Chapter 3

Markov Chains

Suppose we have a system that can be in one of a number of states. For example:

- **A machine**, that either works or is broken.
- **The sky**, that can have
  - no clouds
  - clouds but no precipitation
  - rain clouds but no other precipitation
  - other precipitation, e.g., snow, hail, etc.
- **A queue** with n clients ($n = 0, 1, 2, \dots$).

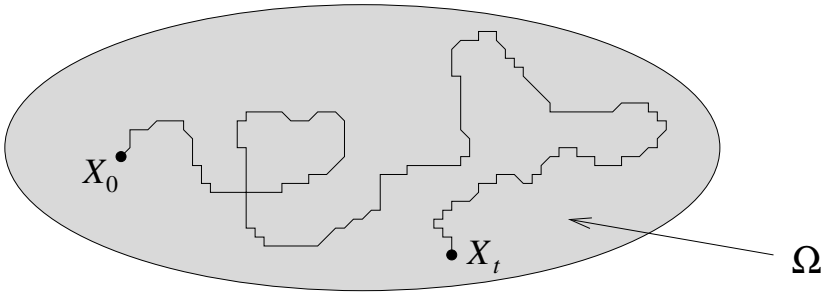
The first two examples have a finite number of states. The third example can have an infinite number of states. But in all three examples, the number of states is countable.

A Markov chain is a random process that models the state of the system at a given discrete time step. Markov chains have a *memoryless* property: the probability of being in a given state is determined *only* by the system's previous state, and is independent of all earlier history. Thus, a Markov chain is specified by a set of transition probabilities, giving the probability of making a transition from one state to another.

3.1 Definitions

Define the following notation:

- Let Ω be the **state space**: a countable set of states $\omega \in \Omega$.
- Let X_t be a random variable denoting the **system state** at time step t.



Then the sequence $X_0, X_1, \ldots$ is a **Markov chain** if for all time steps t,

$$\mathbb{P}\mathrm{r}[X_{t+1} = x_{t+1} | X_0 = x_0, X_1 = x_1, \ldots, X_t = x_t] = \mathbb{P}\mathrm{r}[X_{t+1} = x_{t+1} | X_t = x_t].$$

$\mathbb{P}\mathrm{r}[X_{t+1} = y | X_t = x]$ is called the *transition probability* of going from state x to state y. For most of our discussion, we will restrict ourselves to *homogeneous Markov chains*, where transition probabilities are stationary in time, so

$$\mathbb{P}\mathrm{r}[X_{t+1} = y | X_t = x] = P_{xy},$$

independent of t. We may think of P_{xy} as being an element of the *transition matrix* $\mathbf{P}$. Furthermore, since probabilities are normalized,

$$\sum_{y \in \Omega} P_{xy} = 1,$$

expressing the simple fact that if the system is at state x at a given time step, it must be *somewhere* within the state space Ω at the next time step.

Example 3.1. Consider a machine that either works or is broken. Define

- State 1: the machine works
- State 2: the machine is broken

Imagine that if the machine works, with some probability p it will break at the next time step. If the machine is broken, with some probability q it will be fixed at the next time step. Then

$$P_{12} = p, \quad P_{21} = q,$$

and from normalization,

$$P_{11} = 1 - P_{12} = 1 - p, \quad P_{22} = 1 - P_{21} = 1 - q.$$

(Note that normalization does *not* imply that $p + q = 1$ in general!) So the transition matrix is

$$\mathbf{P} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}.$$

Now consider what happens over more than one time step. It is not hard to see from basic properties of conditional probabilities that

$$\mathbb{P}\mathrm{r}[X_2 = y | X_0 = x] = \sum_{z \in \Omega} \mathbb{P}\mathrm{r}[X_2 = y | X_1 = z, X_0 = x] \, \mathbb{P}\mathrm{r}[X_1 = z | X_0 = x].$$

From the memoryless nature of the Markov chain,

$$\mathbb{P}\mathrm{r}[X_2 = y | X_1 = z, X_0 = x] = \mathbb{P}\mathrm{r}[X_2 = y | X_1 = z],$$

so

$$\begin{aligned} \mathbb{P}\mathrm{r}[X_2 = y | X_0 = x] &= \sum_{z \in \Omega} \mathbb{P}\mathrm{r}[X_2 = y | X_1 = z] \, \mathbb{P}\mathrm{r}[X_1 = z | X_0 = x] \\ &= \sum_{z \in \Omega} P_{xz} P_{zy} \\ &= (\mathbf{P}^2)_{xy}, \end{aligned}$$

where the second equality follows from the transition probabilities being stationary in time. In general, over t time steps,

$$\mathbb{P}\mathrm{r}[X_t = y | X_0 = x] = (\mathbf{P}^t)_{xy},$$

and indeed for any t_0,

$$\mathbb{P}\mathrm{r}[X_{t_0+t} = y | X_{t_0} = x] = (\mathbf{P}^t)_{xy}.$$

So the probability of a t-step transition is given by $\mathbf{P}^t$, the transition matrix to the tth power.

What about the probability of simply being in state x at time step t, in the absence of any prior information? Define

$$\pi_x(t) = \mathbb{P}\mathrm{r}[X_t = x].$$

(Note that we denote time step t using a parenthetical argument of π_x, but a subscript of X.) We may then think of $\pi_x(t)$ as being an element of a vector quantity $\vec{\pi}(t)$, describing the state probabilities at time t. As before, since probabilities are normalized,

$$\sum_{x \in \Omega} \pi_x(t) = 1$$

for all t.

Using the basic property that

$$\mathbb{P}\mathrm{r}[X_1 = y] = \sum_{x \in \Omega} \mathbb{P}\mathrm{r}[X_0 = x] \, \mathbb{P}\mathrm{r}[X_1 = y | X_0 = x],$$

it follows that

$$\pi_y(1) = \sum_{x \in \Omega} \pi_x(0) P_{xy},$$

or in matrix/vector notation,

$$\vec{\pi}(1) = \vec{\pi}(0)\mathbf{P}$$

where $\vec{\pi}(t)$ is a row vector. Likewise,

$$\vec{\pi}(2) = \vec{\pi}(1)\mathbf{P} = \vec{\pi}(0)\mathbf{P}^2,$$

and in general, over t time steps,

$$\vec{\pi}(t) = \vec{\pi}(0)\mathbf{P}^t,$$

or equivalently,

$$\vec{\pi}(t_0 + t) = \vec{\pi}(t_0)\mathbf{P}^t.$$

3.2 Stationary distribution

Let us return to Example 3.1 with the two-state machine. Given the probability vector $\vec{\pi}(0)$ that the machine works or is broken at initial time step 0, we can therefore write the corresponding probability at time t as

$$\begin{aligned} \vec{\pi}(t) &= \vec{\pi}(0)\mathbf{P}^t \\ &= \vec{\pi}(0) \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}^t. \end{aligned}$$

The easiest way to evaluate this is to look at the effect of $\mathbf{P}$ on a given component of $\vec{\pi}$, say π_1:

$$\pi_1(1) = (1-p)\pi_1(0) + q\pi_2(0),$$

and from the normalization property $\pi_2 = 1 - \pi_1$,

$$\begin{aligned} \pi_1(1) &= (1-p)\pi_1(0) + q(1-\pi_1(0)) \\ &= (1-p-q)\pi_1(0) + q. \end{aligned}$$

Iterating for subsequent time steps,

$$\begin{aligned} \pi_1(2) &= (1-p-q)\pi_1(1) + q \\ &= (1-p-q)^2\pi_1(0) + q(1-p-q) + q \\ &\vdots \\ \pi_1(t) &= (1-p-q)^t\pi_1(0) + q(1-p-q)^{t-1} + q(1-p-q)^{t-2} + \cdots + q \\ &= (1-p-q)^t\pi_1(0) + q\sum_{j=0}^{t-1}(1-p-q)^j \\ &= (1-p-q)^t\pi_1(0) + q\,\frac{1-(1-p-q)^t}{p+q} \\ &= \frac{q}{p+q} + (1-p-q)^t\left(\pi_1(0) - \frac{q}{p+q}\right). \end{aligned}$$

If $0 < p+q < 2$, then $(1-p-q)^t$ vanishes at large times t, leading to

$$\lim_{t\to\infty}\pi_1(t) = \frac{q}{p+q},$$

and likewise from normalization,

$$\lim_{t\to\infty}\pi_2(t) = \frac{p}{p+q}.$$

So unless $p = q = 0$ or $p = q = 1$, after the process runs for a long time, the probabilities of being in a certain state are given by the transition probabilities p and q — regardless of the starting state. The Markov chain "forgets" its initial state probabilities, and approaches a *stationary distribution*

$$\vec{\pi} = \left[\tfrac{q}{p+q},\quad \tfrac{p}{p+q}\right].$$

Example 3.2. There is a road where, on average, 4 out of every 5 trucks are followed by a car, and 1 out of every 6 cars is followed by a truck. After a long time, what is the probability that the next vehicle we see on that road will be a truck?

