A Primer on Discrete Time Markov Chains

Pramook Khungurn

January 16, 2022

This is a primer on discrete-time Markov chains. The materials are taken from a number of sources [Häggström, 2002, Lee, 2012b, Lee, 2012a, Kennedy, 2016, Tolver, 2016]. To keep the note relatively short, later sections will contain fewer and fewer proofs. This means that the treatment is the most thorough in the finite state space case and becomes more cursory as the state space becomes more complex.

1 Finite State Spaces

1.1 Basic Definitions and Properties

- **Definition 1.** *A (discrete-time)* **stochastic process** *is a sequence of random variables* $(X_0, X_1, \dots)$.
- In this section, we are interested in stochastic processes on a finite state space $S = \{s_1, s_2, \dots, s_k\}$. In other words, $X_i \in S$ for all i.
- **Definition 2.** *We say that a stochastic process is* **memoryless** *if, for all $t \geq 1$, the probability distribution of X_t depends only on the value of X_{t-1} and not any other past random variables. In other words,*

$$\Pr(X_t = s_j | X_0 = s_{i_0}, X_1 = s_{i_1}, \dots, X_{t-2} = s_{i_{t-1}}, X_{t-1} = s_i) = \Pr(X_t = s_j | X_{t-1} = s_i).$$

- **Definition 3.** *A* **Markov chain** *is a memoryless stochastic process.*
- **Proposition 4 (Chapman–Kolmogorov equation).** *For any $t \geq 0$, $a, b \geq 1$, and any i and j, we have that*

$$\Pr(X_{t+a+b} = s_j | X_t = s_i) = \sum_{\ell=1}^{k} \Pr(X_{t+a+b} = s_j | X_{t+a} = s_\ell) \Pr(X_{t+a} = s_\ell | X_t = s_i).$$

  (It basically says that, in order to get from s_i at time t to s_j at time $t + a + b$, we must pass through an intermediate state s_ℓ at time $t + a$.)
- **Definition 5.** *A Markov chain is said to be* **time homogenous** *if the transition probabilities does not depend on time. In other words,*

$$\Pr(X_t = s_j | X_{t-1} = s_i) = \Pr(X_{t'} = s_j | X_{t'-1} = s_i)$$

  for any t, t', i and j.
- We are only interested in time homogenous Markov chains in this note. So, when we mention a Marko chain, it is always time homogenous.

- **Definition 6.** *For a time-homogenous Markov chain, its* **transition matrix** $P \in \mathbb{R}^{k \times k}$ *is the matrix whose entries are given by:*

$$P_{ij} = \Pr(X_t = s_j | X_{t-1} = s_i)$$

for any $t \geq 1$.

- The transition matrix has the following properties.

  1. Its entries are non-negative.
  2. Entries in each row add up to 1. In other words, $\sum_{j=1}^{k} P_{ij} = 1$ for all i.

- For a time-homogeneous Markov chain, the Chapman–Kolmogorov equation can be rewritten as simply $P^{a+b} = P^a P^b$.

- We can represent the distribution of X^t by a row vector $\boldsymbol{\mu}^{(t)}$ with

$$\boldsymbol{\mu}^{(t)} = (\mu_1^{(t)}, \mu_2^{(t)}, \ldots, \mu_k^{(t)}) = \Big(\Pr(X_t = s_1), \Pr(X_t = s_2), \ldots, \Pr(X_t = s_k) \Big).$$

- **Proposition 7.** *We have that*

$$\boldsymbol{\mu}^{(t)} = \boldsymbol{\mu}^{(0)} P^t.$$

for all $t \geq 0$.

1.2 Asymptotic Behavior

- We are now interested in determining the long term behavior of a Markov chain. In other words, as $t \to \infty$, what does $\boldsymbol{\mu}^{(t)}$ look like? Does it converge to anything? If so, under what conditions?

- The conclusion is this: if the Markov chain is "irreducible" and "aperiodic," there is a unique "stationary distribution" that $\boldsymbol{\mu}^{(t)}$ converges to as $t \to \infty$ no matter what $\boldsymbol{\mu}^{(0)}$ is. We will discuss what the quoted terms mean in this section.

1.2.1 Irreducibility

- **Definition 8.** *We say that a state* s_i **communicates** *with another state* s_j *if there is a positive probability that we can reach* s_j *starting from* s_i. *In other words, there exists an* a *such that*

$$\Pr(X_{t+a} = s_j | X_t = s_i) > 0$$

for all $t \geq 0$. *(Notice that the choice of* t *is irrelevant here because a Markov chain is memoryless and time homogeneous.) We write* $s_i \to s_j$ *to denote the fact that* s_i *communicates with* s_j.

- **Definition 9.** *If* $s_i \to s_j$ *and* $s_j \to s_i$, *we say that* s_i *and* s_j **intercommunicates**, *and we denote this fact by* $s_i \leftrightarrow s_j$. *Moreover, the relation "intercommunicates with" between states is an equivalence relation. So, it partitions* $\mathcal{S}$ *into equivalence classes which are called* **communication classes**.

- **Definition 10.** *A Markov chain is said to be* **irreducible** *if for all* $s_i, s_j \in S$, *we have that* $s_i \leftrightarrow s_j$. *(In other words, it has only one communication class.) Otherwise, we say that the Markov chain is* **reducible**.

- Given a Markov chain, we can construct its **transition graph**. This is a directed graph where (1) S is the vertex set, and (2) there is a directed edge from s_i to s_j if $P_{ij} > 0$. We often assign the weight of P_{ij} to such an edge.

- **Proposition 11.** *We have that $s_i \to s_j$ if and only if there is a directed path from s_i to s_j in the transition graph.*

- **Corollary 12.** *Communication classes are strongly connected components*[1] *in the transition graph. (That is, there is a directed path betweeen every pair of states in a communication class.) A Markov chain is irreducible if and only if the entire transition graph is strongly connected.*

1.2.2 Aperiodity

- For any countable set $\{a_1, a_2, \dots\}$ of positive integers, we use $\gcd\{a_1, a_2, \dots\}$ to denote the greatest common divisor of $a_1, a_2, \dots$.

- **Definition 13.** *For $s_i \in S$, the* **period** *$d(s_i)$ is defined as:*

$$d(s_i) = \gcd\{n \geq 1 : (P^n)[i,i] > 0\}.$$

In other words, it is the greatest common divisor of the number of steps where it is a positive probability to return to s_i after starting from s_i.

