Criticality in Formal Languages

Jonas Peters

April 15, 2025

Contents

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 $S = \{1, \dots, N\}$ satisfying the *Markov property*:

$$P(X_{n+1} = x_{n+1} \mid X_n = x_n, \dots, X_0 = x_0) = P(X_{n+1} = x_{n+1} \mid X_n = x_n) \quad ,$$

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

The transition probabilities are described by a matrix $\boldsymbol{P}$ with entries

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

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

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

Note that each column sums to 1.

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 $\boldsymbol{P}^n$:*

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

Proof. We prove this by induction on n.

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

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

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

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

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

$$\begin{aligned}
P(X_{t+k+1} = i \mid X_t = j) &= \sum_{m \in S} P(X_{t+k+1} = i \mid X_{t+k} = m) \cdot P(X_{t+k} = m \mid X_t = j) \\
&= \sum_{m \in S} \boldsymbol{P}_{im} \cdot (\boldsymbol{P}^k)_{mj} \\
&= (\boldsymbol{P} \cdot \boldsymbol{P}^k)_{ij} = (\boldsymbol{P}^{k+1})_{ij} \quad .
\end{aligned}$$

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

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.

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

$\square$

1.1 Properties

Markov chains are a very simple model. Thus, we can analyze them thoroughly and investigate their 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 can be visualized as in figure 1.

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.

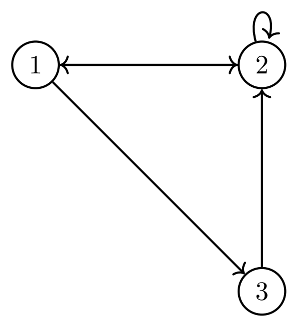

Figure 1: Graph representation of the Markov chain defined in example 1.1.

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.

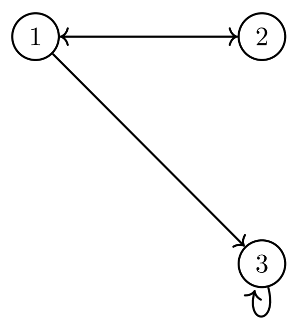

Figure 2: Graph of a reducible Markov chain. Note that once the chain transitions from state 1 to state 3, it will stay at state 3 indefinitely.

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):

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

$k \rightsquigarrow i$ can be show in similar fashion. ☐

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 $C_1, \dots, C_k$ 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. $\square$

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 bigger 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:

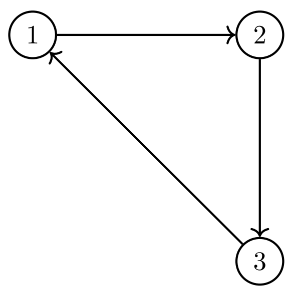

Figure 3: Graph of a periodic Markov chain. Note that once once the starting position is determined, then we also know the state after t steps.

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:

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

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. $\square$

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 3 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:

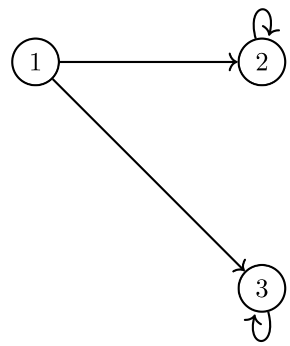

Figure 4: Graph of a reducible Markov chain with two closed communication classes. Note that we might end up stuck at either state 2 or state 3.

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 regarding 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* $m \in \mathbb{N}_{>0}$ *s.t.* $\boldsymbol{P}^m > \mathbf{0}$*, 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 $n_1, \dots, n_r \in \mathbb{N}$ s.t. $(\boldsymbol{P}^{n_1})_{ii} > 0, \dots, (\boldsymbol{P}^{n_r})_{ii} > 0$ and $\gcd\{n_1, \dots, n_r\} = 1$. 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 $n = l_1 n_1 + \dots + l_r n_r$ for suitable $l_1, \dots, l_r \in \mathbb{N}$. Hence:

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

$\square$

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 $t_{ji} \in \mathbb{N}_{>0}$ since $\boldsymbol{P}$ is irreducible. Hence, at some point all the zeros must have vanished. $\square$

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

Corollary 1.2. *The Converse of Theorem 1.3 is also true, that is if $\boldsymbol{P}^m > \mathbf{0}$ for some $m \in \mathbb{N}$, then P is irreducible and aperiodic. Irreducibility directly follows from $\boldsymbol{P}^m > \mathbf{0}$, and since $\boldsymbol{P}^{m+1} > \mathbf{0}$ as well, $\boldsymbol{P}$ must also be aperiodic.*

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 matrix (i.e., all entries of $\boldsymbol{A}$ are non-negative) with the propetry that there exists some $m \in \mathbb{N}$ such that all entries of $\boldsymbol{A}^m$ are strictly 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. Furthermore, we will not provide a rigorous proof, but we will lay the foundation for an 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 $T(\boldsymbol{v}) := \boldsymbol{P}\boldsymbol{v}$. 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 $\boldsymbol{v}^*$. We see that $\boldsymbol{v}^*$ is an eigenvector of $\boldsymbol{P}$ with eigenvalue $\lambda = 1$. Clearly, $\boldsymbol{v}^*$ is positive and is the only eigenvector with non-negative entries, as if there were another one say $\boldsymbol{v'}$, then $\frac{\boldsymbol{v'}}{\|\boldsymbol{v'}\|_1}$ 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 $\boldsymbol{v}^*$.