We can model this as a Markov chain where state 1 represents a vehicle being a car and state 2 represents a vehicle being a truck. Then

$$P_{12} = p = 1/6, \quad P_{21} = q = 4/5,$$

and so

$$\mathbf{P} = \begin{bmatrix} 5/6 & 1/6 \\ 4/5 & 1/5 \end{bmatrix}.$$

The long-time probability of a vehicle being a truck is simply given by the stationary distribution,

$$\lim_{t\to\infty}\pi_2(t) = \frac{p}{p+q} = 5/29.$$

Why is $\vec{\pi}$ called a stationary distribution? If the $t \to \infty$ limit exists, the probabilities will not change when the Markov chain is allowed to run an additional step, since

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

But this means that the vector $\vec{\pi}$ is left untouched when multiplied by the transition matrix, and

$$\vec{\pi} = \vec{\pi}\mathbf{P}$$

We therefore see that the stationary distribution $\vec{\pi}$ of a Markov chain transition matrix $\mathbf{P}$ is the left eigenvector of $\mathbf{P}$ associated with the eigenvalue $\lambda = 1$. Eigenvectors are only defined up to a proportionality constant, but since $\vec{\pi}$ is normalized, it is fully determined.

In the case of our 2-state examples, we can solve the eigenvalue equation directly. In general, this is easier than calculating $\vec{\pi}(t)$ for finite time t and taking the $t \to \infty$ limit. We simply need to solve:

$$\begin{bmatrix}\pi_1, & 1-\pi_1\end{bmatrix} = \begin{bmatrix}\pi_1, & 1-\pi_1\end{bmatrix} \begin{bmatrix}1-p & p \\ q & 1-q\end{bmatrix} = \begin{bmatrix}\pi_1(1-p) + (1-\pi_1)q, & \pi_1 p + (1-\pi_1)(1-q)\end{bmatrix}.$$

It is sufficient to look at the first component,

$$\pi_1 = \pi_1(1-p) + (1-\pi_1)q,$$

from which we recover our earlier result,

$$\pi_1 = \frac{q}{p+q}.$$

3.3 Properties

Let us now establish a few important properties of Markov chains. We often refer to a Markov chain by its transition matrix, so when we say that $\mathbf{P}$ has a certain property, we mean that the Markov chain $X_0, X_1, \ldots$ with transition matrix $\mathbf{P}$ has that property.

Theorem 3.3. *Every Markov chain* $\mathbf{P}$ *has a stationary distribution* $\vec{\pi}$*, such that*

$$\vec{\pi} = \vec{\pi}\mathbf{P}.$$

Proof. Since $\vec{\pi}$ is the left eigenvector solving the eigenvalue equation

$$\vec{\pi}\lambda = \vec{\pi}\mathbf{P}$$

for $\lambda = 1$, it is sufficient to show that $\mathbf{P}$ must have an eigenvalue $\lambda = 1$. This is left as a homework assignment. □

So the existence of a stationary distribution is guaranteed, but is the stationary distribution unique? That depends on whether or not the Markov chain is *irreducible*.

Definition 3.4 (Irreducibility)**.** *A Markov chain* $\mathbf{P}$ *is* **irreducible** *if* $\forall x, y \in \Omega, \exists t$ *such that*

$$(\mathbf{P}^t)_{xy} > 0,$$

i.e., for any two states x *and* y *in the state space of* $\mathbf{P}$*, given a sufficient amount of time, we can get from state* x *to state* y*.*

Irreducibility is an accessibility property. As we will see in a later lecture, if one thinks of the states as being vertices in a *directed* graph with an edge from x to y if a transition $x \to y$ is possible ($P_{xy} > 0$), then irreducibility implies that the graph is strongly connected.

A consequence of this accessibility property is the following lemma.

Lemma 3.5. *If a Markov chain* $\mathbf{P}$ *is irreducible, all elements of its stationary distribution are strictly positive:* $\forall x \in \Omega, \pi_x > 0$.

Proof. From normalization, there must exist at least one positive element of $\vec{\pi}$:

$$\exists x : \pi_x > 0.$$

Since $\vec{\pi}$ is the stationary distribution,

$$\vec{\pi} = \vec{\pi}\mathbf{P} = \vec{\pi}\mathbf{P}^2 = \cdots = \vec{\pi}\mathbf{P}^t \quad \forall t.$$

Now consider any other element y. From irreducibility,

$$\exists t : (\mathbf{P}^t)_{xy} > 0.$$

For that particular t, since $\vec{\pi} = \vec{\pi}\mathbf{P}^t$, it follows that

$$\begin{aligned} \pi_y &= \sum_{z \in \Omega} \pi_z (\mathbf{P}^t)_{zy} \\ &\geq \pi_x (\mathbf{P}^t)_{xy} > 0. \end{aligned}$$

Thus $\forall y, \pi_y > 0$. □

Moreover, irreducibility directly determines the uniqueness of the stationary distribution.

Theorem 3.6. *If a Markov chain* $\mathbf{P}$ *is irreducible, it has a* unique *stationary distribution* $\vec{\pi}$. *(The eigenvalue* $\lambda = 1$ *has multiplicity 1.)*

Proof. We have already proven the existence of the stationary distribution $\vec{\pi}$ in Theorem 3.3. For uniqueness, we proceed by contradiction.

(The following proof has a flaw in it! Can you spot the mistake? Can you see how to fix it?)

Assume that more than one stationary distribution existed for $\mathbf{P}$. Then there would exist $\vec{\pi} \neq \vec{\pi}'$ such that

$$\begin{aligned} \vec{\pi} &= \vec{\pi}\mathbf{P} \\ \vec{\pi}' &= \vec{\pi}'\mathbf{P} \end{aligned}$$

and such that $\vec{\pi}$ and $\vec{\pi}'$ are linearly independent. In that case, $\forall a, b$,

$$a\vec{\pi} - b\vec{\pi}' = (a\vec{\pi} - b\vec{\pi}')\mathbf{P},$$

so $a\vec{\pi} - b\vec{\pi}'$ is also a stationary distribution of $\mathbf{P}$. Now, choose some state $x \in \Omega$, and let

$$b = a\frac{\pi_x}{\pi'_x}.$$

Then $(a\vec{\pi} - b\vec{\pi}')_x = 0$. But according to Lemma 3.5, no component of a stationary distribution can vanish. Since we said that $a\vec{\pi} - b\vec{\pi}'$ is a stationary distribution, we reach a contradiction. □

Now that we have established *whether* a Markov chain can reach a particular state, we consider a property concerning *how often* it reaches that state.

Definition 3.7 (Aperiodicity)**.** *A Markov chain* $\mathbf{P}$ *is* **aperiodic** *if* $\forall x, y \in \Omega$*,*

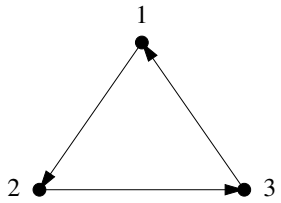
$$\gcd\{t : (\mathbf{P}^t)_{xy} > 0\} = 1,$$

where gcd *denotes the greatest common divisor.*

Aperiodicity implies that the Markov chain cannot get "trapped" in a cycle where, for instance, state y can be reached from state x only after m steps, $2m$ steps, etc. The following Markov chain is an example of this trapping.

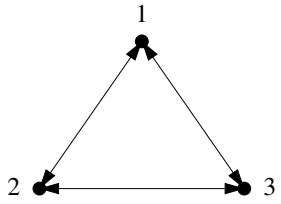
Example 3.8.

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}$$



Here, $\forall x, \gcd\{t : (\mathbf{P}^t)_{xx} > 0\} = 3$, so the aperiodicity condition is violated. On the other hand, if the state transitions are made "bidirectional," e.g.,

$$\mathbf{P} = \begin{bmatrix} 0 & 1/2 & 1/2 \\ 1/2 & 0 & 1/2 \\ 1/2 & 1/2 & 0 \end{bmatrix}$$



then the Markov chain can no longer get trapped in a cycle, and it is aperiodic:

$$\forall x, y \in \Omega : \gcd\{t : (\mathbf{P}^t)_{xy} > 0\} = 1.$$

The next property we consider is ergodicity. As we will see shortly, ergodicity is very closely connected with the use of Markov chains for sampling from a distribution. In what follows, for simplicity of definitions, we restrict ourselves to the case of Markov chains with a finite number of states.

Definition 3.9 (Ergodicity)**.** *A finite Markov chain* $\mathbf{P}$ *is* **ergodic** *if*

$$\exists t : \forall x, y \in \Omega, \quad (\mathbf{P}^t)_{xy} > 0.$$

Note the difference between irreducibility and ergodicity. Irreducibility means that one can get from any given state x to any given state y given enough steps — but the number of steps could differ for every pair. Ergodicity means that for some *fixed* t, *any* transition $x \to y$ can be made in exactly t steps.

Example 3.10. Consider Example 3.8 above. The first (periodic) Markov chain is irreducible, but is certainly not ergodic. The only way to get from state 1 to state 2 is in 1 step, 4 steps, 7 steps, etc. The only way to get from state 1 to state 3 is in 2 steps, 5 steps, 8 steps, etc. And the only

way to get from state 1 back to state 1 again is in 3 steps, 6 steps, 9 steps, etc. Thus, for any value of t, there is some pair of states such that a transition between them is impossible using exactly t steps.