- **Theorem 14.** *All states in a communication class has the same period.*

Proof. For any i, let

$$D_i = \{n \geq 1 : \text{there is a loop starting from } s_i \text{ to } s_i\}.$$

We have that $d(s_i) = \gcd(D_i)$.

Let s_i and s_j be in the same communication class. Choose $k, l \in \mathbb{N}$ such that $(P^k)[i,j] > 0$ and $(P^l)[j,i] > 0$. Let $m = k + l$. We have that $m \in D_i$ and $m \in D_j$. Now, for any $n \in D_i$, we have that there is a loop of length $n+m$ from s_j to s_j, and so $n+m \in D_j$. As a result, by the definition of $d(s_j)$, we have that $d(s_j)|m$ and $d(s_j)|(n+m)$, so it must be the case that $d(s_j)|n$. We have now shown that $d(s_j)$ divides all the numbers of D_i, so $d(s_j) \leq d(s_i)$ because $d(s_i)$ is the greatest number dividing all numbers in D_i. By swapping i and j in the argument we just used, we can show that $d(s_i) \leq d(s_j)$ as well. So, we can conclude that $d(s_i) = d(s_j)$. □

- **Definition 15.** *A state s_i is said to be* **aperiodic** *if $d(s_i) = 1$. A Markov chain is said to be* **aperiodic** *if all of its states are aperiodic. Otherwise, it is said to be* **periodic**.

- **Theorem 16.** *For an aperiodic Markov chain, there exists an $N < \infty$ such that $(P^n)[i,i] > 0$ for all $i \in \{1, 2, \dots, k\}$ and $n \geq N$.*

- The proof of the above theorem uses the following result from number theory.

- **Lemma 17.** *Let $A = \{a_1, a_2, \dots\}$ be a countable set of positive integers with the following properties:*

  1. $\gcd(A) = \gcd\{a_1, a_2, \dots\} = 1$.
  2. *For any $x, y \in A$, we have that $x + y \in A$ as well.*

Then, there is an integer N such that, if $n \geq N$, then $n \in A$.

Proof. Let us assume WLOG that $a_1 < a_2 < a_3 < \cdots$. Define g_i to be $\gcd\{a_1, a_2, \dots, a_i\}$. We have that $a_1 = g_1 \geq g_2 \geq g_3 \geq \cdots \geq 1$. Because $1 = \gcd(A) = \min\{g_1, g_2, \dots\}$, there must be an m such that $g_m = 1$ because $\{g_1, g_2, \dots\}$ is a non-empty set of positive integers so it must contain the minimal element according to the well-ordering principle.[2]

[1] `https://en.wikipedia.org/wiki/Strongly_connected_component`
[2] `https://en.wikipedia.org/wiki/Well-ordering_principle`

Because $\gcd\{a_1, a_2, \ldots, a_m\} = 1$, we can find $c_1, c_2, \ldots, c_m \in \mathbb{Z}$, some of which may not be positive, such that

$$c_1 a_1 + c_2 a_2 + \cdots + c_m a_m = 1.$$

Now, we pick

$$N := a_1 \sum_{i=1}^{m} |c_i| a_i.$$

We will now show that $n \in A$ for any $n \geq N$. To do this, we will first show that N, $N+1$, $N+2$, …, $N + a_1 - 1$ are all members of A. So, let's consider $N + r$ where $0 \leq r < a_1$. We have that

$$N + r = a_1 \sum_{i=1}^{m} |c_i| a_i + r\left(\sum_{i=1}^{m} c_i a_i\right) = \sum_{i=1}^{m} (a_1 |c_i| + r c_i) a_i.$$

Note that

$$a_1 |c_i| + r c_i \geq a_1 |c_i| - r|c_i| = (a_1 - r)|c_i| \geq |c_i| \geq 0.$$

As a result, $N + r \in A$ because it can be written as a linear combination of a_1, a_2, …, a_m with non-negative coefficients.

Lastly, we have that, for any $n \geq N + a_1$, we can find $c \geq 1$ and $0 \leq r < a_1$ such that $n = N + r + c a_1$. Because $N + r \in A$ and $c a_1 \in A$, we have that $n \in A$ as well. □

- *Proof of Theorem 16.* Define $A_i = \{n \geq 1 : (P^n)[i, i] > 0\}$. Because s_i is aperiodic, we have that $\gcd(A_i) = 1$. Now, let $x, y \in A_i$. By the Chapman–Kolmogorov equation, we have that,

$$(P^{x+y})[i, i] = \sum_{j=1}^{k} (P^x)[i, j](P^y)[j, i] \geq (P^x)[i, i](P^y)[i, i] > 0,$$

so $x + y \in A_i$ as well. Applying Lemma 17 to A_i, we have that there exists N_i such that $(P^n)[i, i] > 0$ for all $n \geq N_i$. We can now pick $N = \max\{N_1, N_2, \ldots, N_k\}$. □

- **Corollary 18.** *For an irreducible and aperiodic Markov chain, there exists an integer $M < \infty$ such that $(P^n)[i, j] > 0$ for all $i, j \in \{1, \ldots, k\}$ and $n \geq M$.*

1.2.3 Hitting Time

- **Definition 19.** *The* **hitting time** T_i *is the random variable*

$$T_i = \min\{t \geq 1 : X_t = s_i\}.$$

That is, it is the first time that we reach s_i. Also, define

$$T_{i,j} = \min\{t \geq 1 : X_t = s_j | X_0 = s_i\}$$

to be the first time we reach s_i after starting from s_j. We say that $T_i = \infty$ if we never reach s_i, and $T_{i,j} = \infty$ if we never reach s_j from s_i.

- Note that, if $i \neq i'$, we have that $T_{i,j}$ and $T_{i',j}$ are defined on different probability spaces. So, they should not be mixed.