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

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

shows that $|\mu|\|\boldsymbol{w}\|_1 \leq \|\boldsymbol{P}|\boldsymbol{w}|\|_1 = \|\boldsymbol{w}\|_1$ 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 $\boldsymbol{w}$ are on a line in the complex plane, i.e. we can write $\boldsymbol{w} = c|\boldsymbol{w}|$ for some $c \in \mathbb{C}, |c| = 1$. This would mean that $\boldsymbol{P}\boldsymbol{w} = \boldsymbol{P}c|\boldsymbol{w}| = c\boldsymbol{P}|\boldsymbol{w}| \overset{!}{=} \mu\boldsymbol{w}$. Hence, $\boldsymbol{P}|\boldsymbol{w}| = \frac{\mu}{c}\boldsymbol{w} = |\frac{\mu}{c}||\boldsymbol{w}|$. Thus, we must have $|\boldsymbol{w}| = \boldsymbol{v}^*$ and $|\mu| = |c| = 1$ as already discussed. Hence, $\boldsymbol{w} = c\boldsymbol{v}^*$ (and $\mu = 1$), which is just a different representation of the eigenvector $\boldsymbol{v}^*$ already found. So for every eigenvector $\boldsymbol{w}$ other than $\boldsymbol{v}^*$ with eigenvalue μ we must have that $|\mu| < 1$.

Now, the only thing left to do is to show that T is in fact a contraction mapping. To this end, we must find a $0 \leq k < 1$ s.t. for all $\boldsymbol{x}, \boldsymbol{y} \in \Delta$ we have

$$\|T(\boldsymbol{x}) - T(\boldsymbol{y})\|_1 = \|\boldsymbol{P}(\boldsymbol{x} - \boldsymbol{y})\|_1 \leq k\|\boldsymbol{x} - \boldsymbol{y}\|_1 \quad .$$

The idea is that T maps from Δ into a real subspace $\Delta_{\boldsymbol{P}} \subsetneq \Delta$ defined by the simplex spanned by the columns of $\boldsymbol{P}$, which we call the vertices of the simplex $\Delta_{\boldsymbol{P}}$. $\Delta_{\boldsymbol{P}}$ does not contain any of the border points of Δ. Applying T, we see that T^2 maps into an even smaller sub-simplex $\Delta_{\boldsymbol{P}^2} \subsetneq \Delta_{\boldsymbol{P}}$, and so on and so forth.

Formally, we define k as $k := \frac{\|\Delta_{\boldsymbol{P}}\|_1}{2}$, where $\|\Delta_{\boldsymbol{P}}\|_1$ denotes the maximum L_1 distance between any points in $\Delta_{\boldsymbol{P}}$. We can always measure this distance at two of the vertices of the simplex, as both the L_1 norm and the simplex are convex, so the maximum will be reached at vertices. But since all coordinates of all vertices are strictly positive, we have $\|\Delta_{\boldsymbol{P}}\|_1 < 2$, so k is in fact valid.

Now, for the sake of contradiction, assume $\|\boldsymbol{P}(\boldsymbol{x}-\boldsymbol{y})\|_1 > k\|\boldsymbol{x}-\boldsymbol{y}\|_1$ for some $\boldsymbol{x}, \boldsymbol{y} \in \Delta$. Note that $\boldsymbol{x}-\boldsymbol{y}$ is a vector with entries which sum to 0, which defines a direction tangent to the unit simplex Δ. We can always find two points $\boldsymbol{x}', \boldsymbol{y}' \in \Delta$ with the same direction and $\|\boldsymbol{x}'-\boldsymbol{y}'\|_1 = 2$ (one point will be a vertex, the other will lay on the opposite side). Hence:

$$\begin{aligned}
\|\boldsymbol{P}(\boldsymbol{x}'-\boldsymbol{y}')\|_1 &= \frac{\|(\boldsymbol{x}'-\boldsymbol{y}')\|_1}{\|(\boldsymbol{x}-\boldsymbol{y})\|_1}\|\boldsymbol{P}(\boldsymbol{x}-\boldsymbol{y})\|_1 \\
&> k\frac{\|(\boldsymbol{x}'-\boldsymbol{y}')\|_1}{\|(\boldsymbol{x}-\boldsymbol{y})\|_1}\|\boldsymbol{x}-\boldsymbol{y}\|_1 \\
&= k\|(\boldsymbol{x}'-\boldsymbol{y}')\|_1 = 2k = \|\Delta_{\boldsymbol{P}}\|_1 \quad ,
\end{aligned}$$

a contradiction. □

Corollary 1.3 (Extension to More General Markov Chains)**.** *Based on the proof, we see that we only needed $\boldsymbol{P}$ to be positive in order to, first, guarantee $\|\Delta_{\boldsymbol{P}}\|_1 < 2$. Of course, many other $\boldsymbol{P}$ imply this, as we just need every pair of columns of $\boldsymbol{P}$ to have at least one pair of positive entries at common indices. Thus, the Perron-Frobenius Theorem also applies to Markov chains $\boldsymbol{P}$ where one row of $\boldsymbol{P}$ is strictly positive for example. BUT: The associated eigenvector must not be positive anymore, since that was the second application of $\boldsymbol{P} > \boldsymbol{0}$.*

Corollary 1.4. *Based on the proof, we see that for every $\boldsymbol{v} \in \Delta : \lim_{n\to\infty} \boldsymbol{P}^n\boldsymbol{v} = \boldsymbol{v}^*$. Now, set $\boldsymbol{v} := \boldsymbol{e_i}$. It follows that $\lim_{n\to\infty} \boldsymbol{P}^n = \boldsymbol{P}_{\boldsymbol{v}^*}$, where $\boldsymbol{P}_{\boldsymbol{v}^*}$ is the matrix whose columns all consist of the unique fixed point $\boldsymbol{v}^*$. This convergence is independent of the norm.*