The second (bidirectional) Markov chain, on the other hand, is ergodic. While in one step we cannot get from a state x back to itself, in two steps we can get from any state to any other state.

The precise connection between irreducibility and ergodicity is expressed in the following theorem.

Theorem 3.11. *A finite Markov chain* $\mathbf{P}$ *is ergodic if and only if* $\mathbf{P}$ *is irreducible and aperiodic.*

Proof. Here we will only prove the theorem in the easier direction: that ergodicity implies irreducibility and aperiodicity.

If $\mathbf{P}$ is ergodic, then $\exists t$ such that $\forall x, y \in \Omega$,

$$(\mathbf{P}^t)_{xy} > 0.$$

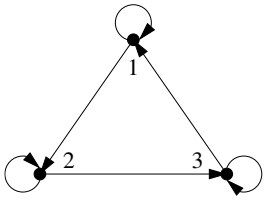
Clearly, this value of t is sufficient to show irreducibility. But moreover, $\forall i$,

$$(\mathbf{P}^{t+i})_{xy} = \sum_{z \in \Omega} (\mathbf{P}^i)_{xz} (\mathbf{P}^t)_{zy} > 0,$$

since from normalization it is impossible that $(\mathbf{P}^i)_{xz} = 0 \; \forall z$. And if $(\mathbf{P}^{t+i})_{xy} > 0$ for all i, it must follow that $\gcd\{t : (\mathbf{P}^t)_{xy} > 0\} = 1$. So $\mathbf{P}$ is aperiodic. □

In many cases, we can show that $\mathbf{P}$ is ergodic simply using the direct method of raising $\mathbf{P}$ to some power and showing that all elements are nonzero. But if that is too difficult, the previous theorem provides a powerful shortcut: show that $\mathbf{P}$ is irreducible and aperiodic, and it follows that $\mathbf{P}$ is ergodic. Since irreducibility is a weaker property than ergodicity, it is typically easier to show. How about aperiodicity? The following example shows a convenient trick to *making* $\mathbf{P}$ aperiodic even if it is not already.

Example 3.12. Consider once again the cyclic Markov chain from Example 3.8 above. This is periodic, with a period of 3. But by adding in *self-loops* at each state ($\forall x : P_{xx} > 0$), we allow the Markov chain to remain at a given state and thus break its periodicity:



Definition 3.13 (Lazy Markov chain)**.** *A* **lazy Markov chain** $\mathbf{P}'$ *is formed from a Markov chain* $\mathbf{P}$ *by adding self-loops to all states:*

$$\mathbf{P}' = \frac{\mathbf{P} + \mathbf{I}}{2}.$$

Note that the self-loop probability does not have to be $1/2$*: it could be any* ε *where* $0 < \varepsilon < 1$*.*

It is not hard to show that $\mathbf{P}'$ is aperiodic. If $\mathbf{P}$ is irreducible, then $\mathbf{P}'$ must be as well, so $\mathbf{P}'$ is ergodic. Furthermore, since

$$\vec{\pi}\mathbf{P}' = \vec{\pi}\frac{\mathbf{P}+\mathbf{I}}{2} = \frac{\vec{\pi}+\vec{\pi}}{2} = \vec{\pi},$$

the lazy Markov chain $\mathbf{P}'$ has the same stationary distribution as $\mathbf{P}$ itself. Consequently, as long as $\mathbf{P}$ is irreducible, we can make sure it will converge to its stationary distribution simply by diluting it with self-loops (at the cost of potentially slowing its convergence).

While self-loops at every state are convenient for preserving the stationary distribution, they are not all needed for aperiodicity. As you will show in the homework, we have the following sufficient condition for an irreducible Markov chain to be aperiodic.

Theorem 3.14. *If, for an irreducible Markov chain* $\mathbf{P}$*,*

$$\exists x \in \Omega : P_{xx} > 0,$$

i.e., $\mathbf{P}$ *has at least* one *self-loop, then* $\mathbf{P}$ *is aperiodic.*

As we mentioned earlier, ergodicity is crucial to the process of sampling from a distribution. This is illustrated by the following theorem.

Theorem 3.15 (Fundamental theorem of Markov chains)**.** *If a finite Markov chain* $\mathbf{P}$ *is ergodic, it converges to its unique stationary distribution:*

$$\lim_{t\to\infty} \vec{\pi}(t) = \vec{\pi}.$$

It follows from the theorem that if we run the Markov chain long enough, it will converge regardless of the state in which we start, as in the two-state machine in Example 3.1. Specifically, since $\vec{\pi}(t) = \vec{\pi}(0)\mathbf{P}^t$, by setting $\vec{\pi}(0)$ to the unit vector $\vec{e}_x$ (meaning that the system starts in state x), we find that

$$\forall x, \quad \lim_{t\to\infty} \vec{e}_x\mathbf{P}^t = \vec{\pi},$$

or

$$\forall x, y, \quad \lim_{t\to\infty} (\mathbf{P}^t)_{xy} = \pi_y.$$

The proof of the fundamental theorem of Markov chains is not simple and will not be given here. Note, however, that ergodicity necessarily implies irreducibility, so if $\mathbf{P}$ is ergodic then its stationary distribution must be unique. But irreducibility does not necessarily imply ergodicity, and a Markov chain could have a unique stationary distribution without converging to it.

Finally, a convenient tool for finding the stationary distribution of an ergodic Markov chain is *reversibility*.

Definition 3.16 (Reversibility)**.** *A Markov chain* $\mathbf{P}$ *is* reversible *with respect to a distribution* $\vec{q}$ *if*

$$\forall x, y \in \Omega, \quad q_x P_{xy} = q_y P_{yx}.$$

This essentially states that the probability flow along an edge of the state graph is equal in both directions. We then have the following theorem.

Theorem 3.17. *If a Markov chain* $\mathbf{P}$ *is reversible with respect to* $\vec{q}$*, then* $\vec{q}$ *is a stationary distribution of* $\mathbf{P}$*.*

Proof. Starting from the reversibility definition, sum over all states x:

$$\begin{aligned}\sum_{x\in\Omega} q_x P_{xy} &= \sum_{x\in\Omega} q_y P_{yx} \\ &= q_y \sum_{x\in\Omega} P_{yx} \\ &= q_y,\end{aligned}$$

or

$$\vec{q}\mathbf{P} = \vec{q}.$$

□

So in cases where it is impractical to solve the eigenvalue problem $\vec{\pi} = \vec{\pi}\mathbf{P}$ directly, reversibilty can provide an alternative means of finding the stationary distribution.

The following two corollaries follow immediately from Theorem 3.17.

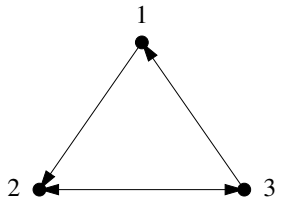
Corollary 3.18. *If a Markov chain* $\mathbf{P}$ *is irreducible and reversible with respect to* $\vec{q}$*, then* $\vec{q}$ *is its* unique *stationary distribution.*

Corollary 3.19. *If a Markov chain* $\mathbf{P}$ *is ergodic and reversible with respect to* $\vec{q}$*, then it converges to the stationary distribution* $\vec{q}$*.*

Note that an ergodic Markov chain can converge to a stationary distribution without being reversible, as illustrated in the following example.

Example 3.20. Consider the Markov chain

$$\mathbf{P} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1/2 & 1/2 & 0 \end{bmatrix}$$



From state 1, a transition is possible only to state 2. From state 2, a transition is possible only to state 3. From state 3, however, transitions can take place either to state 1 or to state 2, with equal probability.

One can confirm that with 5 steps of this Markov chain, it is possible to get from any state to any other:

$$\mathbf{P}^5 = \begin{bmatrix} 1/4 & 1/4 & 1/2 \\ 1/4 & 1/2 & 1/4 \\ 1/8 & 3/8 & 1/2 \end{bmatrix}$$

The Markov chain is thus ergodic, and converges to a unique stationary distribution given by

$$\vec{\pi} = \vec{\pi}\mathbf{P}$$

which is easily shown to be $\vec{\pi} = [1/5, 2/5, 2/5]$. But is $\mathbf{P}$ reversible? Notice that

$$\begin{aligned}\pi_1 P_{12} &> 0 \\ \pi_2 P_{21} &= 0,\end{aligned}$$

so it cannot possibly be reversible. Therefore, reversibility can be a convenient way of finding the stationary distribution, but it is not always applicable.