- **Definition 20.** *The* **mean hitting time** $\tau_{i,j}$ *is the expected value of the hitting time $T_{i,j}$:*

$$\tau_{i,j} = E[T_{i,j}] = E[T_j | X_0 = x_i].$$

- **Lemma 21.** *For any irreducible and aperiodic Markov chain with finite state space, we have that $P(T_{i,j} < \infty) = 1$ for any two states s_i and s_j. Moreover, $\tau_{i,j} = E[T_{i,j}] < \infty$.*

  Proof. By Corollary 18, there exists an integer M such that all the entries of the matrix P^M are strictly positive. Let $\alpha > 0$ be the minimum entry in the matrix P^M. So, for any two states s_i and s_j, we have that

$$\Pr(T_{i,j} > M) \leq \Pr(X_M \neq s_j) \leq 1 - \alpha.$$

  Now, for any value of X_M, there is a probability of at least α that $X_{2M} = s_j$. Thus, for any event A that only concerns outcomes up to X_M, we have that

$$\begin{aligned}
\Pr(T_{i,j} > 2M | T_{i,j} > M) &\leq \Pr(X_{2M} \neq s_j | T_{i,j} > M) \\
&= 1 - \Pr(X_{2M} = s_j | T_{i,j} > M) \\
&= 1 - \sum_{s_\ell} \Pr(X_{2M} = s_j | X_M = s_\ell) \Pr(X_M = s_\ell | T_{i,j} > M) \\
&\leq 1 - \sum_{s_\ell} \alpha \Pr(X_M = s_\ell | T_{i,j} > M) \\
&= 1 - \alpha.
\end{aligned}$$

  So,

$$\Pr(T_{i,j} > 2M) = \Pr(T_{i,j} > M) \Pr(T_{i,j} > 2M | T_{i,j} > M) \leq (1 - \alpha)^2.$$

  Repeating the argument, we can conclude that

$$\Pr(T_{i,j} > \ell M) = (1 - \alpha)^\ell.$$

  As a result, $\lim_{\ell \to \infty} \Pr(T_{i,j} > \ell M) = 0$, and $\Pr(T_{i,j} < \infty) = 1$.

  Next, we have that

$$\begin{aligned}
\tau_{i,j} = E[T_{i,j}] &= \sum_{n=1}^{\infty} \Pr(T_{i,j} \geq n) = \sum_{n=0}^{\infty} P(T_{i,j} > n) \\
&= \sum_{\ell=0}^{\infty} \sum_{n=\ell M}^{(\ell+1)M-1} \Pr(T_{i,j} > n) \\
&\leq \sum_{\ell=0}^{\infty} \sum_{n=\ell M}^{(\ell+1)M-1} \Pr(T_{i,j} > \ell M) = M \sum_{\ell=0}^{\infty} \Pr(T_{i,j} > \ell M) \\
&\leq M \sum_{\ell=0}^{\infty} (1 - \alpha)^\ell = \frac{M}{\alpha} < \infty
\end{aligned}$$

  as required. □

1.2.4 Stationary Distribution

- **Definition 22.** *A distribution, represented by the row vector $\boldsymbol{\pi}$, is said to be a* **stationary distribution** *of a Markov chan if*

$$\boldsymbol{\pi} P = \boldsymbol{\pi}.$$

- **Theorem 23.** *For an irreducible and aperiodic Markov chain over a finite state space, there exists a stionary distribution.*

Proof. Assume that the Markov chain always starts at state s_1. Because the Markov chain is irreducible, we have that it will return to s_1 eventually according to Theorem 1.2.8, and the number of steps taken is $T_{1,1}$ according to the terminology of the last section. Once it returns to s_1, it is like starting over from scratch again and the cycle repeats. (So, it will return to s_1 an infinity number of times.)

The trick is to look in each cycle of departing from s_1 and returning to it. For each state s_i, we can count the number of times R_i the process spends the state. The expected value of the ratio $R_i/T_{1,1}$ should remain unchanged if we apply one more transition step, and this will give us the stationary distribution.

More formally, define the random variable R_i to be the number of times the Markov chain hits state s_i before it returns to s_1 for the first time. Symbolically,

$$R_i = \#\{t : X_t = s_i \wedge t < T_{1,1}\}.$$

We have that

$$R_i = \sum_{t=0}^{\infty} I(X_t = s_i \wedge t < T_{1,1})$$

where $I(\cdot)$ is the indicator function. Also,

$$\rho_i = E[R_i] = \sum_{t=0}^{\infty} \Pr(X_t = s_i \wedge t < T_{1,1}).$$

Note that

$$\begin{aligned}
\sum_{i=1}^{k} \rho_i &= \sum_{i=1}^{k} \sum_{t=0}^{\infty} \Pr(X_t = s_i \wedge t < T_{1,1}) \\
&= \sum_{t=0}^{\infty} \sum_{i=1}^{k} \Pr(X_t = s_i \wedge t < T_{1,1}) \\
&= \sum_{t=0}^{\infty} \Pr(X_t = s_i \wedge t < T_{1,1}) \\
&= E[T_{1,1}] \\
&= \tau_{1,1}.
\end{aligned}$$

By Lemma 21, we have that $\tau_{1,1}$ is finite. Hence, we can define

$$\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots, \pi_k) = \left(\frac{\rho_1}{\tau_{1,1}}, \frac{\rho_2}{\tau_{1,1}}, \ldots, \frac{\rho_k}{\tau_{1,1}} \right).$$

It remains to show that $\boldsymbol{\pi} P = \boldsymbol{\pi}$. To do so, we have to show that

