1 Markov Chains

A **Markov Chain** is a stochastic process $\{X_n\}_{n=0}^{\infty}$ defined on a discrete state space S such that the probability of transitioning to the next state depends only on the present state and not on the sequence of events that preceded it. This property is known as the *Markov property*. For simplicity's sake, we assume S to be finite. Hence, the formal definition:

Definition 1.1 (Markov Chain)**.** A **Markov Chain** is a stochastic process $\{X_n\}_{n=0}^{\infty}$ on a discrete finite state space $S = \{1, \ldots, N\}$ satisfying the *Markov property*:

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_0 = x_0) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n),$$

for all $n \geq 0$ and all $x_0, \ldots, x_{n+1} \in S$.

The transition probabilities are described by a matrix P with entries

$$P_{ij} = \mathbb{P}(X_{n+1} = i \mid X_n = j), \quad \text{where } \sum_{i \in S} P_{ij} = 1 \text{ for all } j \in S.$$

Example 1.1 (Markov Transition Matrix)**.** Let the state space be $S = \{1, 2, 3\}$. A possible transition matrix P is:

$$P = \begin{bmatrix} 0 & 0.7 & 0 \\ 0.5 & 0.3 & 1 \\ 0.5 & 0 & 0 \end{bmatrix}$$

Note that each column sums to 1. Hence, P is a valid column-stochastic Markov matrix according to the given definition.

Remark 1.1. Some sources define $\boldsymbol{P'}$ s.t. $\boldsymbol{P'_{ij}} = P(X_{n+1} = j \mid X_n = i)$. The only difference is that $\boldsymbol{P'} = \boldsymbol{P}^T$.

Based on these simple definitions, we can deduce very useful properties. For example, we can calculate $P(X_{t+n} = i \mid X_t = j)$ algebraically very simple based on the following result:

Lemma 1.1 (n-Step Transition Probabilities)**.** *The probability of transitioning from state j to state i in n steps is given by the (i, j)-th entry of the matrix power P^n:*

$$\mathbb{P}(X_{t+n} = i \mid X_t = j) = (P^n)_{ij}.$$

Proof. We prove this by induction on n.

Base case: When $n = 1$, we have

$$\mathbb{P}(X_{t+1} = i \mid X_t = j) = P_{ij} = (P^1)_{ij}.$$

Inductive step: Assume the claim holds for $n = k$, i.e.,

$$\mathbb{P}(X_{t+k} = i \mid X_t = j) = (P^k)_{ij}.$$

For $n = k + 1$, using the law of total probability and the Markov property:

$$\mathbb{P}(X_{t+k+1} = i \mid X_t = j) = \sum_{m \in S} \mathbb{P}(X_{t+k+1} = i \mid X_{t+k} = m) \cdot \mathbb{P}(X_{t+k} = m \mid X_t = j).$$

$$= \sum_{m \in S} P_{im} \cdot (P^k)_{mj} = (P \cdot P^k)_{ij} = (P^{k+1})_{ij}.$$

Hence, by induction, the result holds for all $n \geq 1$. □

Lemma 1.2 (n-Step Probability Distribution)**.** *If we have a probability distribution vector $\boldsymbol{p}_t$ at time t, meaning $(\boldsymbol{p}_t)_i = P(X_t = i)$, we get $\boldsymbol{p}_{t+n}$ by $\boldsymbol{P}^n \boldsymbol{p}_t$.*

Proof.

$$\begin{aligned}
(\boldsymbol{p}_{t+n})_i &= P(X_{t+n} = i) \\
&= \sum_{j \in S} P(X_{t+n} = i \mid X_t = j) P(X_t = j) \\
&= \sum_{j \in S} (\boldsymbol{P}^n)_{ij} P(X_t = j) \\
&= (\boldsymbol{P}^n \boldsymbol{p}_t)_i
\end{aligned}$$

□

1.1 Properties

At first glance, this model seems to be very simple. However, we can characterize Markov chains by their individual properties. But what could these properties be? Well, we first might want to visualize Markov chains. To this end, we employ a graph $G = (V, E)$ with $V = S$ and $E = \{(u, v) \mid \boldsymbol{P}_{vu} \neq 0\}$.