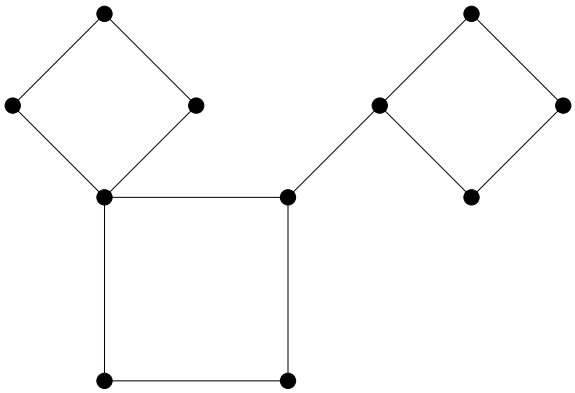
3.4 Sampling

We have established that an ergodic Markov chain $\mathbf{P}$ samples, over long times, from a unique stationary distribution $\vec{\pi}$. But suppose that we have some specific target distribution $\vec{\pi}$, and we wish to *construct* a Markov chain that samples from it. How can we do so?

3.4.1 Undirected graph

We began this course with a graph theoretic example of a discrete modeling problem. Suppose that the states of a Markov chain are the nodes of a graph, and the possible transitions are given by edges on the graph. The edges on the graph are undirected, so if a transition is possible from x to y, it is also possible from y to x. What will be the stationary distribution? And how can we *set* the transition probabilities to ensure a certain stationary distribution — say, a uniform one?

Let us start with the first question, under a straightforward scenario. Consider the following graph:



For illustration purposes, take a very simple kind of transition dynamics on the graph. When the Markov chain is at a node x, let it move to any one of the node's neighbors y on the graph with equal probability $P_{xy} = 1/\deg(x)$, where $\deg(x)$ is the *degree* of the node x: the number of edges incident on it (always 2, 3 or 4 in the example above).

What is the stationary distribution on this graph? From the formulation of the problem, solving the eigenvalue equation does not look simple. Reversibility makes things much easier. From Theorem 3.17,

$$\frac{\pi_x}{\deg(x)} = \frac{\pi_y}{\deg(y)}$$

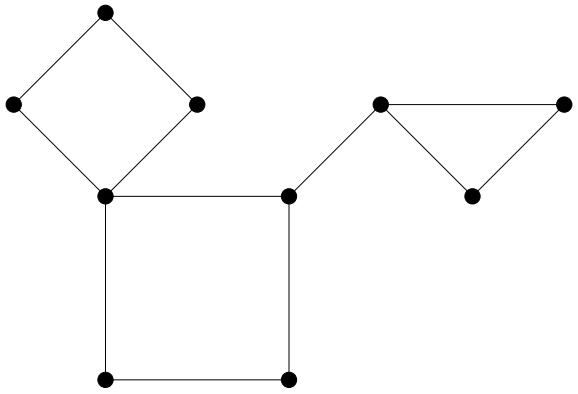
as long as an edge connects x and y. This will be satisfied if $\pi_x \propto \deg(x)$. The constant of proportionality is set by normalization of $\vec{\pi}$, and so

$$\pi_x = \frac{\deg(x)}{\sum_{z\in\Omega} \deg(z)}.$$

Does $\mathbf{P}$ converge to this stationary distribution? As long as the graph is connected, $\mathbf{P}$ is irreducible. Is $\mathbf{P}$ aperiodic as well? It turns out that this is related to the graph coloring problem in Chapter 1. If the graph can be colored with exactly 2 colors (as we will see later, this is called a *bipartite* graph), $\mathbf{P}$ is periodic. To see this, let one color be red and the other blue. At each step, the Markov chain alternates between a red and a blue state. Thus, after an even number of steps it always ends up on the starting color, and after an odd number of steps it always ends up on

the opposite color. It has a period of 2. That is the case for the graph in the example above: its chromatic number is 2, and so the corresponding Markov chain will not be ergodic.

On the other hand, if more than 2 colors are required anywhere on an undirected graph, that will break the periodicity. (Note that this is not necessarily true for a *directed* graph, as we saw in Example 3.8.) An example is the graph above, with a small modification:



The triangle is sufficient to force the chromatic number to be 3. The Markov chain is now irreducible and aperiodic, hence ergodic, and converges to its stationary distribution. Note that we could also have enforced aperiodicity in the original example by adding self-loops at each node, making $\mathbf{P}$ lazy.

But now, suppose that we want to construct a Markov chain such that after running for a long time, we are equally likely to be at any node of the graph. This means that we need π_x to be a constant, independent of x. Reversibility then tells us that as long as an edge connects x and y,

$$P_{xy} = P_{yx}.$$

How can we accomplish this? We can set all nonzero transition probabilities P_{xy} to the same constant, independent of x and y — but we need to make sure probabilities remain normalized. So if $\deg_{\max}$ is the maximum degree in the graph (4 in our example), for all x and y connected by an edge, set

$$P_{xy} = \frac{1}{\deg_{\max}}.$$

From the normalization condition

$$\sum_{y \in \Omega} P_{xy} = 1,$$

this will result in a self-loop probability

$$P_{xx} = 1 - \frac{\deg(x)}{\deg_{\max}}.$$

Due to Theorem 3.14, as long as not all nodes in the graph have $\deg(x) = \deg_{\max}$, aperiodicity is guaranteed. So even with the first graph example above (with chromatic number 2), the Markov chain is ergodic and samples nodes uniformly at random on the graph.

3.4.2 Card shuffling

A beautiful demonstration of the use of Markov chains for sampling is the problem of shuffling cards. When a deck of cards is shuffled, it is very important that the order of the cards be completely

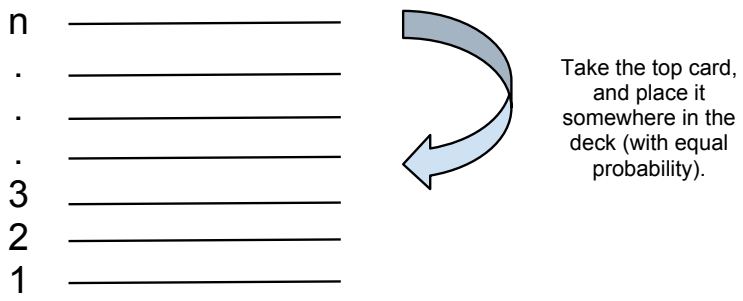
random, with no kind of bias. (The business model of essentially every legal casino relies on this!) How can we ensure it?

Ideally, given n cards, we would simply like to pick one of the $n!$ possible permutations uniformly at random. But when $n = 52$, $n! \approx 8 \times 10^{67}$. If one thinks of each permutation as being a state of the system, it is hard to even contemplate such a large number of states, let alone devise a method of directly picking one of them uniformly at random.

So let us model the shuffling process as a Markov chain transition. Can we show that, given a long enough time, this Markov chain samples from a uniform distribution over the states? And even more crucially, can we estimate how long we need to run the Markov chain before we are "sufficiently close to" the stationary distribution? We address the first question here, and the second question in Section 3.5.

Top-in-at-random shuffle

We start with the crudest kind of shuffle one can imagine. At each step, take the top card in the deck, and place it somewhere randomly in the deck.



This is of course a very slow form of shuffling, but it will help us in analyzing more sophisticated methods. Let a state $x \in \Omega$ be a given permutation of the n cards. If the cards are numbered in increasing order from the bottom to the top of the deck, a transition consists of taking card n from its present position and inserting it below what was previously card i, where i is drawn uniformly at random from 1 to n. So there can be transitions to n possible states y, each of which is equally likely:

$$P_{xy} = \frac{1}{n}.$$

Note that if $i = n$, we are putting the top card right back where it was. Thus, self-loops occur, with $P_{xx} = 1/n$.

Is this Markov chain ergodic? It is not hard to see that from any starting permutation, one can always generate a target permutation by repeatedly inserting the top card in the desired place in the deck, so $\mathbf{P}$ is irreducible. Furthermore, there are self-loops, so $\mathbf{P}$ is aperiodic. $\mathbf{P}$ is thus ergodic, and converges to its stationary distribution.

What is its stationary distribution? Solving the eigenvalue problem looks daunting. How about reversibility? Unfortunately, almost all transitions are asymmetric: if $P_{xy} > 0$, then unless $i = n$ (self-loop) or $i = n - 1$, $P_{yx} = 0$. This Markov chain cannot possibly be reversible.

Fortunately, another property can be used to find $\vec{\pi}$. If $\mathbf{P}$ satisfies not only the normalization

condition

$$\sum_{y \in \Omega} P_{xy} = 1 \quad \forall x$$

but also

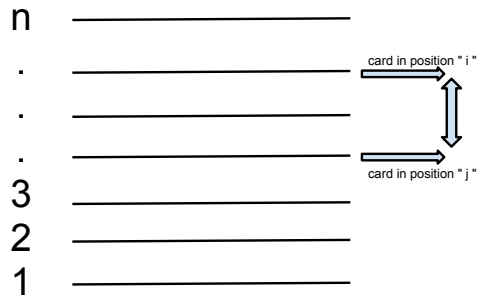
$$\sum_{x \in \Omega} P_{xy} = 1 \quad \forall y,$$

it is said to be **doubly stochastic**, in that not only is $\mathbf{P}$ a valid Markov chain but its transpose $\mathbf{P}^T$ is as well: it is a Markov chain when run both forwards and backwards.