$$\pi_j = \sum_{i=1}^{k} \pi_i P_{ij}$$

for all j. There are two cases. First, when $j \neq 1$, we have that

$$\begin{aligned}
\pi_j &= \frac{\rho_j}{\tau_{1,1}} \\
&= \frac{1}{\tau_{1,1}} \sum_{t=0}^{\infty} \Pr(X_t = s_j \wedge t < T_{1,1}) \\
&= \frac{1}{\tau_{1,1}} \sum_{t=1}^{\infty} \Pr(X_t = s_j \wedge t < T_{1,1}) \\
&= \frac{1}{\tau_{1,1}} \sum_{t=1}^{\infty} \Pr(X_t = s_j \wedge t - 1 < T_{1,1}) \\
&= \frac{1}{\tau_{1,1}} \sum_{t=1}^{\infty} \sum_{i=1}^{k} \Pr(X_t = s_j \wedge X_{t-1} = s_i \wedge t - 1 < T_{1,1}) \\
&= \frac{1}{\tau_{1,1}} \sum_{t=1}^{\infty} \sum_{i=1}^{k} \Pr(X_t = s_j \wedge X_{t-1} = s_i \wedge t - 1 < T_{1,1}) \\
&= \frac{1}{\tau_{1,1}} \sum_{t=1}^{\infty} \sum_{i=1}^{k} \Pr(X_{t-1} = s_i \wedge t - 1 < T_{1,1}) \Pr(X_t = s_j | X_{t-1} = s_i) \\
&= \frac{1}{\tau_{1,1}} \sum_{t=1}^{\infty} \sum_{i=1}^{k} \Pr(X_{t-1} = s_i \wedge t - 1 < T_{1,1}) P_{ij} \\
&= \frac{1}{\tau_{1,1}} \sum_{i=1}^{k} \left(\sum_{t=1}^{\infty} \Pr(X_{t-1} = s_i \wedge t - 1 < T_{1,1}) \right) P_{ij} \\
&= \frac{1}{\tau_{1,1}} \sum_{i=0}^{k} \left(\sum_{t=0}^{\infty} \Pr(X_t = s_i \wedge t < T_{1,1}) \right) P_{ij} \\
&= \frac{1}{\tau_{1,1}} \sum_{i=0}^{k} \rho_i P_{ij} = \sum_{i=0}^{k} \frac{\rho_i}{\tau_{1,1}} P_{ij} = \sum_{i=0}^{k} \pi_i P_{ij}.
\end{aligned}$$

Next, when $j = 1$, we have that $\rho_1 = 1$.

$$\begin{aligned}
\rho_1 = 1 = \Pr(T_{1,1} < \infty) &= \sum_{t=1}^{\infty} \Pr(T_{1,1} = t) \\
&= \sum_{t=1}^{\infty} \sum_{i=1}^{k} \Pr(X_{t-1} = s_i \wedge T_{1,1} = t) \\
&= \sum_{t=1}^{\infty} \sum_{i=1}^{k} \Pr(X_{t-1} = s_i \wedge t - 1 < T_{1,1}) \Pr(X_t = s_1 | X_{t-1} = s_i) \\
&= \sum_{t=1}^{\infty} \sum_{i=0}^{k} \Pr(X_{t-1} = s_i \wedge t - 1 < T_{1,1}) P_{i1} \\
&= \sum_{i=1}^{k} \left(\sum_{t=1}^{\infty} \Pr(X_{t-1} = s_i \wedge t - 1 < T_{1,1}) \right) P_{i1} \\
&= \sum_{i=1}^{k} \left(\sum_{t=0}^{\infty} \Pr(X_t = s_i \wedge t < T_{1,1}) \right) P_{i1} \\
&= \sum_{i=1}^{k} \rho_i P_{i1}.
\end{aligned}$$

As a result,

$$\pi_1 = \frac{\rho_1}{\tau_{1,1}} = \frac{1}{\tau_{1,1}} \sum_{i=1}^{k} \frac{\rho_i}{\tau_{1,1}} P_{i1} = \sum_{i=1}^{k} \pi_i P_{i1}$$

as required. □

- We will now show that $\boldsymbol{\mu}^{(t)}$ converges to the stationary distribution $\boldsymbol{\pi}$ as $t \to \infty$ regardless of what $\boldsymbol{\mu}^{(0)}$ is. To do this, we need a notion of how two distributions would converge to each other.
- **Definition 24.** *Given two row vectors* $\mathbf{a} = (a_1, \ldots, a_k)$ *and* $\mathbf{b} = (b_1, \ldots, b_k)$*, the* **total variation distance** *between* $\mathbf{a}$ *and* $\mathbf{b}$ *is given by:*

$$d_{\mathrm{TV}}(\mathbf{a}, \mathbf{b}) = \frac{1}{2} \sum_{i=1}^{k} |a_i - b_i|.$$

- It follows that, if $d_{\mathrm{TV}}(\mathbf{a}, \mathbf{b}) = 0$, then $\mathbf{a} = \mathbf{b}$.
- **Definition 25.** *A sequence of probability distributions* $\boldsymbol{\nu}^{(0)}$, $\boldsymbol{\nu}^{(1)}$, ... *is said to* **converge to a distribution** $\boldsymbol{\nu}$ **in total variation** *if*

$$\lim_{t \to \infty} d_{\mathrm{TV}}(\boldsymbol{\nu}^{(t)}, \boldsymbol{\nu}) = 0.$$

  We denote this by $\boldsymbol{\nu}^{(t)} \xrightarrow{\mathrm{TV}} \boldsymbol{\nu}$.
- Before we prove the convergence theorem for Markov chains, we need to introduce the notion of a **simulation** of a Markov chain. This is just basically how you would like a computer program to do it. More precisely:
  - We have a source of randomness that gives us an infinite sequence of random variables $(\xi_0, \xi_1, \ldots)$ where ξ_i is i.i.d. sampled uniformly from the interval $[0, 1]$.

– We have a function $\psi : [0,1] \to S$ that maps ξ_0 to the initial state X_0. In other words, we start the simulation with:

$$X_0 \leftarrow \psi(\xi_0).$$

Here, we require that $\Pr(X_0 = s_i)$ be equal to $\mu_i^{(0)}$ for our choice of $\boldsymbol{\mu}^{(0)}$.

– Next, we have another function $\phi : S \times [0,1] \to S$ that maps (X_{t-1}, ξ_t) to X_t in such a way that agrees with the transition matrix P. In other words, having sampled X_{t-1}, we compute X_t as follows:

$$X_t \leftarrow \phi(X_{t-1}, \xi_t)$$

In this way, the Markov chain can be viewed as a function that turns $(\xi_0, \xi_1, \dots)$ to $(X_0, X_1, \dots)$.

- **Theorem 26.** *Consider a irreducible, aperiodic Markov chain on a finite state space with whose stationary distribution is denoted by $\boldsymbol{\pi}$. Starting from an arbitrary distribution $\boldsymbol{\mu}^{(0)}$, we have that*

$$\boldsymbol{\mu}^{(t)} \xrightarrow{\text{TV}} \boldsymbol{\pi}.$$

Proof. Consider two simulations of the Markov chain using two sequences of random numbers $(\xi_0, \dots)$ and $(\xi'_0, \dots)$ where all random numbers are independent of any other numbers. Let us say that the simulations result in two outcomes $(X_0, X_1, \dots)$ and $(X'_0, X'_1, \dots)$, respectively.