Lemma 1.5. *If the Perron-Frobenius Theorem applies to both $\boldsymbol{A}^m$ and $\boldsymbol{A}^{m+1}$ for some $m \in \mathbb{N}_{>0}$, then it will also apply to $\boldsymbol{A}$.*

Proof. Let $\lambda_1, \ldots, \lambda_n$ by the eigenvalues of $\boldsymbol{A}$ allowing for multiplicity and ordering them s.t. $|\lambda_1| \geq \cdots \geq |\lambda_n|$. We apply the Perron-Frobenius Theorem on $\boldsymbol{A}^m$ to get the eigenvalues $\lambda_1^{(m)}, \ldots, \lambda_n^{(m)}$ s.t. $\lambda_1^{(m)} > |\lambda_2^{(m)}| \geq \cdots \geq |\lambda_n^{(m)}|$. Now, assume that we have ordered $\lambda_1, \ldots, \lambda_n$ perfectly s.t. $\lambda_i^m = \lambda_i^{(m)}$, as such an order always exists. From this, we immediately see that $|\lambda_1|$ is strictly bigger than all other eigenvalues of $\boldsymbol{A}$. It also must be real and positive, as both $\lambda_1^{(m)}$ and $\lambda_1^{(m+1)}$ are real and positive, and thus $\lambda_1 = \frac{\lambda_1^{(m+1)}}{\lambda_1^{(m)}}$ is real and positive. So statements (1), (2), and (4) of theorem 1.4 follow.

Let $\boldsymbol{v}^{(m)}$ be the (unique up to scaling) eigenvector of $\boldsymbol{A}^m$ with eigenvalue $\lambda_1^{(m)}$. By the Perron-Frobenius Theorem we know that $\boldsymbol{v}^{(m)}$ is positive. Also, we know that $\boldsymbol{A}$ has an eigenvector $\boldsymbol{v}^*$ with eigenvalue λ_1, since λ_1 is unique. This eigenvector will not change for $\boldsymbol{A}^r$ for every $r \in \mathbb{N}$, and the only matching eigenvector for $\boldsymbol{v}^*$ when $r = m$ is $\boldsymbol{v}^{(m)}$ based on the eigenvalues. And hence $\boldsymbol{v}^* := \boldsymbol{v}^{(m)}$ itself is positive, which was the last claim (3). □

Corollary 1.5. *The Perron-Frobenius Theorem holds for irreducible aperiodic Markov chain transition matrices* $\boldsymbol{P}$*, and* $\lambda_{max} = 1$*. The associated eigenvector is the stationary probability distribution* $\boldsymbol{\mu}$*.*

Remark 1.6. $T : \Delta \to \Delta, T(\boldsymbol{v}) \coloneqq \boldsymbol{P}\boldsymbol{v}$ is not a contraction mapping in general for irreducible aperiodic Markov chain tranisition matrices $\boldsymbol{P}$. But we still have $\Delta_{\boldsymbol{P}^{r+1}} \subsetneq \Delta_{\boldsymbol{P}^r}$, so intuitively $T^r(\boldsymbol{v}) = \boldsymbol{P}^r\boldsymbol{v}$ itself will converge to $\boldsymbol{\mu}$. To formally prove this, let $m \in \mathbb{N}$ be s.t. $\boldsymbol{P}^m > \boldsymbol{0}$. Then also $\boldsymbol{P}^{m+1} > \boldsymbol{0}$. $((\boldsymbol{P}^m)^r)_{r\in\mathbb{N}}$ and $((\boldsymbol{P}^{m+1})^r)_{r\in\mathbb{N}}$ have the common subsequence $((\boldsymbol{P}^{m^2+m})^r)_{r\in\mathbb{N}}$ and hence must converge to the same $\boldsymbol{P}_{\boldsymbol{\mu}}$. From this, it follows that $(\boldsymbol{P}^r)_{r\in\mathbb{N}}$ itself converges to $\boldsymbol{P}_{\boldsymbol{\mu}}$.

Remark 1.7. Again, let $m \in \mathbb{N}$ be s.t. $\boldsymbol{P}^m > \boldsymbol{0}$. Based on the proof, we see that

$$\|\boldsymbol{P}^{mr}\boldsymbol{v} - \boldsymbol{\mu}\|_1 \leq \frac{\|\Delta_{\boldsymbol{P}^m}\|_1^r}{2^r}\|(\boldsymbol{v} - \boldsymbol{\mu})\|_1 \in \mathcal{O}\left(\left[\frac{\|\Delta_{\boldsymbol{P}^m}\|_1}{2}\right]^r\right) \quad .$$

In other words, we have *exponential decay* in the distance between $\boldsymbol{P}^{mr}\boldsymbol{v}$ and $\boldsymbol{\mu}$ with respect to r. Hence, we also have

$$\|\boldsymbol{P}^r\boldsymbol{v} - \boldsymbol{\mu}\|_1 \in \mathcal{O}\left(\left[\frac{\|\Delta_{\boldsymbol{P}^m}\|_1}{2}\right]^{\dfrac{r}{m}}\right) = \mathcal{O}\left(\left[\left(\frac{\|\Delta_{\boldsymbol{P}^m}\|_1}{2}\right)^{\dfrac{1}{m}}\right]^r\right) \quad .$$

Hypothesis 1.1. Maybe

$$\lim_{m\to\infty} \left[\|\Delta_{\boldsymbol{P}^m}\|_1\right]^{\frac{1}{m}}$$

evaluates to something interesting like $|\lambda_2|$ of $\boldsymbol{P}$, or equivalently $\lim_{m\to\infty}$: $|\lambda_2^{(m)}|^{\frac{1}{m}} = \|\Delta_{\boldsymbol{P}^m}\|_1^{\frac{1}{m}}$ where $\lambda_2^{(m)}$ is the second largest eigenvalue of $\boldsymbol{P}^m$. Maybe in the proof there is another (complex) eigenvector hidden. Maybe we also use another norm, as we have convergence independent of the norm, but the contraction mapping is not as trivial.