It is not hard to see that the top-in-at-random shuffle is doubly stochastic. There are n possible transitions leading to a given state, since any one of the n cards could have been the top card in the previous state, and each of these transitions takes place with probability $1/n$. As you will show in the homework, the stationary distribution of a doubly stochastic Markov chain is uniform. So after running for long enough, the top-in-at-random shuffle will converge to a uniform distribution, where it samples each permutation of the deck of cards with equal probability.

Random transposition shuffle

A slightly more sophisticated shuffle is as follows. At each step, choose two positions i and j randomly in the deck. Take the cards at those positions, and swap them.



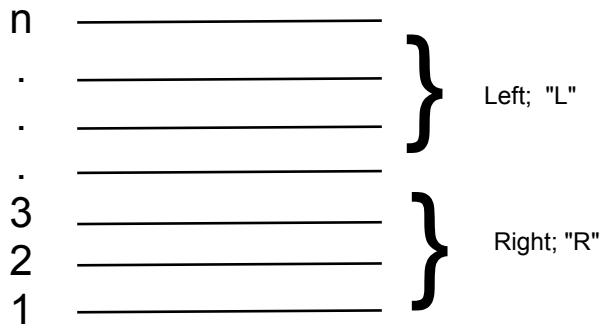
If the Markov chain $\mathbf{P}$ is defined by these transitions, is $\mathbf{P}$ irreducible? Yes: it is not hard to show that with a sequence of n pair swaps, we can generate any permutation that we want. Is $\mathbf{P}$ aperiodic? Yes, if we allow the possibility of i and j being the same card, i.e., pick positions i and j at random *with replacement*, because then we have self-loops. $\mathbf{P}$ is thus ergodic, and converges to its stationary distribution.

What is the stationary distribution? This Markov chain *is* reversible, and moreover $\mathbf{P}$ is completely symmetric: it is easy to see that if $x \to y$ is an allowable transition, $P_{xy} = P_{yx}$. We therefore know the stationary distribution without even having to calculate P_{xy}: as in the case of the undirected graph, reversibility tells us that $\pi_x = \pi_y$. (Actually, any symmetric Markov chain must also be doubly stochastic, so that is already sufficient to ensure a uniform stationary distribution.) Thus, after running long enough, the random transposition shuffle also samples each permutation of the deck of cards with equal probability.

Riffle shuffle

Finally we reach the classic card shuffle. In the riffle, or "dovetail" shuffle, one cuts the deck into two stacks, and then (with skillful use of one's thumbs and index fingers) one interleaves the two

stacks with each other, while preserving the *relative* order of cards within each stack. Call the two stacks "left" and "right."



Now the transition is somewhat more subtle. We make certain idealizations in modeling the process, based on an approach proposed by Gilbert and Shannon (1955) and independently by Reeds (1981).

To start with, if the process of cutting the deck is a fair one, we assume the size of each stack to be a binomial random variable, so the number of cards in (say) the left stack is $L \sim \text{Bin}(n, 1/2)$. This is reasonable, since it is what we would obtain if we simply placed each card with equal probability into the left or right stack.

For the interleaving procedure, we want to mix the two stacks as uniformly as we can. There are $\binom{n}{L}$ ways that we can place the cards from the left stack into the shuffled deck (while preserving the relative order of cards in the left stack). We assume that any one of these ways occurs with equal probability.

Note that it is conceivable (albeit unlikely) that when first cutting the deck, we will end up with a left stack containing only one card. But in that case, the procedure is in fact identical to the top-in-at-random shuffle. Since top-in-at-random is irreducible and aperiodic, it therefore follows that the riffle shuffle is as well. So the riffle shuffle is ergodic, and converges to its stationary distribution.

As you will show in the homework, this Markov chain turns out to be doubly stochastic. So once again, the stationary distribution is uniform, and the riffle shuffle samples each permutation of the deck of cards with equal probability.

3.4.3 Markov chain Monte Carlo

The idea behind the Monte Carlo method is the use of a random process, or "numerical experiments," to perform calculations. This method is fundamental to computer simulation, which must typically run a model under randomly chosen scenarios. But efficiently drawing a random sample from a very specific probability distribution is not simple, especially when the distribution may not even have a clear analytical description. This is where Markov chains are useful. So we return to the original question of this section: how can we construct a Markov chain to sample from a *desired* distribution — particularly one that in general is not uniform?

The **Metropolis-Hastings algorithm** is a frequently used method for doing this. Suppose that for any state x, we are given a set of allowable transitions to states $y \in N(x)$. We call $N(x)$ the *neighborhood class* or *move class* of x. Assume the move class is symmetric, so $y \in N(x)$ implies $x \in N(y)$. Define

$$N_{\max} = \max_{x \in \Omega} |N(x)|,$$

the largest possible number of neighbors. Suppose we also know that there always exists some sequence of moves taking us from any state x to any other state y. In order to sample from a target distribution $\vec{\pi}$ over these states, the algorithm proceeds as follows:

1. From current state x: with probability $p|N(x)|$, attempt a transition to another state, where

$$p = \frac{1}{N_{\max}}.$$

   If no transition is attempted, remain at x.

2. If a transition is attempted, choose any neighbor $y \in N(x)$ with equal probability. With probability

$$\min\left(1, \frac{\pi_y}{\pi_x}\right),$$

   accept the move to y. Otherwise, remain at x. Thus, if $\pi_y > \pi_x$ the move is always accepted, whereas if $\pi_y < \pi_x$ it may not be accepted.

3. Repeat at step 1. as long as a stopping criterion has not yet been met.

Why does this Markov chain sample exactly from the stationary distribution $\vec{\pi}$? Note first of all that the move class is constructed so that the Markov chain is irreducible. Furthermore, due to step 2., as long as $\vec{\pi}$ is not a uniform distribution, there must exist states where there is some chance of rejecting a move to a neighbor, so there must be at least one self-loop: the Markov chain is aperiodic. Thus, it is ergodic. As for the transition probabilities, if $y \in N(x)$,

$$P_{xy} = p \, \min\left(1, \frac{\pi_y}{\pi_x}\right).$$

Assume without loss of generality that $\pi_y \leq \pi_x$. Then

$$P_{xy} = p \, \frac{\pi_y}{\pi_x} \qquad \text{and} \qquad P_{yx} = p,$$

so

$$P_{xy}\pi_x = p \, \pi_y = P_{yx}\pi_y.$$

From reversibility, it follows that $\vec{\pi}$ is the stationary distribution.

Interestingly, the algorithm does not require knowledge of the exact stationary distribution, but only of some weight function w_x that is proportional to π_x, since it only uses the ratio π_y/π_x. This is significant, as the number of states may be so large that in practice we cannot enumerate them all, and so we cannot explicitly compute the normalization factor for $\vec{\pi}$. In such cases, the Markov chain Monte Carlo method allows us nevertheless to sample according to the weights.

3.5 Convergence time

We have seen how to construct a Markov chain that converges to a given stationary distribution. But how long does it take to converge? A Markov chain is a stochastic process, so in general we cannot state with certainty that it is sampling *precisely* from its stationary distribution. Instead, we ask that its distribution $\vec{\pi}(t)$ be "sufficiently close" to the limiting stationary distribution $\vec{\pi}$. How do we define "sufficiently close"? And how long it does it take to reach that point?

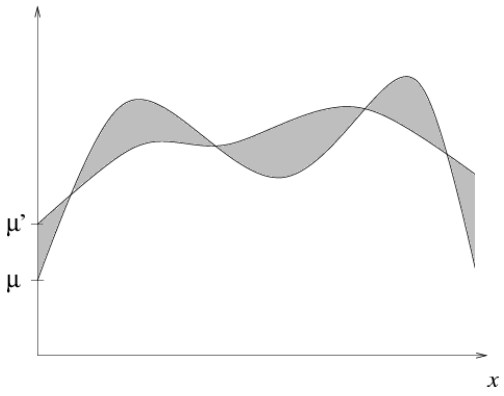
3.5.1 Mixing time

Consider two probability distributions over the state space Ω of the Markov chain. For instance, one of them might be the actual set of state probabilities $\vec{\pi}(t)$ at time t, and the other might be the stationary probabilities $\vec{\pi}$. The *total variation (TV) distance* between the two distributions is one of many possible quantities that measure how far apart they are from each other. Formally, we define it as follows.

Definition 3.21 (TV distance)**.** *Given two probability distributions $\vec{\mu}$ and $\vec{\mu}'$, the total variation (TV) distance between them is*

$$\left\|\vec{\mu} - \vec{\mu}'\right\| = \frac{1}{2}\sum_{x\in\Omega}\left|\mu_x - \mu'_x\right|.$$

Visually, if we plot the two distributions $\vec{\mu}$ and $\vec{\mu}'$ as a function of the state x, then $\|\vec{\mu} - \vec{\mu}'\|$ is simply equal to one-half of the shaded area below:



The TV distance is also the largest possible difference in probability that the two distributions can assign to the same event. This is expressed by the following theorem, which you will show in the homework.