For the first outcome $(X_0, X_1, \dots)$, we pick the initial function ψ such that

$$\Pr(X_0 = s_i) = \Pr(\psi(\xi_0) = s_i) = \mu_i^{(0)}.$$

For the second outcome, we pick the initial function ψ' so that

$$\Pr(X'_0 = s_i) = \Pr(\psi'(\xi'_0) = s_i) = \pi_i.$$

Note that the transition function ϕ and ϕ' for the two outcomes are exactly the same, so we will just denote them with ϕ. We also have that, if we denote the distribution X'_t with $\boldsymbol{\pi}^{(t)}$, we have that $\boldsymbol{\pi}^{(t)} = \boldsymbol{\pi}$ for all $t \geq 0$ because $\boldsymbol{\pi}$ is the stationary distribution.

We will show that the two simulation will "meet" with probability 1. More precisely, define the random variable T to the first meeting time:

$$T = \min\{t : X_t = X'_t\}.$$

We will show that $\Pr(T < \infty) = 1$.

Because our Markov chain is irreducible and periodic, there exists a positive integer M such that $(P^M)[i,j] > 0$ for i and j. Let α be the smallest entry if P^M. We have that $\alpha > 0$. We have that

$$\begin{aligned}
\Pr(T \leq M) &\geq \Pr(X_M = X'_M) \\
&\geq \Pr(X_M = s_1 \wedge X'_M = s_1) \\
&= \Pr(X_M = s_1) \Pr(X'_M = s_1) \\
&= \left(\sum_{i=1}^{k} P(X_0 = s_i \wedge X_M = s_1) \right) \left(\sum_{i=1}^{k} P(X'_0 = s_i \wedge X'_M = s_1) \right) \\
&= \left(\sum_{i=1}^{k} P(X_0 = s_i) \Pr(X_M = s_1 | X_0 = s_i) \right) \left(\sum_{i=1}^{k} P(X'_0 = s_i) \Pr(X'_M = s_1 | X'_0 = s_i) \right) \\
&= \left(\sum_{i=1}^{k} P(X_0 = s_i) \alpha \right) \left(\sum_{i=1}^{k} P(X'_0 = s_i) \alpha \right) \\
&= \alpha^2.
\end{aligned}$$

As a result,

$$\Pr(T > M) \leq 1 - \alpha^2.$$

Using the same argument that we used to derive a lower bound for $\Pr(T \leq M)$, we can show that

$$\Pr(X_{2M} = X'_{2M} | T > M) \geq \alpha^2,$$

and, as a result,

$$\Pr(X_{2M} \neq X'_{2M} | T > M) \leq 1 - \alpha^2.$$

So,

$$\begin{aligned}\Pr(T > 2M) &= P(T > M)P(T > 2M | T > M) \\ &\leq P(T > M) \Pr(X_{2M} \neq X'_{2M} | T > M) \\ &= (1 - \alpha^2)^2.\end{aligned}$$

Repeating the argument, we can conclude that

$$\Pr(T > \ell M) \leq (1 - \alpha^2)^\ell,$$

and so

$$\lim_{\ell \to \infty} \Pr(T > \ell M) = 0.$$

In other words, $\Pr(T < \infty) = 1$.

We now construct a new outcome $(X''_0, X''_1, \dots)$ such that

$$X''_t = \begin{cases} X_t, & \text{if } t < T \\ X'_t, & \text{if } t \geq T \end{cases}.$$

Note that the outcome can be easily generated from the following algorithm:

$X_0 \leftarrow \psi(\xi_0)$
$X'_0 \leftarrow \psi'(\xi'_0)$
flag $\leftarrow X_0 = X'_0$
for $t \leftarrow 0, 1, \dots$ **do**
  if flag **then**
    $X''_t \leftarrow X'_t$
  else
    $X''_t \leftarrow X_t$
  end if
  $X_{t+1} \leftarrow \phi(X_t, \xi_t)$
  $X'_{t+1} \leftarrow \phi(X_t, \xi'_t)$
  flag $\leftarrow$ flag $\wedge (X_t = X'_t)$
end for

We observe that the outcome $(X''_0, X''_1, \dots)$ is a Markov chain whose transition probabilities are given by the transition matrix P.

Next, we have that X''_0 has distribution $\boldsymbol{\mu}^{(0)}$. So, for any $t \geq 0$, X''_t has distribution $\boldsymbol{\mu}^{(t)}$. For any i, we have that

$$\mu_i^{(t)} - \pi_i = \Pr(X''_t = s_i) - \Pr(X'_t = s_i) \leq \Pr(X''_t = s_i \wedge X'_t \neq s_i) \leq \Pr(X''_t \neq X'_t) = \Pr(T > t).$$

Using the same argument, we can also say that $\pi_i - \mu_i^{(t)} \leq \Pr(T > t)$, and so $|\pi_i - \mu_i^{(t)}| \leq \Pr(T > t)$. Hence, $\lim_{t\to\infty} |\pi_i - \mu_i^{(t)}| = \lim_{t\to\infty} \Pr(T > t) = 0$, and we can finally conclude that

$$\lim_{t\to\infty} d_{\mathrm{TV}}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\pi}) = \lim_{t\to\infty} \sum_{i=1}^{k} |\pi_i - \mu_i^{(t)}| = \sum_{i=1}^{k} \lim_{t\to\infty} |\pi_i - \mu_i^{(t)}| = 0$$

as desired.

- **Theorem 27.** *The stationary distribution of an irreducible, aperiodic Markov chain on a finite state space is unique.*

  Proof. Suppose there are two stationary distributions $\boldsymbol{\pi}$ and $\boldsymbol{\pi}'$. We can start a Markov chain simulation with X_0 being distributed according to $\boldsymbol{\mu}^{(0)} = \boldsymbol{\pi}'$. By the last theorem, we have that

  $$0 = \lim_{t\to\infty} d_{\mathrm{TV}}(\boldsymbol{\mu}^{(t)}, \boldsymbol{\pi}) = \lim_{t\to\infty} d_{\mathrm{TV}}(\boldsymbol{\pi}', \boldsymbol{\pi}) = d_{\mathrm{TV}}(\boldsymbol{\pi}', \boldsymbol{\pi}),$$

  which means that $\boldsymbol{\pi}' = \boldsymbol{\pi}$. □

- The theorem above gives an alternative expression for the stationary distribution:

  $$\boldsymbol{\pi} = \left(\frac{1}{\tau_{1,1}}, \frac{1}{\tau_{2,2}}, \ldots, \frac{1}{\tau_{k,k}}\right)$$

  This is because, in the proof of Theorem 23, we can use any state s_i instead of s_1 to define the stationary probability. So, we have that

  $$\pi_i = \frac{\#\{t : X_t = s_i \wedge t < T_{i,i} | X_0 = s_i\}}{\tau_{i,i}} = \frac{1}{\tau_{i,i}}.$$

- When $\boldsymbol{\mu}^{(t)}$ converges (i.e., close enough to the stationary distribution), we say that the Markov chain is in **equilibrium**.