2 Mutual Information in Markov Chains

2.1 Exponential Decay in Irreducible Aperiodic Markov Chains

If we have a Markov chain defined by the matrix $\boldsymbol{M}$ (we adopt the notation in the paper and write $\boldsymbol{M}$ instead of $\boldsymbol{P}$), which is *irreducible* and *aperiodic*, and has a finite state space $S = \{1, \ldots, n\}$, then we have that

$$\lim_{i \to \infty} \boldsymbol{M}^i = \boldsymbol{M}_{\boldsymbol{\mu}} \quad ,$$

where $\boldsymbol{M}_{\boldsymbol{\mu}}$ is the matrix whose columns all consist of the unique stationary probability distribution $\boldsymbol{\mu}$.

Now, let us consider two random variables X and Y, which will denote the state of the Markov chain at times t_0 and $t_0 + \tau$ respectively. We assume that we measure these variables very late in the process, where we already have that $\boldsymbol{M}^{t_0} \approx \boldsymbol{M}_{\boldsymbol{\mu}}$. We will use this "equality" later.

Our goal now is to quantify the mutual information of X and Y, that is, the discrepancy between the joint probability distribution $P(X,Y)$ and the one defined by the product of the two marginalized distributions, that is $P'(X,Y) \coloneqq P(X) \cdot P(Y)$. We use the Kullback-Leibler divergence, so our target expression becomes

$$D(P(X,Y) \,\|\, P'(X,Y)) \quad .$$

Note that of course this divergence $I(X,Y) \coloneqq D(P(X,Y) \,\|\, P'(X,Y))$ depends on the properties of M, as well as on τ. Because $\boldsymbol{M}$ is irreducible and aperiodic, it follows that $|\lambda_2| < 1$. The claim is that

$$I(X,Y) \in \mathcal{O}(|\lambda_2|^\tau) \quad .$$

There is a lot of math involved, so let us first get an intuition for what is going on. When considering Markov chains, we consider a set of states, say $S = \{A, B, C\}$, and for each time $t \in \mathbb{N}$ we assign a probability to the random variable $X_t \in S$. So let us consider the following Markov chain in figure 5.

If $\tau = 1$, i.e. we consider the mutual information of to consecutive states, we get a large value of $I(X,Y)$, as if X_{t_0} is either A or C, then X_{t_0+1} is uniquely determined, so we have a strong dependency between the two random variables. If, however, we have $\tau = 5$, then we can reach every state independent of the starting position. To see this, note that we can reach every state from A in four steps:

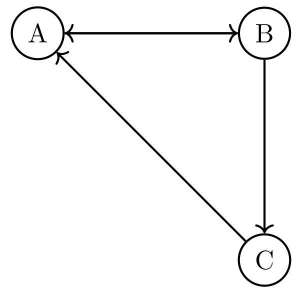

Figure 5: A simple irreducible aperiodic Markov chain. Note that if $X_{t_0} = C$, then we know that $X_{t_0+1} = A$.

- $A \to B \to C \to A \to B$
- $A \to B \to A \to B \to C$
- $A \to B \to A \to B \to A$

The last step can then be used to go around in a cycle. If we on the other hand started at B or C, then we could go to A in one step, and consequently to every other state in the following four. Hence, the probability distribution will "wash out" over time and converge to the stationary one, which results in a decline of $I(X,Y)$ for increasing τ.

Because we measure our X very late in time, meaning t_0 is very large, we will have that $P(X = a) \approx \boldsymbol{\mu}_a$ because of this "washing out". Similarly, we have $P(Y = b) \approx \boldsymbol{\mu}_b$, since the probability distribution will only get attracted more towards $\boldsymbol{\mu}$. As we now increase τ, $P(Y = b \,|\, X = a)$ itself will converge to $\boldsymbol{\mu}_b$ exactly due to the same "washing out" reason. Note that $P(Y = b \,|\, X = a) = (\boldsymbol{M}^\tau)_{ba} \xrightarrow{\tau\to\infty} \boldsymbol{\mu}_b$. And, of course, if $P(X = a, Y = b) = P(X = a) \cdot P(Y = b \,|\, X = a) = \boldsymbol{\mu}_a \cdot \boldsymbol{\mu}_b$, we have $I(X,Y) = 0$. Hence, in a sense the theorem describes how fast $\boldsymbol{M}^\tau \boldsymbol{p}_0$ converges to $\boldsymbol{\mu}$, or, equivalently, $\boldsymbol{M}^\tau$ towards $\boldsymbol{M}_{\boldsymbol{\mu}}$.

Now it's time to dive into the math. In the following, we try to reconstruct the arguments given in the paper. We also adopt the notation $P(a,b) \equiv P(X = a, Y = b)$. By definition of the Kullback-Leibler divergence, we have

$$D(P(X,Y) \,\|\, P'(X,Y)) = \sum_{(a,b)\in S^2} P(a,b) \log_B \frac{P(a,b)}{P(a)P(b)} \quad .$$

The idea is now that $log_B(\bullet)$ is *concave*. Hence, we can upper bound it by its

Taylor expansion of the first degree at the point $x_0 = 1$:

$$\begin{aligned}\log_B(x) &\leq \log_B(x_0) + \log_B'(x_0)(x - x_0)\\ &= 0 + \frac{\ln'(x_0)}{\ln(B)}(x-1)\\ &= \frac{\frac{1}{x_0}}{\ln(B)}(x-1)\\ &= \frac{x-1}{\ln(B)} \quad .\end{aligned}$$