Theorem 3.22. *For a subset of the state space $A \subseteq \Omega$, define $\mu_A = \sum_{x\in A}\mu_x$. Then*

$$\left\|\vec{\mu} - \vec{\mu}'\right\| = \max_{A\subseteq\Omega}\left|\mu_A - \mu'_A\right|.$$

A consequence is that

$$0 \le \left\|\vec{\mu} - \vec{\mu}'\right\| \le 1.$$

There is another interpretation of TV distance, as follows.

Theorem 3.23.

$$\left\|\vec{\mu} - \vec{\mu}'\right\| = 1 - \sum_{x\in\Omega}\min(\mu_x, \mu'_x).$$

Proof. Starting from the identity

$$\min(\mu_x, \mu'_x) = \frac{\mu_x + \mu'_x}{2} - \frac{|\mu_x - \mu'_x|}{2},$$

sum over all $x \in \Omega$:

$$\sum_{x\in\Omega}\min(\mu_x, \mu'_x) = \sum_{x\in\Omega}\frac{\mu_x + \mu'_x}{2} - \sum_{x\in\Omega}\frac{|\mu_x - \mu'_x|}{2}.$$

From normalization, the first sum on the right is 1. The second sum on the right is simply the TV distance, so

$$\left\|\vec{\mu} - \vec{\mu}'\right\| = 1 - \sum_{x \in \Omega} \min(\mu_x, \mu'_x).$$

□

Notice that if we have one random variable X (which could be the state of a Markov chain) distributed according to $\vec{\mu}$, i.e., $\mathbb{P}\mathrm{r}[X = x] = \mu_x$, and another random variable X' distributed according to $\vec{\mu}'$, then regardless of any correlations between X and X', $\mathbb{P}\mathrm{r}[X \neq X']$ must be equal to **at least** the TV distance $\|\vec{\mu} - \vec{\mu}'\|$. That is because, for any two events A and B,

$$\mathbb{P}\mathrm{r}[A \cap B] \leq \min(\mathbb{P}\mathrm{r}[A], \mathbb{P}\mathrm{r}[B]),$$

so

$$\begin{aligned}
\mathbb{P}\mathrm{r}[X \neq X'] &= 1 - \mathbb{P}\mathrm{r}[X = X'] \\
&= 1 - \sum_{x \in \Omega} \mathbb{P}\mathrm{r}[(X = x) \cap (X' = x)] \\
&\geq 1 - \sum_{x \in \Omega} \min(\mu_x, \mu'_x) \\
&= \left\|\vec{\mu} - \vec{\mu}'\right\|.
\end{aligned}$$

An illustration of this property comes from card shuffling again.

Example 3.24 (Card shuffling)**.** Consider a perfectly shuffled deck of n cards. Then each of the $n!$ permutations (states) is equally probable, and so the distribution is uniform:

$$\mu_x = \frac{1}{n!}.$$

Now consider another deck where we know what the top card is, but the remaining $n-1$ cards are perfectly shuffled. In that case, only $(n-1)!$ states can occur, and the distribution is

$$\mu'_x = \begin{cases} 1/(n-1)! & \text{if } x \text{ has known card on top} \\ 0 & \text{otherwise.} \end{cases}$$

The TV distance between these two distributions is given by

$$\left\|\vec{\mu} - \vec{\mu}'\right\| = 1 - \sum_{x \in \Omega} \min(\mu_x, \mu'_x) = 1 - \frac{(n-1)!}{n!} = 1 - \frac{1}{n},$$

so knowledge of *only* one out of the n cards puts us very close to the maximum distance! It would seem, then, that TV distance is a very unforgiving measure. But really this just reflects the fact that the probability of sampling the same state in the two decks is quite low. A state drawn in the first deck, where the top card could be anything, is usually not even an option in the second deck, since there we have fixed the value of the top card.

Let us apply all of this formalism to Markov chain convergence. If a Markov chain starts in state x at $t = 0$, then $\vec{\pi}(0) = \vec{e}_x$, the unit vector whose xth component is 1 and all other components are 0. Therefore,

$$\vec{\pi}(t) = \vec{e}_x \mathbf{P}^t,$$

i.e., row x of matrix $\mathbf{P}^t$. In order to measure how far the Markov chain can be from converging after t time steps, we therefore define the following quantity.

Definition 3.25. *The quantity $\Delta(t)$ is the largest possible TV distance between $\vec{\pi}(t)$ and the stationary distribution $\vec{\pi}$:*

$$\Delta(t) = \max \|\vec{\pi}(t) - \vec{\pi}\| .$$

Lemma 3.26. *The TV distance in $\Delta(t)$ is maximized when the Markov chain starts in a pure state:*

$$\Delta(t) = \max_{x \in \Omega} \left\| \vec{e}_x \mathbf{P}^t - \vec{\pi} \right\| .$$

Equipped with $\Delta(t)$, we now define the *mixing time*.

Definition 3.27 (Mixing time)**.** *The mixing time τ_{mix} of a Markov chain is the number of time steps needed for $\Delta(t)$ to fall below a specified threshold value. Conventionally,*

$$\tau_{\text{mix}} = \min \left\{ t : \Delta(t) \leq \frac{1}{2e} \right\} .$$

Why $1/2e$ for the threshold value? The choice is somewhat arbitrary, but it actually does not much matter what constant we choose. $\Delta(t)$ typically falls exponentially in t: recall the two-state machine in Example 3.1, where the part of $\vec{\pi}(t)$ that depends on $\vec{\pi}(0)$ vanishes exponentially. This means that dividing the threshold by some constant α would merely result in adding a quantity logarithmic in α to the mixing time. We really do not much care if our estimate of the mixing time is off by a small constant. What we care about is the distinction between a mixing time of $\Theta(n)$, $\Theta(n^2)$, etc.

3.5.2 Spectral analysis

If we can calculate the eigenvalues and eigenvectors of an ergodic Markov chain $\mathbf{P}$, then this offers a powerful approach to estimating its mixing time.

Let $\mathbf{P}$ have dimensions $n \times n$, with eigenvalues $\lambda_1, \ldots, \lambda_n$. Let $\vec{a}_k$ be the eigenvector associated with eigenvalue λ_k, so

$$\forall k, \quad \vec{a}_k \mathbf{P} = \vec{a}_k \lambda_k.$$

We now express $\vec{\pi}(0)$ as a linear combination of these eigenvectors:

$$\vec{\pi}(0) = \sum_{k=1}^{n} c_k \vec{a}_k.$$

(Strictly speaking, this requires $\mathbf{P}$ to be diagonalizable, which does not always hold. If not, one instead needs to express the matrix in Jordan canonical form and use its *generalized eigenvectors*, which is technically more complicated but leads to similar results.) Then,

$$\begin{aligned} \vec{\pi}(1) &= \vec{\pi}(0)\mathbf{P} \\ &= \sum_{k=1}^{n} c_k \vec{a}_k \mathbf{P} \\ &= \sum_{k=1}^{n} c_k \lambda_k \vec{a}_k. \end{aligned}$$

Likewise,

$$\vec{\pi}(2) = \vec{\pi}(1)\mathbf{P} = \sum_{k=1}^{n} c_k (\lambda_k)^2 \vec{a}_k,$$

and in general, over t time steps,

$$\vec{\pi}(t) = \sum_{k=1}^{n} c_k (\lambda_k)^t \vec{a}_k.$$

Clearly, $|\lambda_k| \leq 1$ for all k, or else $\vec{\pi}(t)$ would diverge for sufficiently large t and could not possibly be a probability distribution. Furthermore, we know that since $\mathbf{P}$ is irreducible, there is a unique eigenvalue of 1 (call this λ_1) whose associated eigenvector $\vec{a}_1$ is the stationary distribution $\vec{\pi}$. Thus,

$$\vec{\pi}(t) = c_1 \vec{\pi} + \sum_{k=2}^{n} c_k (\lambda_k)^t \vec{a}_k.$$

If $\mathbf{P}$ is ergodic, $\vec{\pi}(t)$ converges to $\vec{\pi}$ for $t \to \infty$, which requires $c_1 = 1$ and $|\lambda_k| < 1$ for $k > 1$. Let us then order the remaining eigenvalues so that

$$1 = \lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_n|.$$

This means

$$\vec{\pi}(t) = \vec{\pi} + \sum_{k=2}^{n} c_k (\lambda_k)^t \vec{a}_k,$$

where for large t, the sum is dominated by the $k = 2$ term. (To be more precise: it is dominated by all terms for which $|\lambda_k| = |\lambda_2|$. There could be several of them, as in the irreversible case of Example 3.20 which has complex eigenvalues $\lambda_2 = (-1+i)/2$ and $\lambda_3 = (-1-i)/2$, so $|\lambda_2| = |\lambda_3| = \sqrt{2}/2$.) Hence, the speed of convergence to the stationary distribution is governed primarily by the second-largest eigenvalue in magnitude.