1.3 Reversibility

- **Definition 28.** *Consider a Markov chain with finite state space $S = \{s_1, \ldots, s_k\}$ and transition matrix P. A probability distribution $\boldsymbol{\pi}$ is* **reversible** *for the chain if, for all i, j, we have that*

  $$\pi_i P_{i,j} = \pi_j P_{j,i}. \tag{1}$$

  A Markov chain is reversible if it has a reversible distribution.

- The condition in Equation (1) is called **detailed balance**. The LHS can be interpreted as the probabilty mass going from s_i to s_j, and the RHS is the probability mass going in the opposite direction. This suggests a strong form of equilibrium.

- **Theorem 29.** *If a Markov chain with finite state space has a reversible distribution, then the distribution is also stationary.*

  Proof. Let $\boldsymbol{\pi}$ be a reversible distribution for the chain. We have that

  $$\pi_j = \pi_j \sum_{i=1}^{k} P_{j,i} = \sum_{i=1}^{k} \pi_j P_{j,i} = \sum_{i=1}^{k} \pi_i P_{i,j}$$

  for all j. This implies that $\boldsymbol{\pi}$ is stationary. □

- When a reversible Markov chain reaches equilibrium, it looks exactly the same whether time runs forward or backward.

- Reversible Markov chains show up a lot in the context of Markov chain Monte Carlo (MCMC) algorithms. This is why it is important to mention.

2 Countably Infinite State Spaces

- We are now concerned with a state space $S = \{s_1, s_2, \dots\}$ which is countable but infinite.
- The Chapman–Kolmogorov equation is pretty much the same:

$$\Pr(X_{t+a+b} = s_j | X_t = s_i) = \sum_{\ell=1}^{\infty} \Pr(X_{t+a+b} = s_j | X_{t+a} = s_\ell) \Pr(X_{t+a} = s_\ell | X_t = s_i).$$

- Obviously, the transition matrix P is now infinite.
- The detailed balance condition is still the same in the countably infinite state space. So, Theorem 29 holds in thsi case too.

2.1 Recurrence and Transience

- The definition for the hitting time T_i is the same:

$$T_i = \min\{t > 0 : X_t = s_i\}.$$

- However, properties of states with regards to the hitting times are more complicated in the countably infinite case. We need to classify states in new ways.
- **Definition 30.** *A state $s_i \in S$ is said to be* **recurrent** *if*

$$\Pr(T_{i,i} < \infty) = \Pr(T_i < \infty | X_0 = s_i) = 1.$$

Otherwise, the state is said to be **transient**. *We say a Markov chain is recurrent if all of its states are recurrent.*

- **Theorem 31 (Criterion for recurrence #1).** *For a Markov chain on a countably infinite state space, a state s_i is recurrent if and only if*

$$\sum_{i=t}^{\infty} (P^t)[i,i] = \infty.$$

- **Theorem 32.** *All states in a communication class are either all recurrent or all transient.*
- When the state space is finite, Lemma 21 implies that all states of an irreducible and aperiodic Markov chain are recurrent. However, this is not true for the countably infinite case: we can see from the last theorem that all states can be transient.
- Let N_i denote the total number of visits to state s_i after starting the Markov chain at s_i:

$$N_i = \sum_{t=1}^{\infty} I(X_t = s_i | X_0 = s_i).$$

- **Theorem 33.** *If s_i is a recurrent state, then*

$$Pr(N_i = \infty) = 1.$$

In other words, a Markov chain will return to a recurrent state infinitely many times with probability 1 if it ever reaches that state. On the other hand, if s_i is transient, we have that

$$P(N_i = k) = (1 - q)^k q$$

for all $k \in \mathbb{N} \cup \{0\}$. Here, $q = P(T_i = \infty | X_0 = s_i)$ is the probability that the Markov chain never returns to s_i.

- **Theorem 34 (Criterion for recurrence #2).** *Let s_i be an arbitrary state in an irreducible Markov chain with transition matrix P. Consider the system of equations*

$$\alpha_j = \sum_{k \neq i} \alpha_k P_{j,k}$$

for all $j \neq i$. The Markov chain is recurrent if and only if the only bounded solution of the above system of equations is $\alpha_j = 0$ for all $j \neq i$.

2.2 Stationary Distribution

- **Theorem 35.** *For an irreducible, recurrent, and aperiodic Markov chain with countably infinite states, it holds that*

$$\lim_{t \to \infty} P(X_t = s_i) = \frac{1}{E[T_i | X_0 = s_i]} = \frac{1}{\tau_{i,i}}$$

for any initial distribution of X_0. If $\tau_{i,i} = \infty$, then the limit on the right side is defined to be 0.

- The above theorem highlights a difference between the finite and countably infinite case.
  - In the finite case, by Lemma 21, we have that $\tau_{i,i}$ is finite for any irreducible and aperiodic Markov chain. (Note that we do not need recurrence because it is implied by irreducibility and aperiodicy in the finite case.) As a result, the asymptotic probability of s_i is non-zero.
  - On the other hand, when the state space is infinite, it might be the case that $\tau_{i,i}$ is infinite even though $\Pr(T_i < \infty | X_0 = s_i) = 1$. As a result, the asymptotic probability of s_i might be zero.
- **Definition 36.** *A recurrent state s_i is said to be* **positive recurrent** *if the mean return time $\tau_{i,i}$ is finite. Otherwise, it is said to be* **null recurrent**. *We say that a Markov chain is positive recurrent (or null recurrent) if all states are positive recurrent (or null recurrent, respectively).*
- **Proposition 37.** *All states in the same recurrent communication class are either all positive recurrent or all null recurrent.*
- **Proposition 38.** *A non-negative vector $\boldsymbol{\nu}$ that satisfies*

$$\nu_j = \sum_{i=1}^{\infty} \nu_i P_{i,j}$$

for all j is called an **invariant measure** *or a* **stationary measure**.