For simplicity, we set $B \coloneqq e$. So our expression becomes

$$\begin{aligned}D(P(X,Y)\,\|\,P'(X,Y)) &\leq \frac{1}{\ln(B)} \sum_{(a,b)\in S^2} P(a,b)\left(\frac{P(a,b)}{P(a)P(b)} - 1\right)\\ &= \sum_{(a,b)\in S^2} P(a,b)\left(\frac{P(a,b)}{P(a)P(b)} - 1\right)\\ &= \left(\sum_{(a,b)\in S^2} P(a,b)\frac{P(a,b)}{P(a)P(b)}\right) - 1\\ &= \left(\sum_{(a,b)\in S^2} \frac{P(a,b)^2}{P(a)P(b)}\right) - 1\\ &\eqqcolon I_R(X,Y) \quad .\end{aligned}$$

The authors of the paper coin this definition for $I_R(X,Y)$ the *rational mutual information*, as it has some useful properties. As discussed, we can approximate $P(a) \approx \boldsymbol{\mu}_a$ and $P(b) \approx \boldsymbol{\mu}_b$, and also $P(b\,|\,a) = (\boldsymbol{M}^\tau)_{ba}$. Thus:

$$\begin{aligned}I_R(X,Y) + 1 &= \sum_{(a,b)\in S^2} \frac{P(a,b)^2}{P(a)P(b)}\\ &= \sum_{(a,b)\in S^2} \frac{P(b\,|\,a)^2 P(a)^2}{P(a)P(b)}\\ &= \sum_{(a,b)\in S^2} \frac{\boldsymbol{\mu}_a}{\boldsymbol{\mu}_b}\left[(\boldsymbol{M}^\tau)_{ba}\right]^2 \quad .\end{aligned}$$

Let us now focus on $(\boldsymbol{M}^\tau)_{ba}$. For simplicity, we consider the case that $\boldsymbol{M}$ is diagonalizable (for the general case see section 2.1.1). Note that since M is irreducible and aperiodic, we have that $1 = \lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_n|$. Hence, let

$$\boldsymbol{M} = \boldsymbol{B}\boldsymbol{D}\boldsymbol{B}^{-1}$$

be the diagonalization of $\boldsymbol{M}$. Of course, we immediately see that $\boldsymbol{M}^\tau =$

$\boldsymbol{B}\boldsymbol{D}^\tau\boldsymbol{B}^{-1}$. Hence, it is easy to verify that

$$(\boldsymbol{M}^\tau)_{ba} = \sum_{c=1}^{n} \lambda_c^\tau \boldsymbol{B}_{bc}(\boldsymbol{B}^{-1})_{ca} \quad .$$

Okay, that was a lot of math. Now it is a good time to reassure ourselves what we actually have achieved. What do we expect $(\boldsymbol{M}^\tau)_{ba}$ to look like for $\tau \to \infty$? $\boldsymbol{\mu}_b$ of course. What does $\boldsymbol{B}$ look like? Well, this is very hard to tell, it at least should have a scaled version of $\boldsymbol{\mu}$ in its first column. But we cannot really infer any information about $\boldsymbol{B}^{-1}$. But we know

$$\begin{aligned}
\boldsymbol{\mu}_b &= \lim_{\tau\to\infty} (\boldsymbol{M}^\tau)_{ba} \\
&= \lim_{\tau\to\infty} \sum_{c=1}^{n} \lambda_c^\tau \boldsymbol{B}_{bc}(\boldsymbol{B}^{-1})_{ca} \\
&= \lambda_1 \boldsymbol{B}_{b1}(\boldsymbol{B}^{-1})_{1a} \quad .
\end{aligned}$$

So we know that

$$(\boldsymbol{M}^\tau)_{ba} = \boldsymbol{\mu}_b \pm \mathcal{O}(|\lambda_2|^\tau) \quad .$$

Note that this is informal writing. It would be more precise to state that $|(\boldsymbol{M}^\tau)_{ba} - \boldsymbol{\mu}_b| \in \mathcal{O}(|\lambda_2|^\tau)$.

This is looking promising, as this means that the discrepancy between $(\boldsymbol{M}^\tau)_{ba}$ and $\boldsymbol{\mu}_b$ decays exponentially. The only thing left to do is translating this exponential decay to the mutual independence measure $I_R(X, Y)$. To this end, we plug our results back into our previous equation. Thus:

$$\begin{aligned}
I_R(X,Y) &= \left(\sum_{(a,b)\in S^2} \frac{\boldsymbol{\mu}_a}{\boldsymbol{\mu}_b} \left[(\boldsymbol{M}^\tau)_{ba}\right]^2 \right) - 1 \\
&= \sum_{(a,b)\in S^2} \left(\frac{\boldsymbol{\mu}_a}{\boldsymbol{\mu}_b} \left[(\boldsymbol{M}^\tau)_{ba}\right]^2 - \boldsymbol{\mu}_a\boldsymbol{\mu}_b \right) \\
&= \sum_{(a,b)\in S^2} \left(\frac{\boldsymbol{\mu}_a}{\boldsymbol{\mu}_b} \left[\boldsymbol{\mu}_b \pm \mathcal{O}(|\lambda_2|^\tau)\right]^2 - \boldsymbol{\mu}_a\boldsymbol{\mu}_b \right) \\
&= \sum_{(a,b)\in S^2} \left(\frac{\boldsymbol{\mu}_a}{\boldsymbol{\mu}_b} \left[\boldsymbol{\mu}_b^2 \pm \mathcal{O}(|\lambda_2|^\tau)\right] - \boldsymbol{\mu}_a\boldsymbol{\mu}_b \right) \\
&= \pm \sum_{(a,b)\in S^2} \frac{\boldsymbol{\mu}_a}{\boldsymbol{\mu}_b} \mathcal{O}(|\lambda_2|^\tau) \quad ,
\end{aligned}$$

