1 Mutual Information in 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 $\boldsymbol{\mu}$. In case the reader is unfamiliar with these terms or this result, one can read them in appendix **??**.

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 fact later.

Our goal now is to quantify the mutual information $I(X;Y)$ between 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 P_X and P_Y, that is $P' \coloneqq P_X \otimes P_Y$. We use the Kullback-Leibler divergence, so our target expression becomes

$$I(X;Y) = D_{\mathrm{KL}}(P_{(X,Y)} \,\|\, P_X \otimes P_Y) \quad .$$

For further details, see appendix **??**.

1.1 Exponential Decay in Irreducible Aperiodic Markov Chains

Note that $I(X;Y)$ 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:

Theorem 1.1 (No Power-Law in Markov Chains)**.** *Let X and Y be random variables from an irreducible aperiodic Markov chain at times t_0 and $t_0 + \tau$ respectively. Let $\boldsymbol{M}$ be the transition matrix, and let $|\lambda_2|$ denote the second largest absolute value of its eigenvalues. Then:*

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

Remark 1.1. Based on theorem **??**, it follows that Markov chains are incapable of weak power-law behavior (and hence also strong power-law behavior).

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 1.

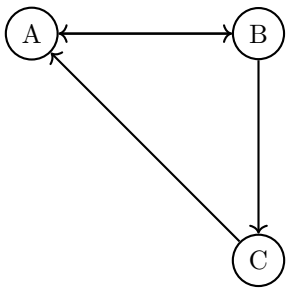

Figure 1: A simple irreducible aperiodic Markov chain. Note that if $X_{t_0} = C$, then we know that $X_{t_0+1} = A$.

If $\tau = 1$, i.e. we consider the mutual information of two 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:

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

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 spread out over time and converge to the stationary probability distribution, 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. 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 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 $\boldsymbol{M}_{\boldsymbol{\mu}}$.

Proof. 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

$$I(X;Y) = D_{\mathrm{KL}}(P_{(X,Y)} \,\|\, P_X \otimes P_Y) = \sum_{(a,b)\in S^2} P(a,b) \log_B \frac{P(a,b)}{P(a)P(b)} \quad .$$

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

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

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 1.1.1). 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 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 1.2. The above proof should also work without the approximation $P(a) \approx \boldsymbol{\mu}_a$, so t_0 doesn't have to be be large.

Remark 1.3. 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 **??**.

1.1.1 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

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

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

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

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

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

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 $\mathcal{O}(|\lambda_2^+|^\tau)$, where λ_2^+ is defined s.t. $|\lambda_2^+| = |\lambda_2| + \epsilon$ for some $\epsilon \in \mathbb{R}_{>0}$. 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.

1.2 No Markov Chain with Power-Law Behavior

We are interested in cases where $I(X;Y)$ decays to 0, as this is implied by power-law behavior (see definition **??** or definition **??**). This means that for increasing τ, we get in the limit $\tau \to \infty$ that $P(a,b) = P(a)P(b \mid a) = P(a)(\boldsymbol{M}^\tau)_{ba} \overset{!}{=} P(a)P(b)$, 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 1.1 (Convergence of P(b) is Necessary for Power-Law Behavior)**.** *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). □

Theorem 1.2 (No Markov Chain with Power-Law Behavior)**.** *There is no Markov chain with weak power-law behavior (and hence also strong power-law behavior).*

Proof. For the sake of contradiction, assume that $\boldsymbol{M}$ has weak lower bound power-law behavior and decaying behavior ($\Longleftarrow$ power-law behavior). Due to the previous results, we know that $\boldsymbol{M}^\tau$ converges to a stationary matrix $\boldsymbol{M_\mu}$. (The following discussion is facultative; it aims to enhance the reader's understanding). $\boldsymbol{M_\mu}$ must contain 0-entries, as if it didn't, it would mean that $\boldsymbol{M}$ is irreducible and aperiodic based on corollary **??**, 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 submatrix 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 submatrix 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 **??** 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 $\boldsymbol{v} \in \mathbb{C}^{|S_C|}$ s.t. $\boldsymbol{M}_{S_C}\boldsymbol{v} = \lambda_2 v$. Now, extend $\boldsymbol{v}$ to

an eigenvector $\boldsymbol{v}'$ 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:

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

and for $i \in S_O$ it follows that

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

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 **??** we can apply the Perron-Frobenius Theorem (with the exception that the eigenvector is non-negative instead of positive) to $\boldsymbol{M}^m$ and $\boldsymbol{M}^{m+1}$, and thanks to lemma **??** 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 $\lim_{\tau\to\infty}(\boldsymbol{M}^\tau)_{ba} = \boldsymbol{\mu}_b$ ($\coloneqq \lim_{\tau\to\infty} P(b)$). But those were the only two assumptions made in section 1.1.1. Thus, by the same logic, we get $I(X_{t_0}; X_{t_0+\tau}) \in \mathcal{O}(|\lambda_2^+|^\tau)$. □