diag}(\sqrt{\boldsymbol{\pi}}) \\ &= \operatorname{diag}(\sqrt{\boldsymbol{\pi}}^{-1}) \mathbf{V} \operatorname{diag}(\boldsymbol{\lambda})^t \mathbf{V}^{\mathrm{T}} \operatorname{diag}(\sqrt{\boldsymbol{\pi}}) \\ &= \operatorname{diag}(\sqrt{\boldsymbol{\pi}}^{-1}) \mathbf{V} \operatorname{diag}(\boldsymbol{\lambda})^t \mathbf{V}^{\mathrm{T}} \operatorname{diag}(\sqrt{\boldsymbol{\pi}}).\end{aligned}$$

From this representation of $\mathbf{P}^t$, we conclude that

$$[\mathbf{P}^t]_{jj} = \sum_i \lambda_i^t (\mathbf{v}_i)_j^2.$$

It can be shown by the Perron-Frobenius theorem that $\lambda_1 = 1$ and $|\lambda_k| < 1$ for $k < 1$ and the first eigenvector of $\mathbf{P}$ is $\boldsymbol{\pi}$. $\mathbf{M}$ is similar[3] to $\mathbf{P}$ hence they have the same eigenvalues and their eigenspaces have the same dimensions. By construction, an eigenvector $\mathbf{u}$ of $\mathbf{P}$ implies that $\mathbf{u} \operatorname{diag}(\sqrt{\boldsymbol{\pi}}^{-1})$ is an eigenvector of $\mathbf{M}$ with the same eigenvalue. Thus $\boldsymbol{\pi} \operatorname{diag}(\sqrt{\boldsymbol{\pi}}^{-1})$ is eigenvector of $\mathbf{M}$ and because the eigenspace corresponding to the eigenvalue 1 has dimension 1, $\mathbf{v}_1 \propto \boldsymbol{\pi} \operatorname{diag}(\sqrt{\boldsymbol{\pi}}^{-1})$. Furthermore, $\|\mathbf{v}_1\|_2 = 1$ because $\mathbf{V}$ is orthogonal, so we conclude that $\mathbf{v}_1 = \pm \boldsymbol{\pi} \operatorname{diag}(\sqrt{\boldsymbol{\pi}}^{-1})$. Using this, we can simplify the expression for $[\mathbf{P}^{2t}]_{jj}$ and upper bound it via

$$\begin{aligned}[\mathbf{P}^{2t}]_{jj} = \pi_j + \sum_{i=2}^{n} \lambda_i^{2t} [\mathbf{v}_i]_j^2 &\leq \pi_j + \lambda_2^{2t} \sum_{i=2}^{L} [\mathbf{v}_i]_j^2 \\ &\leq \pi_j + \lambda_2^{2t} \sum_{i=1}^{L} [\mathbf{v}_i]_j^2 \\ &= \pi_j + \lambda_2^{2t},\end{aligned}$$

because $\mathbf{V}$ is orthogonal. Putting this into our earlier expression yields

$$\left(\frac{[\mathbf{P}^{2t}]_{jj}}{\pi_j} - 1 \right)^{1/2} \leq \left(\frac{\pi_j + \lambda_2^{2t}}{\pi_j} - 1 \right)^{1/2} = \left(\frac{\lambda_2^{2t}}{\pi_j} \right)^{1/2}.$$

[3] Matrices $\mathbf{A}$ and $\mathbf{B}$ are similar if $\mathbf{A} = \mathbf{C}^{-1} \mathbf{B} \mathbf{C}$ for some invertible matrix $\mathbf{C}$.

Putting all of the bounds together we have

$$|\boldsymbol{\mu}\mathbf{P}^t - \boldsymbol{\pi}|_{\mathrm{TV}} \le \max_j \left(\frac{\lambda_2^{2t}}{\pi_j}\right)^{1/2} = \lambda_2^t \left(\frac{1}{\min_j \pi_j}\right)^{1/2}.$$

Setting the right hand side equal to ϵ and solving for t gives

$$t = \frac{\log \epsilon + \frac{1}{2}\log(\min_j \pi_j)}{\log \lambda_2}.$$

This gives an upper bound on the time it takes the Markov chain to mix so that $|\boldsymbol{\mu}\mathbf{P}^t - \boldsymbol{\pi}|_{\mathrm{TV}} < \epsilon$. Because we used inequalities to arrive at t, we can only conclude that

$$T_{\mathrm{mix}}(\epsilon) \le t = \frac{\log \epsilon + \frac{1}{2}\log(\min_j \pi_j)}{\log \lambda_2} = \frac{\log \frac{1}{\epsilon} + \frac{1}{2}\log \frac{1}{\min_j \pi_j}}{\log 1/\lambda_2} \le \frac{\log \frac{1}{\epsilon} + \frac{1}{2}\log \frac{1}{\min_j \pi_j}}{1 - \lambda_2}.$$