where we have used multiple facts about $\boldsymbol{\mu}$. For instance, $\sum_{a\in S} \boldsymbol{\mu}_a = 1$ and thus $\sum_{(a,b)\in S^2} \boldsymbol{\mu}_a \boldsymbol{\mu}_b = 1$, as well as $0 < \boldsymbol{\mu}_a < 1$ for all $a \in S$ (at least for $|S| > 1$). We now use the latter inequality again: We see that we can always bound $\frac{\boldsymbol{\mu}_a}{\boldsymbol{\mu}_b}$ from above, i.e. there exists $\alpha \in \mathbb{R}$ s.t. for all $(a,b) \in S^2$ we have $\frac{\boldsymbol{\mu}_a}{\boldsymbol{\mu}_b} < \alpha$. Hence:

$$\begin{aligned}
|I_R(X,Y)| &\in \sum_{(a,b)\in S^2} \frac{\boldsymbol{\mu}_a}{\boldsymbol{\mu}_b} \mathcal{O}(|\lambda_2|^\tau) \\
\implies |I_R(X,Y)| &\in \sum_{(a,b)\in S^2} \alpha \mathcal{O}(|\lambda_2|^\tau) \\
\implies |I_R(X,Y)| &\in n^2 \alpha \mathcal{O}(|\lambda_2|^\tau) \\
\implies |I_R(X,Y)| &\in \mathcal{O}(|\lambda_2|^\tau) \quad .
\end{aligned}$$

Of course, $I_R(X,Y) \geq 0$, so really $I_R(X,Y) \in \mathcal{O}(|\lambda_2|^\tau)$. Since $0 \leq I(X,Y) \leq I_R(X,Y)$, we also have $I(X,Y) \in \mathcal{O}(|\lambda_2|^\tau)$.

Remark 2.1. The above proof should also work without the approximation $P(a) \approx \boldsymbol{\mu}_a$, so t_0 musn't be big.

Remark 2.2. Based on the proof, we see that if the distance between $\boldsymbol{M}^\tau$ and $\boldsymbol{M}_{\boldsymbol{\mu}}$ experiences exponential decay, we can translate this exponential decay to the mutual information measure $I_R(X,Y)$. Note that we have already established that *all* irreducible aperiodic Markov chains have this property in remark 1.7.

2.1.1 The Defective Case

Nonetheless, we will prove the case that $\boldsymbol{M}$ is not diagonalizable separately and establish the connection to λ_2. The idea is that while not every matrix is diagonalizable, every square matrix over the complex numbers can be put into *Jordan normal form*, which resembles diagonalization. In this form, the matrix is nearly diagonal, except that for each repeated eigenvalue, there may be 1s on the superdiagonal (just above the main diagonal), indicating the presence of generalized eigenvectors.

For example, if there are only three distinct eigenvalues and λ_2 is threefold degenerate, the the Jordan form of $\boldsymbol{M}$ would be

$$\boldsymbol{B}^{-1}\boldsymbol{M}\boldsymbol{B} = \begin{bmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 1 & 0 & 0 \\ 0 & 0 & \lambda_2 & 1 & 0 \\ 0 & 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 0 & 0 & \lambda_3 \end{bmatrix} =: \boldsymbol{D} \quad .$$

Thus, again $\boldsymbol{M}^\tau = \boldsymbol{B}\boldsymbol{D}^\tau\boldsymbol{B}^{-1}$, and the claim is that for our example $\boldsymbol{D}^\tau$ reads as

$$\boldsymbol{D} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2^\tau & \binom{\tau}{1}\lambda_2^{\tau-1} & \binom{\tau}{2}\lambda_2^{\tau-2} & 0 \\ 0 & 0 & \lambda_2^\tau & \binom{\tau}{1}\lambda_2^{\tau-1} & 0 \\ 0 & 0 & 0 & \lambda_2^\tau & 0 \\ 0 & 0 & 0 & 0 & \lambda_3^\tau \end{bmatrix} .$$

All the entries except the ones in the blocks with the binomial coefficient terms are trivial. So let us quickly verify those. For $\tau \coloneqq 1$ it obviously holds when setting $\binom{\tau}{n} \coloneqq 0$ for $n > \tau$. So assume the claim holds for $\tau \coloneqq k$. Then we have

$$\begin{aligned} \boldsymbol{D}^{k+1} &= \boldsymbol{D}^k\boldsymbol{D} \\ &= \begin{bmatrix} 1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2^k & \binom{k}{1}\lambda_2^{k-1} & \binom{k}{2}\lambda_2^{k-2} & 0 \\ 0 & 0 & \lambda_2^k & \binom{k}{1}\lambda_2^{k-1} & 0 \\ 0 & 0 & 0 & \lambda_2^k & 0 \\ 0 & 0 & 0 & 0 & \lambda_3^k \end{bmatrix} \begin{bmatrix} \lambda_1 & 0 & 0 & 0 & 0 \\ 0 & \lambda_2 & 1 & 0 & 0 \\ 0 & 0 & \lambda_2 & 1 & 0 \\ 0 & 0 & 0 & \lambda_2 & 0 \\ 0 & 0 & 0 & 0 & \lambda_3 \end{bmatrix} . \end{aligned}$$