- Note that, if $\boldsymbol{\nu}$ is an invariant measure, then $c\boldsymbol{\nu}$ is also an invariant measure for any positive real number c.
- If an invariant measure $\boldsymbol{\nu}$ has an additional property that $\sum_i \nu_i = 1$, it becomes a stationary distribution.
- An invariant measure $\boldsymbol{\nu}$ can be normalized to a stationary distribution only if $\sum_i \nu_i$ is finite.
- **Theorem 39.** *For an irreducible, recurrent, aperiodic Markov chain with countably infinite state space, there is a unique (up to multiplication) invariant measure $\boldsymbol{\nu}$ given by*

$$\nu_j = E\left[\sum_{t=0}^{\infty} I(X_t = s_j \wedge t < T_{i,i}) \;\middle|\; X_0 = s_i \right]$$

for any arbitrary state $s_i \in S$. The invariant measure can be normalized to a stationary distribution if and only if $\tau_{i,i}$ is finite; that is, if the Markov chain is positive recurrent.

- Note the difference difference between the finite case (Theorem 23) and the countably infinite case (Theorem 39).
  - For the finite case, positive recurrence is a consequence of irreducibility and aperiodicity. So, any invariant measure is always normalizable, and we always have an stationary distribution.
  - However, positive recurrence is not automatically gauranteed, and it is required for the Markov chain to have a stationary distribution.
- To discuss convergence, we need the total variation distance, which remains pretty much the same for the countably infinite state space:

$$d_{\mathrm{TV}}(\mathbf{a}, \mathbf{b}) = \frac{1}{2} \sum_{i=1}^{\infty} |a_i - b_i|.$$

- **Theorem 40.** *Consider an irreducible, aperiodic, and positive recurrent Markov chain on a countably infinite state space. Starting from any distribution $\boldsymbol{\mu}^{(0)}$, we have that*

$$\boldsymbol{\mu}^{(t)} \xrightarrow{\mathrm{TV}} \boldsymbol{\pi}$$

where $\boldsymbol{\pi}$ is the unique stationary distribution

$$\boldsymbol{\pi} = \left(\frac{1}{\tau_{1,1}}, \frac{1}{\tau_{2,2}}, \dots \right).$$

- The following theorem is useful for computing functions on the states.

  Theorem 41. *Consider an irreducible, aperiodic, and positive recurrent Markov chain on a countably infinite state space with stationary distribution $\boldsymbol{\pi}$. Let $f : \mathcal{S} \to \mathbb{R}$ be a function on the state space such that*

$$\sum_{i=1}^{\infty} |f(s_i)| \pi_i < \infty.$$

  Then, for any initial distribution,

$$\Pr\left(\lim_{T \to \infty} \frac{1}{T} \sum_{t=0}^{T_1} f(X_t) = \sum_{i}^{\infty} f(s_i) \pi_i \right) = 1.$$

3 Continuous State Spaces

3.1 Basic Notions

- For continuous state spaces, the transition matrix is replaced by the transition "kernel."

  Definition 42. *Let S be a continuous state space, and $\mathcal{S}$ be a σ-algebra on S. A* **transition kernel** *K is a function from $S \times \mathcal{S}$ to $[0, 1]$ such that*

  1. *For all $x \in S$, the function $K(x, \cdot)$ is a probability measure on the measurable space $(S, \mathcal{S})$.*
  2. *For all $A \in \mathcal{S}$, the function $K(\cdot, A)$ is a measurable function on S.*

- Suppose that the space space is a subset of $\mathbb{R}^d$. Suppose also that, for all s, the measure $K(s, \cdot)$ is absolutely continuous with respect to the Lebesgue measure. The Random–Nikodym theorem tells us that there exists a non-negative function $k(x, y)$ such that

$$K(x, A) = \int_A k(x, y) \, \mathrm{d}y$$

  for all $A \in \mathcal{S}$. In this case, we refer to $k(x, y)$ as the **transition function**.

- The transition function must satisfy

$$\int_S k(x,y)\,\mathrm{d}y = 1$$

for all $x \in S$.

- **Definition 43.** *Let K be a transition kernel on the state space $(S,\mathcal{S})$. Let μ be a probability measure on $(S,\mathcal{S})$. A sequence of random variables X_0, X_1, X_2, ..., is a* **Markov chain transition kernel K and initial distribution μ** *if, for all $k = 0,1,2,\ldots$,*

$$P(X_{k+1} \in A | X_0, X_1, \ldots, X_k) = P(X_{k+1} \in A | X_k) = \int_A K(X_k, x)\,\mathrm{d}x$$

and the distribution of X_0 is μ.

- To save space, we abbreviate $\int_A K(X_k,x)\,\mathrm{d}x$ as $\int_A K(X_k,\mathrm{d}x)$.
- Examples.
  - **(Random walk)** Let (ξ_n) be an i.i.d. sequence of random variables. Let

$$X_n = \sum_{i=1}^{n} \xi_n.$$

    We have that $X_{n+1} = X_n + \xi_{n+1}$. The kernel is given by

$$K(x,A) = \mu_\xi(\{y - x : y \in A\})$$

    where $\mu_\xi(\cdot)$ is the probability measure of the values of the ξ_k's.
  - **(AR(1))** Again, let (ξ_n) be an i.i.d. sequence of random variables. Let θ be a real constant. Define

$$X_{n+1} = \theta X_n + \xi_{n+1}.$$

    The transition kernel is

$$K(x,A) = \mu_\xi(\{y - \theta x : y \in A\}).$$

- Let $K^n(\cdot,\cdot) : S \times \mathcal{S} \to [0,1]$ be defined by

$$K^n(x,A) = \Pr(X_n \in A | X_0 = x).$$

It is called the **n-step transition kernel**.

- **Proposition 44 (Kolmogorov–Chapman equation).** *Let m,n be positive integers and $A \in \mathcal{S}$. Then,*

$$K^{m+n}(x,A) = \int_S K^n(y,A) K^m(x,dy)$$

for $x \in S$ and $A \in \mathcal{S}$.