In order to illustrate how to calculate the mixing time from the eigenvector expansion, we restrict ourselves for the present purposes to Markov chains with only two states. In that case, the eigenvector expansion above simplifies to

$$\vec{\pi}(t) = \vec{\pi} + c_2 (\lambda_2)^t \vec{a}_2,$$

or at time $t = 0$,

$$\vec{\pi}(0) = \vec{\pi} + c_2 \vec{a}_2,$$

so we may write

$$\vec{\pi}(t) = \vec{\pi} + (\lambda_2)^t [\vec{\pi}(0) - \vec{\pi}].$$

But then, since $\Delta(t)$ is the largest possible TV distance between $\vec{\pi}(t)$ and $\vec{\pi}$,

$$\begin{aligned} \Delta(t) &= \max \|\vec{\pi}(t) - \vec{\pi}\| \\ &= |\lambda_2|^t \max \|\vec{\pi}(0) - \vec{\pi}\| \\ &= |\lambda_2|^t \, \Delta(0), \end{aligned}$$

reflecting our observation that $\Delta(t)$ typically decays exponentially in t (and the rate of decay is controlled directly by λ_2). $\Delta(0)$ is given by

$$\begin{aligned} \Delta(0) &= \max_{x \in \Omega} \|\vec{e}_x - \vec{\pi}\| \\ &= \max \left\{ \|[1,0] - \vec{\pi}\|, \|[0,1] - \vec{\pi}\| \right\} \\ &= \max\{\pi_1, \pi_2\}, \end{aligned}$$

so

$$\Delta(t) = |\lambda_2|^t \max\{\pi_1, \pi_2\}.$$

$\Delta(t) = 1/2e$ when

$$t = \frac{1 + \ln 2 + \ln \max\{\pi_1, \pi_2\}}{-\ln |\lambda_2|},$$

so

$$\tau_{\text{mix}} = \left\lceil \frac{1 + \ln 2 + \ln \max\{\pi_1, \pi_2\}}{-\ln |\lambda_2|} \right\rceil.$$

For instance, consider Example 3.2 with cars and trucks on a road.

Example 3.28 (Cars and trucks, revisited)**.** The transition matrix is

$$\mathbf{P} = \begin{bmatrix} 5/6 & 1/6 \\ 4/5 & 1/5 \end{bmatrix}.$$

The eigenvalues are given by

$$\left(\frac{5}{6} - \lambda\right)\left(\frac{1}{5} - \lambda\right) - \frac{4}{5} \cdot \frac{1}{6} = 0,$$

which has solutions $\lambda_1 = 1$ (necessarily, since it is a Markov chain) and $\lambda_2 = 1/30$. We calculated in Example 3.2 that the stationary distribution is $\vec{\pi} = \begin{bmatrix} 24/29, & 5/29 \end{bmatrix}$. Consequently,

$$\Delta(t) = |\lambda_2|^t \max\{\pi_1, \pi_2\} = \frac{24}{29}\left(\frac{1}{30}\right)^t.$$

In this case, already at $t = 1$, $\Delta(t)$ is far less than $1/2e$!!! So here, $\tau_{\text{mix}} = 1$: convergence is essentially instantaneous.

The more general two-state case is Example 3.1, with the machine that either works or is broken.

Example 3.29 (Two-state machine, revisited)**.** The transition matrix is

$$\mathbf{P} = \begin{bmatrix} 1-p & p \\ q & 1-q \end{bmatrix}$$

which has eigenvalues $\lambda_1 = 1$ and $\lambda_2 = 1 - p - q$, and stationary distribution

$$\vec{\pi} = \begin{bmatrix} \frac{q}{p+q}, & \frac{p}{p+q} \end{bmatrix}.$$

Thus,

$$\Delta(t) = |\lambda_2|^t \max\{\pi_1, \pi_2\} = |1 - p - q|^t \frac{\max\{p, q\}}{p + q},$$

and

$$\tau_{\text{mix}} = \left\lceil \frac{1 + \ln 2 + \ln \max\{p, q\} - \ln(p + q)}{-\ln |1 - p - q|} \right\rceil.$$

We have not needed to calculate explicitly the second eigenvector $\vec{a}_2$ for any of this analysis, but it nevertheless instructive to do so. Take the eigenvector expansion

$$\vec{\pi}(t) = \vec{\pi} + c_2(\lambda_2)^t \vec{a}_2,$$

and sum over the two states:

$$\pi_1(t) + \pi_2(t) = \pi_1 + \pi_2 + c_2(\lambda_2)^t[(a_2)_1 + (a_2)_2].$$

Normalization sets $\pi_1(t) + \pi_2(t) = 1$ and $\pi_1 + \pi_2 = 1$. Therefore, $c_2(\lambda_2)^t[(a_2)_1 + (a_2)_2]$ must be zero, so

$$\vec{a}_2 = \begin{bmatrix} 1, & -1 \end{bmatrix}$$

up to an arbitrary constant of proportionality. If we want to find the constant c_2 as well, that is determined by the initial state $\vec{\pi}(0)$, since

$$\vec{\pi}(0) = \vec{\pi} + c_2 \begin{bmatrix} 1, & -1 \end{bmatrix},$$

so

$$c_2 = \pi_1(0) - \pi_1 = \pi_2 - \pi_2(0).$$

All of this is for a Markov chain with two states. What if there are more than two states? We can still use λ_2 to bound the mixing time, as expressed by the following theorem.

Theorem 3.30. *Let* $\pi_{\min} = \min_{x \in \Omega} \pi_x$. *Then*

$$\tau_{\text{mix}} \leq \left\lceil \frac{1 + \ln 2 - \ln \pi_{\min}}{1 - |\lambda_2|} \right\rceil.$$

While we will not prove the theorem here, a few observations are in order. First of all, we can confirm that this bound is consistent with the mixing time we calculated for the two-state Markov chain. Since the logarithm of a probability can never be positive,

$$\ln \max\{\pi_1, \pi_2\} \leq -\ln \pi_{\min},$$

and it is always true that

$$\frac{1}{-\ln |\lambda_2|} \leq \frac{1}{1 - |\lambda_2|}.$$

Second of all, the theorem illustrates that the convergence time of the Markov chain is not only governed primarily by λ_2, but is even formally bounded by it. More specifically, since $\lambda_1 = 1$, the bound uses the *spectral gap* $|\lambda_1| - |\lambda_2|$. If that gap is large, the mixing time will be small, and the Markov chain will converge rapidly.

Our entire discussion on convergence time has assumed, not surprisingly, that the Markov chain does actually converge. But say we have a Markov chain that is not ergodic. How is that reflected in the spectral analysis? Consider the following dramatic example.

Example 3.31. At a certain company, each employee works every other day. Define

- State 1: the employee works
- State 2: the employee is off

It is rather silly to model this as a Markov chain, since everything is completely deterministic, but let us try anyway. The transition matrix is

$$\mathbf{P} = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix},$$

which of course identical to the Markov chain with the two-state machine, if $p = q = 1$. This Markov chain has a period of 2 and cannot possibly be ergodic: its stationary distribution is $\vec{\pi} = \begin{bmatrix} 1/2, & 1/2 \end{bmatrix}$, but in general it will never converge. How can we see this in its eigenvalues? Here, $\lambda_2 = -1$, so

$$\Delta(t) = |-1|^t \max\{\pi_1, \pi_2\} = \max\{\pi_1, \pi_2\},$$

i.e., $\Delta(t)$ remains constant and does not decay as t increases. Furthermore,

$$\vec{\pi}(t) = \vec{\pi} + (-1)^t[\vec{\pi}(0) - \vec{\pi}].$$

So for instance, if 60% of employees work on the first day, then $\vec{\pi}(0) - \vec{\pi} = \begin{bmatrix} 0.6 - 0.5, & 0.4 - 0.5 \end{bmatrix} = \begin{bmatrix} 0.1, & -0.1 \end{bmatrix}$, and

$$\vec{\pi}(t) = \begin{bmatrix} 0.5 + (-1)^t 0.1, & 0.5 - (-1)^t 0.1 \end{bmatrix}.$$

In the long run, we expect 50% of employees to work on any given day, but we will never approach this stationary distribution. We keep cycling between 60% of employees working and 40% of employees working. Negative eigenvalues always result in oscillatory behavior, and when the eigenvalue is -1, that is sufficient to prevent convergence altogether. More generally, periodic behavior reflects the presence of an eigenvalue λ where $|\lambda| = 1$ but $\lambda \neq 1$: if the period is k, λ is the kth root of unity. Clearly, if $k > 2$, this implies that λ is complex.

3.5.3 Strong stationary time

Another general technique for finding (or at least bounding) the mixing time of a Markov chain is as follows. Suppose there is a certain identifiable event, signaling that the Markov chain has converged. If we can express the probability of this event taking place by time t, we can bound how close we are (in probability) to convergence at time t. That is the idea behind a strong stationary time.

Definition 3.32 (Strong stationary time (SST))**.** *Given a Markov chain with states $X_0, \ldots, X_t$, the SST is a random variable T satisfying*

$$\forall x, \quad \Pr[X_t = x | T = t] = \pi_x.$$

T is then a "stopping time" at which we are sure to have converged to the stationary distribution.