Thus, at $j \geq i,\ d \coloneqq j - i$, $(\boldsymbol{D}^k)_{i,j} = \binom{k}{d}\lambda_2^{k-d}$, and hence we have

$$\begin{aligned} (\boldsymbol{D}^{k+1})_{i,j} &= (\boldsymbol{D}^k)_{i,j-1}(\boldsymbol{D})_{j-1,j} + (\boldsymbol{D}^k)_{i,j}(\boldsymbol{D})_{j,j} \\ &= (\boldsymbol{D}^k)_{i,i+(d-1)} + (\boldsymbol{D}^k)_{i,i+d}\lambda_2 \\ &= \binom{k}{d-1}\lambda_2^{k-d+1} + \binom{k}{d}\lambda_2^{k-d}\lambda_2 \\ &= \binom{k}{d-1}\lambda_2^{k-d+1} + \binom{k}{d}\lambda_2^{k-d+1} \\ &= \left(\binom{k}{d-1} + \binom{k}{d}\right)\lambda_2^{k-d+1} \\ &\overset{\checkmark}{=} \binom{k+1}{d}\lambda_2^{k+1-d} \qquad , \end{aligned}$$

just as expected.

This was just an example, but it is easy to see that we can generalize this, and we get that the absolute value of every entry in $\boldsymbol{D}^\tau$, except the top left 1, is $\mathcal{O}(|\lambda_2|^\tau)$. Note that $|\binom{\tau}{d}\lambda_2^{\tau-d}| \in \mathcal{O}(|\lambda_2|^\tau)$ for each $d \in \mathbb{N}$.

The rest is trivial, as all the the entries in $\boldsymbol{B}$ and $\boldsymbol{B}^{-1}$ are really just constants, and hence when calculating $(\boldsymbol{M}^\tau)_{ba} = (\boldsymbol{B}\boldsymbol{D}^\tau\boldsymbol{B}^{-1})_{ba}$, we have

$$
\begin{aligned}
(\boldsymbol{B}\boldsymbol{D}^\tau\boldsymbol{B}^{-1})_{ba} &= \left(\boldsymbol{B}\left(\begin{bmatrix}1 & 0 & \cdots & 0\\ 0 & 0 & \cdots & 0\\ \vdots & \vdots & \ddots & \vdots\\ 0 & 0 & \cdots & 0\end{bmatrix} + \begin{bmatrix}c_{11} & c_{12} & \cdots & c_{1n}\\ c_{21} & c_{22} & \cdots & c_{2n}\\ \vdots & \vdots & \ddots & \vdots\\ c_{n1} & c_{n2} & \cdots & c_{nn}\end{bmatrix}\mathcal{O}(|\lambda_2|^\tau)\right)\boldsymbol{B}^{-1}\right)_{ba}\\
&= \boldsymbol{\mu}_b \pm \mathcal{O}(|\lambda_2|^\tau) \quad ,
\end{aligned}
$$

for some $c_{ij} \in \mathbb{C},\ |c_{ij}| = 1$. The rest of the proof is identical to the one given.

2.2 No Markov Chain with Power-Law Decay

We are interested in cases where $I(X,Y)$ decays to 0, as this is implied by a power-law decay. This means that for increasing τ, we get in the limit $\tau \to \infty$ that $P(a,b) = P(a)P(b \mid a) = P(a)(\boldsymbol{M}^\tau)_{ba} \stackrel{!}{=} P(a)P(b)$, and hence $(\boldsymbol{M}^\tau)_{ba} = P(b)$ for every $a \in S$. Clearly, this means that $\boldsymbol{M}^\tau$ must converge to a stationary matrix $\boldsymbol{M}_{\boldsymbol{\mu}}$, where all columns are equal, at least given the case that $P(b)$ converges for all $b \in S$. But thanks to the following lemma, we can assume that this is the case:

Lemma 2.1 (Convergence of P(b) is Necessary for Power-Law Decay)**.** *In order for $I(X,Y)$ to converge to 0 for increasing τ, we must have that $\lim_{t\to\infty} \boldsymbol{M}^t\boldsymbol{p}_0 = \boldsymbol{\mu}$ for every $\boldsymbol{p}_0 \in \Delta$ (and hence $P(b)$ converges).*

Proof. In order for $I(X,Y)$ to converge to 0, we still must have that $|(\boldsymbol{M}^\tau)_{ba} - P(X_{t_0+\tau} = b)| \xrightarrow{\tau\to\infty} 0$ for every $a \in S$. But note that for this to happen we must have close to equal columns for increasing τ, because $(\boldsymbol{M}^\tau)_{ba}$ must be independent of a. But such a matrix must be stationary (and hence $P(b)$ converges), so we cannot have $I(X,Y) = 0$ if $P(b)$ does not converge. ☐

Theorem 2.1 (No Markov Chain with Power-Law Decay)**.** *There is no Markov chain $\boldsymbol{M}$ with $I(X_{t_0}, X_{t_0+\tau}) \in \mathcal{O}(\tau^{-\alpha})$ and $I(X_{t_0}, X_{t_0+\tau}) \in \Omega(\tau^{-\beta})$ for some $\alpha, \beta \in \mathbb{R}_{>0}$.*

Proof. For the sake of contradiction, assume that $\boldsymbol{M}$ produces power-law decay. Due to the previous results, we know that $\boldsymbol{M}^\tau$ converges to a stationary matrix $\boldsymbol{M}_{\boldsymbol{\mu}}$. (The following discussion is facultative; it aims to enhance the reader's understanding). $\boldsymbol{M}_{\boldsymbol{\mu}}$ must contain 0-entries, as if it didn't, it would mean that $\boldsymbol{M}$ is irreducible and aperiodic based on corollary 1.2, and hence $\boldsymbol{M}$ would have exponential decay.