Hence, the Markov chain from example 1.1 becomes:

Note that we do not care about the magnitude of the transition probabilities, it only matters whether it is possible to transition from one state to another.

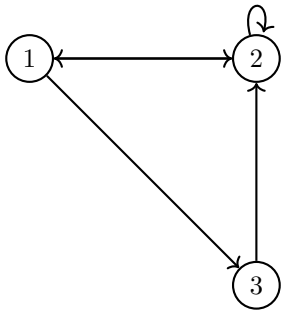

Figure 1: Graph representation of the Markov chain defined in example 1.1.

1.1.1 Irreducibility

Already we can see that we can reach every every state v from every other state u, i.e. there exists a path of length n starting at u and ending at $v \iff (\boldsymbol{P}^n)_{vu} > 0$. Such a Markov chain is called *irreducible*. The importance of this is that the Markov chain cannot trap itself in a subclass of states, like for example in figure 2.

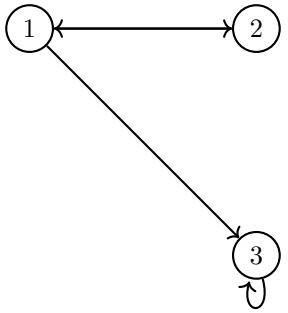

Figure 2: Graph of a reducible Markov chain. Note that once the chain transitions from state 1 to state 3, it will stay at state 3 indefinitely.

Before defining this property formally, we may first introduce a very related concept of *communication classes*:

Definition 1.2. We say that $i \in S$ *leads to* state $j \in S$ iff there exists $n \in \mathbb{N}_{>0}$ s.t. $(\boldsymbol{P}^n)_{ji} > 0$. We use the notation $i \rightsquigarrow j$. Also, for all $i \in S$: $i \rightsquigarrow i$.

Definition 1.3 (Communication between States)**.** States $i, j \in S$ *communicate* iff $i \rightsquigarrow j$ and $j \rightsquigarrow i$. We use the notation $i \leftrightsquigarrow j$.

Theorem 1.1. *$i \leftrightsquigarrow j$ is an equivalence relation.*

Proof. Both reflexivity and symmetry follow directly from the definitions. For transitivity we have assuming $i \neq j$ and $j \neq k$ (the other cases are trivial):

$$
\begin{aligned}
& i \leftrightsquigarrow j \text{ and } j \leftrightsquigarrow k \\
& \implies i \rightsquigarrow j \text{ and } j \rightsquigarrow i \\
& \implies \exists_{m,n \in \mathbb{N}_{>0}} (\boldsymbol{P}^m)_{ji} > 0 \text{ and } (\boldsymbol{P}^n)_{kj} > 0 \\
& \implies \exists_{m,n \in \mathbb{N}_{>0}} P(X_m = j \mid X_0 = i) > 0 \text{ and } P(X_{m+n} = k \mid X_m = j) > 0 \\
& \implies P(X_{m+n} = k \mid X_0 = i) > 0 \\
& \implies (\boldsymbol{P}^{m+n})_{ki} > 0 \\
& \implies i \leftrightsquigarrow k
\end{aligned}
$$

□

Based on this result, the following definition suggests itself:

Definition 1.4 (Communication Class)**.** The *communication class* of state $i \in S$ is the set $\{j \in S : i \leftrightsquigarrow j\}$. This set consists of all states j that communicate with i.

Remark 1.2. Since communication of states is an equivalence relation, the state space S can be decomposed into a disjoint union of communication classes (also called a *partition*). Any two communication classes either coincide completely or are disjoint sets.

Example 1.2. The partition of figure 1 is $\{\{1, 2, 3\}\}$ and of figure 2 we have $\{\{1, 2\}, \{3\}\}$.

Finally, we can state the concept of *irreducibility* formally:

Definition 1.5 (Irreducibility)**.** A Markov chain is *irreducible* iff every two states communicate. Hence, an irreducible Markov chain consists of exactly one communication class.

We will mostly focus on irreducible Markov chains, but for the completeness' sake we also define the following concepts:

Definition 1.6 (Open and Closed Communication Class)**.** A communication class C is *open* iff there exists a state $i \in C$ and a state $k \notin C$ s.t. $i \rightsquigarrow k$. Otherwise, C is called *closed*.