How do we relate T to the mixing time? It can be shown that

$$\Delta(t) \leq \Pr[T > t],$$

which is not surprising since $\Delta(t)$ and $\Pr[T > t]$ are both bounded above by 1, and the Markov chain is close to convergence for values of t where $\Pr[T > t]$ is small, so $\Delta(t)$ is small there as well. But then, given the inequality, all we need is knowledge of the tail of the distribution of T: if $\Pr[T > t] \leq 1/2e$, it follows that $\Delta(t) \leq 1/2e$, and

$$\tau_{\text{mix}} \leq t.$$

To demonstrate the use of the SST, we return to the problem of card shuffling, and now ask the question of *how many* shuffles are needed for the deck to be sufficiently well mixed.

Top-in-at-random shuffle

In order to understand the SST for this kind of shuffle, consider the card b that is initially at the bottom of the deck. As we repeatedly remove the top card and reinsert it randomly, b will slowly rise through the deck. But as it rises, everything *below* it will have been completely randomized. So once b arrives at the top of the deck, we perform one last step of removing it and reinserting it at random, and then the entire deck will be completely randomized. All memory of the initial configuration will be erased, and we will have converged to the stationary distribution. The SST is therefore

$$\begin{aligned} T &= 1 + \text{time until } b \text{ reaches top of deck} \\ &= T_1 + \cdots + T_{n-1} + 1, \end{aligned}$$

where

$$T_i = \text{time to rise from position } i \text{ to position } i+1.$$

T_i is a geometric random variable: $\Pr[T_i = t]$ is the probability that the top card will be placed in one of the $n-i$ upper positions $t-1$ times, followed by the top card being placed in one of the i lower positions once. Thus,

$$\Pr[T_i = t] = \left(\frac{n-i}{n}\right)^{t-1} \frac{i}{n}.$$

From this, we can explicitly compute the probability $\Pr[T = t]$. Fortunately, the problem has already been solved for us. It is precisely the **coupon collector problem**: at each time step, we receive one out of n possible coupons *with replacement*, and we ask how many steps it will take until we have gotten at least one of all n coupons. The time it takes between getting coupon i and coupon $i+1$ is precisely T_{n-i}, so the total time to receive all coupons is T.

For the coupon collector problem, it is known that a simple union bound argument gives

$$\Pr[T > n \ln n + cn] \leq e^{-c}$$

for any $c > 0$. Letting $c = 2$,

$$\Pr[T > n \ln n + 2n] \leq e^{-2} < \frac{1}{2e},$$

which, from our earlier argument, implies that

$$\tau_{\text{mix}} < n \ln n + 2n.$$

So thanks to the SST, we are able to obtain a quantitative bound on how long it takes for this process to produce a "sufficiently random" deck of cards. Substituting $n = 52$, what we find is not surprising: $\tau_{\text{mix}} < 310$, so top-in-at-random could be a very slow way to shuffle!

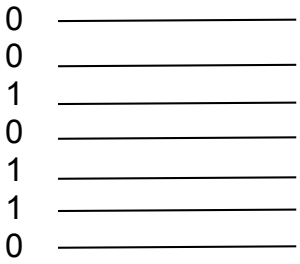
Riffle shuffle

Let us skip directly to the final kind of shuffling: the riffle shuffle, where we cut the deck into two stacks and then interleave them. What is the SST in this case, and how many shuffles are needed to randomize the deck?

As you will show in the homework, this Markov chain is doubly stochastic: it is a valid Markov chain when run in both the forwards and reverse direction. Furthermore, the SST is identical in both directions. The trick is then to analyze the mixing time for the reverse Markov chain.

The reverse process amounts to the following:

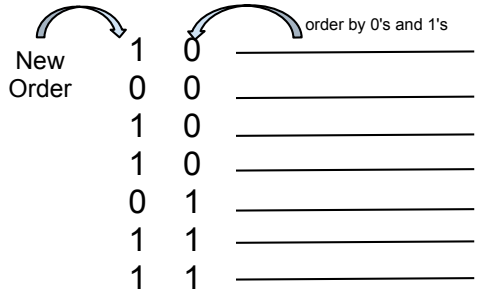
1. Randomly assign a digit 0 or 1 to each card in the deck:

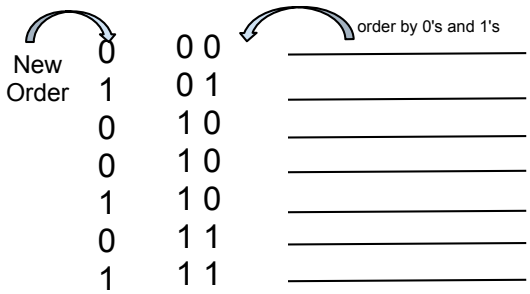


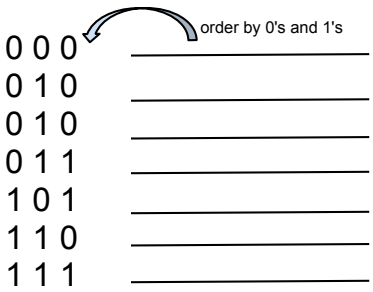
2. Separate the deck into one stack containing cards labeled 0 and the other stack containing cards labeled 1, while preserving the relative order of cards within each stack. Alternatively, one can think of this step as randomly dealing out the cards, in order, to player "0" and player "1."

3. Place the 0 stack on top of the 1 stack.

4. Repeat the random digit assignment at step 1, but now "prepending" this new random digit to the existing label:







After t steps, each card will be labeled with a t-bit string of 0s and 1s. That is simply a binary number. But notice that any two cards with the *same* number are still in their *original* relative order and have not yet had the opportunity to be exchanged. So only when all n cards have n *distinct* numbers is the configuration completely random. We have therefore found our SST:

$$T = \text{time for all cards to have distinct numbers.}$$

Note, first of all, a *lower bound* on convergence time. At time step t, there are 2^t possible distinct numbers. So if $t \leq 5$, at most 32 distinct numbers are available. But in that case, given $n = 52$ cards, the pigeonhole principle states that at least two of them will need to have the same number! So in 5 shuffles or less, *it is completely impossible to randomize the deck of cards:* certain configurations will simply be inaccessible. This conclusion came as quite a surprise to certain casinos that had routinely been using only 4 riffle shuffles to mix their decks.

If 5 shuffles are not enough, how many are? Since each step involves a assignment of labels 0 and 1 uniformly at random to each card, the t-step process is equivalent to assigning each card one of the 2^t possible numbers uniformly at random. The probability that each card will have a distinct number at time t is identical to the **birthday problem**: given n people in a room, how likely is it that no two will share the same birthday?

For the case of birthdays, a straightforward application of the product rule gives

$$\begin{aligned} \mathbb{P}\text{r}[\text{distinct birthdays}] &= \frac{365}{365} \cdot \frac{364}{365} \cdots \frac{365-n+1}{365} \\ &= \left(1 - \frac{1}{365}\right)\left(1 - \frac{2}{365}\right) \cdots \left(1 - \frac{n-1}{365}\right) \end{aligned}$$

For the case of card shuffling, instead of 365 possible days we have 2^t possible numbers, so given n cards,

$$\mathbb{P}\text{r}[\text{distinct numbers}] = \left(1 - \frac{1}{2^t}\right)\left(1 - \frac{2}{2^t}\right) \cdots \left(1 - \frac{n-1}{2^t}\right).$$

Expanding the polynomial and keeping only the leading order terms gives a bound:

$$\begin{aligned}\mathbb{P}\mathrm{r}[\text{distinct numbers}] &= 1 - \sum_{i=1}^{n-1} \frac{i}{2^t} + \cdots \\ &> 1 - \sum_{i=1}^{n-1} \frac{i}{2^t} \\ &= 1 - \frac{n(n-1)}{2} 2^{-t} \\ &> 1 - \frac{n^2}{2} 2^{-t}.\end{aligned}$$

This is the probability that convergence occurs by time t, so

$$\mathbb{P}\mathrm{r}[T > t] < \frac{n^2}{2} 2^{-t}.$$

If we let $t = \log_2(n^2 e)$, then

$$\mathbb{P}\mathrm{r}[T > \log_2(n^2 e)] < \frac{1}{2e},$$

and consequently,

$$\tau_{\mathrm{mix}} < \log_2(n^2 e) = \frac{1 + 2\ln n}{\ln 2}.$$

We obtain a dramatic improvement over top-in-at-random: the mixing time is reduced from $\Theta(n \ln n)$ to $\Theta(\ln n)$. Numerically, for $n = 52$, we have $\tau_{\mathrm{mix}} < 13$.

With a little bit of work, one can even improve the bound. In 1983, Aldous showed that asymptotically (for large n), $\tau_{\mathrm{mix}} \approx \log_2 n^{3/2}$. Subsequently, in a delightful 1992 paper full of discussions of legendary magicians and their card tricks ("Trailing the dovetail shuffle to its lair"), Bayer and Diaconis found $\Delta(t)$ explicitly for $n = 52$, obtaining the numerical values

t	1	2	3	4	5	6	7	8	9
$\Delta(t)$	1.00	1.00	1.00	1.00	0.92	0.61	0.33	0.17	0.09

Using a $\Delta(t)$ threshold value of $1/2$ rather than $1/2e$ for the mixing time led to the well-publicized conclusion that, in practice, 7 riffle shuffles are sufficient to mix a deck of cards.