So, as expected, the mixing time depends on the difference between the largest and second largest eigenvalues. In this case, Cheeger's celebrated inequality states that

$$\frac{1}{1-\lambda_2} \le \frac{2}{\Phi^2},$$

where the *conductance* Φ of $\mathbf{P}$ is defined as

$$\Phi = \Phi(\mathbf{P}) = \min_{\mathcal{S} \subset \mathcal{X}} \frac{\sum_{i \in \mathcal{S}, j \in \mathcal{S}^{\mathrm{c}}} \pi_i P_{ij}}{\left(\sum_{k \in \mathcal{S}} \pi_k\right)\left(\sum_{k \in \mathcal{S}^{\mathrm{c}}} \pi_k\right)}.$$

Conductance takes the minimum over $\mathcal{S}$ of the probability of starting in $\mathcal{S}$ and transitioning to $\mathcal{S}^{\mathrm{c}}$ in one time step normalized by the "sizes" of $\mathcal{S}$ and $\mathcal{S}^{\mathrm{c}}$. If the conductance is small, then there is a set $\mathcal{S}$ such that transitioning out of $\mathcal{S}$ is difficult, so if the Markov chain gets stuck in $\mathcal{S}$, it will be unlikely to leave $\mathcal{S}$, hence we would expect the mixing time to be large. See Fig. 4.

Thus we conclude that

$$T_{\mathrm{mix}}(\epsilon) \le \frac{2}{\Phi^2}\left(\log \frac{1}{\min_i \pi_i} + \log \frac{1}{\epsilon}\right).$$

In fact, it can be shown that without converting our Markov chain to a lazy Markov chain, we improve the bound by a factor of 2

$$T_{\mathrm{mix}}(\epsilon) \le \frac{1}{\Phi^2}\left(\log \frac{1}{\pi_{min}} + \log \frac{1}{\epsilon}\right).$$

Example 4. Consider the simple Markov chain depicted in Fig. 5 with a single binary variable $\mathsf{x} \in \mathcal{X} = \{0, 1\}$. By symmetry, the stationary distribution $\boldsymbol{\pi} = [1/2, 1/2]$. It's clear that Φ is minimized when $\mathcal{S} = \{0\}$ and $\mathcal{S}^{\mathrm{c}} = \{1\}$ so that

$$\Phi = \frac{\delta/2}{(1/2)\cdot(1/2)} = 2\delta.$$

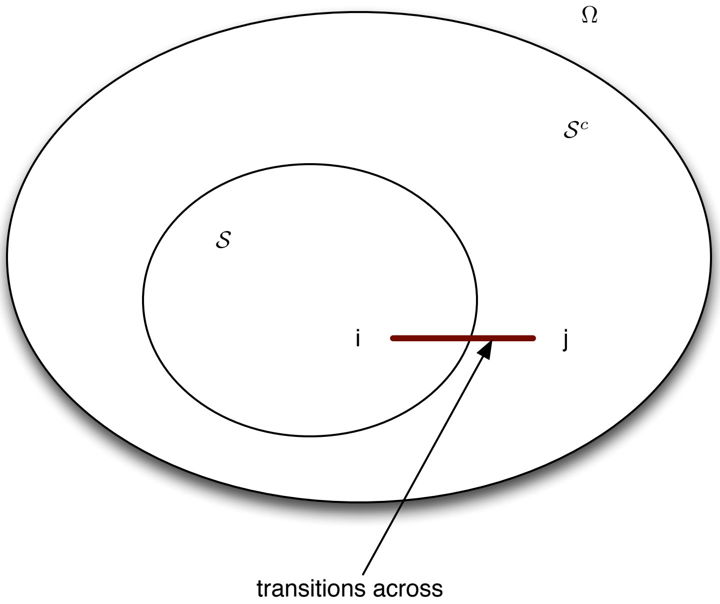

Figure 4

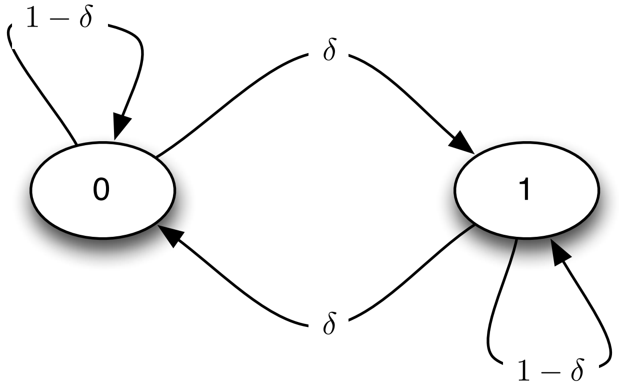

Figure 5