Remark 1.3. An irreducible Markov chain has exactly one closed communication class.

If a Markov chain once arrived in a closed communication class, it will stay in this class forever. This is exactly what happens in figure 2.

Theorem 1.2 (Existence of Closed Communication Class)**.** *There is always at least one closed communication class.*

Proof. Assume all communication classes $C_1, \ldots, C_k$ are open. Hence, we can traverse these classes. But at some point we must complete a cycle, but that is a contradiction, as this would imply that those communication classes forming the cycle are really just one big communication class, which is not the case. □

1.1.2 Aperiodicity

We will now analyze the second important property of Markov chains. The reader might ask, *important for what?* Well, the answer will make more sense in the big picture later, but the short answer is that these two properties will guarantee for a convergence to a unique stationary probability distribution.

To motivate the following discussion, say we had the following Markov chain:

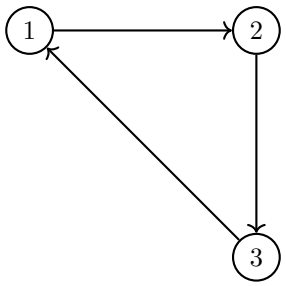

Figure 3: Graph of a periodic Markov chain. Note that once once the starting position is determined, then we also know the state after t steps.

We notice that the behavior of this chain is periodic with a period of length 3. Of course, this is very informal speaking, but we will now define this idea precisely.

Definition 1.7 (Period and Aperiodicity)**.** The *period* of state $i \in S$ is defined as

$$\gcd\{n \in \mathbb{N}_{>0} : (\boldsymbol{P}^n)_{ii} > 0\} \quad .$$

Here, gcd stands for the greatest common divisor. A state $i \in S$ is called aperiodic iff its period is equal to 1. Otherwise, the state i is called periodic.

Remark 1.4. This definition is not well defined in all cases, as it could happen that $\{n \in \mathbb{N}_{>0} : (\boldsymbol{P}^n)_{ii} > 0\} = \emptyset$. However, we mostly care about Markov chains being aperiodic in closed communication classes, especially in irreducible Markov chains. And for closed communication classes, this set can never be empty. In fact, for the set to be empty, we must have a communication class consisting of only one state $i \in S$ with $\boldsymbol{P}_{ii} = 0$. This communication class is obviously open.

Lemma 1.3 (Periodicity and Aperiodicity are Class Properties)**.** *If state $i \in S$ is aperiodic and $i \leftrightsquigarrow j$, then j is also aperiodic.*

Proof. Since i is aperiodic, we can find an $n \in \mathbb{N}$ s.t. both $(\boldsymbol{P}^n)_{ii} > 0$ and $(\boldsymbol{P}^{n+1})_{ii} > 0$ due to results from number theory. Since $j \rightsquigarrow i$, we can go from j to i in t_{ji} steps, and from i to j in t_{ij} steps since $i \rightsquigarrow j$. Thus:

$$\{t_{ji} + n + t_{ij}, t_{ji} + n + 1 + t_{ij}\} \subseteq \gcd\{n \in \mathbb{N}_{>0} : (\boldsymbol{P}^n)_{jj} > 0\} \quad .$$

Since $\gcd\{t_{ji} + n + t_{ij}, t_{ji} + n + 1 + t_{ij}\} = 1$, we conclude that $\gcd\{n \in \mathbb{N}_{>0} : (\boldsymbol{P}^n)_{jj} > 0\} = 1$, and hence j is aperiodic. □

This result leads to the following definitions:

Definition 1.8 (Aperiodic Markov Chain)**.** An irreducible Markov chain is called aperiodic iff some (and hence, all) states in this chain are aperiodic.

With these basics covered, we can now focus on establishing important results we will need later.

1.2 Irreducible Aperiodic Markov Chains

We are interested in *stationary* probability distributions $\boldsymbol{\mu}$ satisfying $\boldsymbol{P}\boldsymbol{\mu} = \boldsymbol{\mu}$.