Thus, let's focus on all the rows of $\boldsymbol{M_\mu}$ with non-0-entries with associated states S_C. We see that for all $i \in S_C, j \in S$ we have that $(\boldsymbol{M_\mu})_{ij} > 0$ (especially for $j \in S_C$). Hence, the set S_C describes a closed communication class. Furthermore, there cannot be another closed communication class, since in a closed communication class the transition probabilities between any two states cannot approach 0. Thus, all other states $S \setminus S_C =: S_O$ are in an open communication class.

For indexing reasons, we assume without loss of generality that $S_C = \{1, \dots, n'\} \subseteq S$. Since S_C is a closed communication class, we have that for all $j \in S_C, i \in S_O, t \in \mathbb{N}_{>0} : (\boldsymbol{M}^t)_{ij} = 0$ (especially for $t \coloneqq 1$). Thus for all $i, j \in S_C, t \in \mathbb{N}_{>0}$:

$$\begin{aligned}
(\boldsymbol{M}^{t+1})_{ij} &= (\boldsymbol{M}^t \boldsymbol{M})_{ij} \\
&= \sum_{k \in S} (\boldsymbol{M}^t)_{ik} (\boldsymbol{M})_{kj} \\
&= \sum_{k \in S_C} (\boldsymbol{M}^t)_{ik} (\boldsymbol{M})_{kj} \quad .
\end{aligned}$$

Now, let $\boldsymbol{M}_{S_C}$ be the sub-matrix of $\boldsymbol{M}$ containing and only containing the transition probability entries for the states in S_C (with our assumption $\boldsymbol{M}_{S_C}$ is the top left sub-matrix of $\boldsymbol{M}$). From our result, we see that $(\boldsymbol{M}^t)_{S_C} = (\boldsymbol{M}_{S_C})^t$.

But $(\boldsymbol{M}^t)_{S_C}$ converges to a positive matrix, and hence so must $(\boldsymbol{M}_{S_C})^t$. Hence, $\boldsymbol{M}_{S_C}$ must be irreducible aperiodic based on corollary 1.2 again. Thus, $(\boldsymbol{M}_{S_C})^t$ converges with exponential decay (with a basis of $|\lambda_2| < 1$ of $\boldsymbol{M}_{S_C}$), and hence so does $(\boldsymbol{M}^t)_{S_C}$.

Quickly note that λ_2 of $\boldsymbol{M}_{S_C}$ is an eigenvalue of $\boldsymbol{M}$ as well: There must be (at least) one associated eigenvector $v \in \mathbb{C}^{|S_C|}$ s.t. $\boldsymbol{M}_{S_C} v = \lambda_2 v$. Now, extend v to an eigenvector v' of $\boldsymbol{M}$ by adding zeros in all the places associated with states in S_O (here: at the end). For $i \in S_C$ we have:

$$\begin{aligned}
(\boldsymbol{M}v')_i &= \sum_{j \in S} \boldsymbol{M}_{ij} v'_j \\
&\underset{\text{0-entries of v'}}{=} \sum_{j \in S_C} \boldsymbol{M}_{ij} v'_j \\
&= \sum_{j \in S_C} \boldsymbol{M}_{ij} v_j \\
&\underset{i \in S_C}{=} \sum_{j \in S_C} (\boldsymbol{M}_{S_C})_{ij} v_j \\
&= (\boldsymbol{M}_{S_C} v)_i \\
&= \lambda_2 v_i \overset{\checkmark}{=} \lambda_2 v'_i \quad ,
\end{aligned}$$

and for $i \in S_O$ it follows that

$$\begin{aligned}(\boldsymbol{M}v')_i &= \sum_{j\in S} \boldsymbol{M}_{ij} v'_j \\ &\underset{\text{0-entries of v'}}{=} \sum_{j \in S_C} \boldsymbol{M}_{ij} v'_j \\ &\underset{i\in S_O}{=} \sum_{j\in S_C} 0 \cdot v'_j \\ &= 0 \overset{\checkmark}{=} \lambda_2 v'_i \quad .\end{aligned}$$

The only things left to do are, first, to verify that $|\lambda_2|$ of $\boldsymbol{M}_{S_C}$ is less than or equal to the absolute value of the second largest eigenvalue of $\boldsymbol{M}$; and second, to show the convergence of $\boldsymbol{M}_{S_O}$ and that it also exhibits exponential decay (with a base less than or equal to $|\lambda_2|$ of $\boldsymbol{M}$). Luckily, we achieve all our goals thanks to the following observation:

(Start of the actual proof.) Since $\boldsymbol{M}^t$ converges to $\boldsymbol{M}_{\boldsymbol{\mu}}$, there must exist an $m \in \mathbb{N}_{>0}$ s.t. $\boldsymbol{M}^m$ and $\boldsymbol{M}^{m+1}$ both have a positive row. Thus, thanks to corollary 1.3 we can apply the Perron-Frobenius Theorem (with the exception that the eigenvector must not be positive) to $\boldsymbol{M}^m$ and $\boldsymbol{M}^{m+1}$, and thanks to lemma 1.5 also to $\boldsymbol{M}$ itself. Of course, $\lambda_{max} = 1$ with the associated eigenvector $\boldsymbol{\mu}$. So we know that $|\lambda_2| < 1$ of $\boldsymbol{M}_{S_C}$ is less than or equal to the absolute value of the second largest eigenvalue of $\boldsymbol{M}$.

Furthermore, we of course have that $\lim_{\tau\to\infty}(\boldsymbol{M}^\tau)_{ba} = \boldsymbol{\mu}_b$ $(:= \lim_{\tau\to\infty} P(b))$. But those were the only two assumptions made in section 2.1.1. Thus, by the same logic, we get $I(X_{t_0}, X_{t_0+\tau}) \in \mathcal{O}(|\lambda_2|^\tau)$. $\square$