- Let $A_0, A_2, \ldots, A_n \in \mathcal{S}$. We have that

$$\Pr(X_0 \in A_0, X_1 \in A_1, \ldots, X_n \in A_n) = \left(\int_{A_0} \mathrm{d}\mu(x_0)\right)\left(\int_{A_1} K(x_0,\mathrm{d}x_1)\right)\cdots\left(\int_{A_n} K(x_{n-1},dx_n)\right).$$

- **Theorem 45 (Weak Markov property).** *Then, for any initial distribution and any positive integer k,*

$$E[h(X_{k+1}, X_{k+2}, \ldots) | X_0 = x_0, X_1 = x_1, \ldots, X_k = x_k] = E[h(X_1, X_2, \ldots) | X_0 = x_k].$$

for any function h such that the expectation exists.

3.2 Asymptotic Behavior

To discuss asymptotic behavior, we need to discuss terms like irreducibility, recurrence, aperiodicity, and convergence. These notions are quite different from their counterparts in the discrete cases.

3.2.1 Irreducibility

- **Definition 46.** *Let ψ be a non-zero σ-finite measure on $(S, \mathcal{S})$. A Markov chain is ψ-irreducible if, for every $A \in \mathcal{S}$ with $\psi(A) > 0$ and every $x \in S$, there exists a positive integer n such that $K^n(x, A) > 0$.*
- By the above definition, it doesn't matter if a set of measure zero is unreachable. The definition does not care about sets of measure zero.

3.2.2 Recurrence

- **Definition 47.** *Let $A \in \mathcal{S}$. The* **hitting time** *T_A is the time the chain first enters A.*

$$T_A = \inf\{n \geq 1 : X_n \in A\}.$$

- **Definition 48.** *Let $A \in \mathcal{S}$. Define N_A to be the number of times the chain visits A.*

$$N_A = \sum_{n=1}^{\infty} I(X_n \in A).$$

- **Definition 49.** *A Markov chain is* **recurrent** *if the following conditions hold.*
  - *(a) There exists a non-zero σ-finite measure ψ where the chain is ψ-irreducible.*
  - *(b) For all $A \in \mathcal{S}$ with $\psi(A) > 0$, we have that $E[N_A | X_0 = x] = \infty$ for all $x \in A$.*
- The above definition, however, has pathological cases. This is because we can have $E[N_A | X_0 = x] = \infty$ when $0 < \Pr(N_A = \infty | X_0 = x) < 1$.
- **Definition 50.** *A Markov chain is* **Harris recurrent** *if the following conditions hold.*
  - *(a) There exists a non-zero σ-finite measure ψ where the chain is ψ-irreducible.*
  - *(b) For all $A \in \mathcal{S}$ with $\psi(A) > 0$, we have that $\Pr(T_A < \infty | X_0 = x) = 1$ for all $x \in A$.*
- **Definition 51.** *A σ-finite measure π is* **invariant** *for a Markov chain with transition kernel K if*

$$\pi(A) = \int_S K(x, A) \, \mathrm{d}\pi(x)$$

  for all $A \in \mathcal{S}$. If there is an invariant measure which is a probability measure, then we say that the Markov chain is **positive recurrent**.
- **Theorem 52.** *Every recurrent Markov chain has an invariant σ-finite measure. It is unique up to a multiplicative constant.*
- **Proposition 53.** *If a Markov chain is positive recurrent, then it is recurrent.*
- A recurrent chain that is not positive recurrent is called **null recurrent**.

3.2.3 Aperiodicity

- **Definition 54.** *A Markov chain is said to be* **periodic** *if there exists a sequence of non-empty disjoint measurable sets A_1, A_2, ..., A_n with $n \geq 2$ such that*

  (a) $K(x, A_{j+1}) = 1$ for all $x \in A_j$ and $j = 1, 2, \ldots, n-1$, and

  (b) $K(x, A_1) = 1$ for all $x \in A_n$.

  If a Markov chain is not periodic, it is said to be **aperiodic**

- **Proposition 55.** *Consider a Markov chain with transition function $k(\cdot, \cdot)$. If $k(x, \cdot)$ is positive in a neighborhood of x for all $x \in S$, then the Markov chain is a periodic.*

  This is true because the chain can remain in that neighborhood for an arbitrary number of times before visiting other areas.

3.2.4 Total Variation Distance

- **Definition 56.** *The* **total variation distance** *between two probabiliy measures μ and ν on measurable space $(S, \mathcal{S})$ is given by:*

$$d_{\mathrm{TV}}(\mu, \nu) = \sup_{A \in \mathcal{S}} |\mu(A) - \nu(A)|.$$

- We note that this definition is equivalent to Definition 24.

3.2.5 Convergence Theorems

- **Theorem 57.** *Consider a Markov chain with transition kernel K. If the chain is ψ-irreducible, aperiodic, and has an invariant probability measure (i.e., a stationary distribution) π, then*

$$\lim_{n \to \infty} d_{\mathrm{TV}}(K^n(x, \cdot), \pi(\cdot)) = 0$$

  for π-almost every $x \in S$ regardless of the initial distribution of X_0. Moreover, if the chain is Harris recurrent, then the convergence holds for all $x \in S$.

3.3 Reversibility

- **Definition 58.** *Consider a Markov chain with transition function k. The chain satisfies* **detailed balance** *if there is a non-negative function $p : S \to \mathbb{R}$ such that*

$$p(y)k(y, x) = p(x)k(x, y)$$

  for all $x, y \in S$.

- **Proposition 59.** *Consider a Markov chain that satisfies detailed balance through a function p. If p is integrable with respect to some measure μ on $(S, \mathcal{S})$, then its integral $\pi(A) = \int_A p \, \mathrm{d}\mu$ is an invariant measure.*

References

[Häggström, 2002] Häggström, O. (2002). *Finite Markov Chains and Algorithmic Applications.*

[Kennedy, 2016] Kennedy, T. (2016). Markov chain background. `https://www.math.arizona.edu/~tgk/mc/book_chap7.pdf`. Accessed: 2022-01-15.

[Lee, 2012a] Lee, S. S. (2012a). Markov chains on continuous state space. https://www.webpages.uidaho.edu/~stevel/565/lectures/5d%20MCMC.pdf. Accessed: 2022-01-15.

[Lee, 2012b] Lee, S. S. (2012b). Markov chains on countable state space. https://www.webpages.uidaho.edu/~stevel/565/lectures/5c%20Markov%20chain.pdf. Accessed: 2022-01-15.

[Tolver, 2016] Tolver, A. (2016). An introduction to markov chains. http://web.math.ku.dk/noter/filer/stoknoter.pdf. Accessed: 2022-01-15.