Does such a stationary probability distribution always exist? Well, for a finite state space maybe. More interestingly, we may also ask whether a Markov chain will converge towards $\boldsymbol{\mu}$ regardless of the initial probability distribution vector, i.e. for all $\boldsymbol{p} : \lim_{n\to\infty} \boldsymbol{P}^n\boldsymbol{p} = \boldsymbol{\mu}$.

In general, the answer to this is no. Consider the Markov chain in figure 4 again. Clearly, if we start at certain state, say state 1, then we will always hop around the states and never converge to $\boldsymbol{\mu}$. So our intuition might be that we need the Markov chain to be aperiodic in order for it to converge.

Furthermore, we also might ask whether the stationary probability distribution is unique. Well, in general the answer is no as well. To see this, consider this Markov chain:

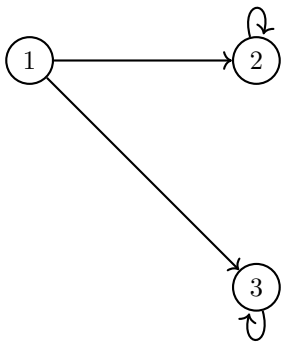

Figure 4: Graph of a reducible Markov chain with two closed communication classes. Note that we might end up stuck at either state 2 or state 3.

Clearly, we have two stationary probability distributions with all their weight in either state 2 or state 3. Once again, our intuition tells us we might require an irreducible Markov chain.

Now it's time to specify our intuitions precisely. The following result is for irreducible aperiodic Markov chains is very significant.

Theorem 1.3 (Positive n-Step Transition Matrix for Irreducible Aperiodic Markov Chains)**.** *For every irreducible aperiodic Markov chain specified by* $\boldsymbol{P}$*, there exists an* $n \in \mathbb{N}_{>0}$ *s.t.* $\boldsymbol{P}^n > \mathbf{0}$*, where the comparison is element-wise.*

We first prove the following auxiliary lemma:

Lemma 1.4. *Let* $i \in S$ *be an aperiodic state. Then there exists an* $L \in \mathbb{N}$ *s.t. for all* $n > L : (\boldsymbol{P}^n)_{ii} > 0$*.*

Proof. Since state i is aperiodic, we can find $n_1, \ldots, n_r \in \mathbb{N}$ s.t. $(\boldsymbol{P}^{n_1})_{ii} > 0, \ldots, (\boldsymbol{P}^{n_r})_{ii} > 0$ and $\gcd\{n_1, \ldots, n_r\} = 1$. From number theory, we know that for $L \coloneqq \prod_{k=1}^{r} n_k$ we can write every natural number $n > L$ in the form $n = l_1 n_1 + \cdots + l_r n_r$ for suitable $l_1, \ldots, l_r \in \mathbb{N}$. Hence:

$$(\boldsymbol{P}^{l_1 n_1 + \cdots + l_r n_r})_{ii} \geq ((\boldsymbol{P}^{n_1})^{l_1})_{ii} \cdot \cdots \cdot ((\boldsymbol{P}^{n_r})^{l_r})_{ii} > 0 \quad .$$

□

Remark 1.5. The converse of lemma 1.4 is also true for obvious reasons.

Proof of Theorem 1.3. Let L' be defined as the maximum of all L defined like in Lemma 1.4 when looping over all $i \in S$. Then for every $n > L'$, we have that $\boldsymbol{P}^n$ has positive entries along its diagonal. It follows that if $(\boldsymbol{P}^{t_{ji}})_{ij} > 0$ for some $t_{ji} \in \mathbb{N}_{>0}$, then we have $(\boldsymbol{P}^{n+t_{ji}+\tau})_{ij} > 0$ as well for every $\tau \in \mathbb{N}$. Furthermore, for every $i, j \in S$ we have $(\boldsymbol{P}^{t_{ji}})_{ij} > 0$ at some point $t_{ji} \in \mathbb{N}_{>0}$ due to aperiodicity. Hence, at some point all the zeros must have vanished. □

Corollary 1.1. *Once $\boldsymbol{P}^n > \boldsymbol{0}$, then for all $\tau \in \mathbb{N} : \boldsymbol{P}^{n+\tau} > \boldsymbol{0}$.*

The reader might question the importance of Theorem 1.3. Clearly, we can see the interplay of the two properties of the Markov chain being irreducible and aperiodic in the proofs. But how does it help us finding a stationary probability distribution? Well, to answer this, we need another fundamental result, which we will cover next.

1.3 Perron-Frobenius Theorem

To come straight to the point, the Perron-Frobenius Theorem reads as follows:

Theorem 1.4 (Perron-Frobenius)**.** *Let $\boldsymbol{A}$ be a non-negative, irreducible matrix (i.e., all entries of $\boldsymbol{A}$ are non-negative, and there exists some n such that all entries of $\boldsymbol{A}^n$ are positive). Then the following hold:*

1. *The matrix $\boldsymbol{A}$ has a unique largest non-negative eigenvalue $\lambda_{\max}$, and this eigenvalue is simple (it has algebraic multiplicity 1).*
2. *The eigenvalue $\lambda_{\max}$ is real and positive.*
3. *There is a corresponding positive eigenvector $\boldsymbol{v}$ (i.e., all entries of $\boldsymbol{v}$ are positive) associated with $\lambda_{\max}$.*
4. *Any other eigenvalue λ of $\boldsymbol{A}$ satisfies $|\lambda| < \lambda_{\max}$.*

This is a mouthful, and we will not prove this theorem in this general form, as it is not trivial to do so. Instead, we focus on the case of $\boldsymbol{A}$ being a Markov chain transition matrix, meaning all columns sum to 1. To this end, we will write $\boldsymbol{P}$ again instead of $\boldsymbol{A}$. Additionally, we assume $\boldsymbol{P}$ to be strictly positive for now instead of irreducible. Furthermore, we will not provide a rigorous proof, but we will lay the foundation for a solid intuitive understanding.

Proof of Perron-Frobenius for Positive Markov Chain Transition Matrices. Consider the mapping $T : \Delta \to \Delta$ from the unit simplex Δ onto itself defined by $T(v) \coloneqq \boldsymbol{P}v$. We want to show that T is a contraction mapping with respect to the L_1 norm. We do this step last in order to understand the line of argument better.

Thus, assume that T is a contraction mapping with respect to the L_1 norm. By the Banach Fixed-Point Theorem we know that this mapping has a *unique* fixed point v^*. We see that v^* is an eigenvector of $\boldsymbol{P}$ with eigenvalue $\lambda = 1$. Clearly, v^* is the only eigenvector with non-negative (actually positive!) entries, as if there were another one say v', then $\frac{v'}{\|v'\|_1}$ must be one as well, but this point lies on Δ, hence it must have an eigenvalue of 1 and must be a fixed point of T, a contradiction to the uniqueness of v^*.

So every other eigenvector w with eigenvalue μ must have a coordinate-entry which is negative or truly complex. Write $|w|$ for the vector with coordinates $|w_j|$. The computation

$$|\mu||w|_i = |\mu w_i| = |\sum_j \boldsymbol{P}_{ij} w_j| \leq \sum_j |\boldsymbol{P}_{ij}||w_j| = \sum_j \boldsymbol{P}_{ij}|w_j| = (\boldsymbol{P}|w|)_i$$

shows that $|\mu|\|w\|_1 \leq \|\boldsymbol{P}|w|\|_1 = \|w\|_1$ and hence $|\mu| \leq 1$. Now, the final trick is that we can assume the "$\leq$" in the equation above to actually be "$<$".

To see this, note that we only have equality iff all entries of w are on a line in the Gauss plane, i.e. we can write $w = c|w|$ for some $c \in \mathbb{C}, |c| = 1$. This would mean that $\boldsymbol{P}w = \boldsymbol{P}c|w| = c\boldsymbol{P}|w| \stackrel{!}{=} \mu w$. Hence, $\boldsymbol{P}|w| = \frac{\mu}{c}w = |\frac{\mu}{c}||w|$. Thus, we must have $|w| = v^*$ and $|\mu| = |c| = 1$ as already discussed. Hence, $w = cv^*$ (and $\mu = 1$), which is just a different representation of the eigenvector v^* already found. So for every eigenvector w other than v^* with eigenvalue μ we must have that $|\mu| < 1$. $\square$