1 Mutual Information in Markov Chains

1.1 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 $S = \{1, \dots, n\}$, then we have that

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

where $\boldsymbol{M_\mu}$ 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 $\boldsymbol{M}^{t_0} \approx \boldsymbol{M_\mu}$. 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 1.

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:

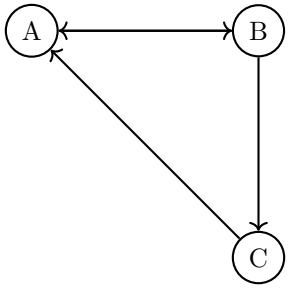

Figure 1: A simple irreducible aperiodic Markov chain. Note that if $X_{t_0} = C$, then we know that $X_{t_0+1} = A$.

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

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

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

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. Note that since M is irreducible and aperiodic, we have that $1 = \lambda_1 > |\lambda_2| \geq \cdots \geq |\lambda_n|$. The authors provide proof for the other case as well. But for now, 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. Note that this step deviates from the procedure in the paper (own interpretation, informal!). 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 1.1. The above proof should also work without the approximation $P(a) \approx \boldsymbol{\mu}_a$, so t_0 musn't be big.

Remark 1.2. 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 **??**.

1.1.1 The Defective Case

Nonetheless, we will prove the case that $\boldsymbol{M}$ is not diagonalizable seperately 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

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

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 $\mathcal{O}(|\lambda_2|^\tau)$. 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

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

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

1.2 Other Cases

We are interested in cases where $I(X,Y)$ decays to 0. This means that for incrasing τ, we get in the limit $\tau \to \infty$ that $P(a,b) = P(a)P(a \mid b) = P(a)(\boldsymbol{M}^\tau)_{ba} \stackrel{!}{=} P(a)P(b)$, 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 $b \in S$.

1.2.1 Case: P(b) Does Not Converge

The problem is that the convergence of $P(b)$ might depend on the initial probability distribution vector $\boldsymbol{p}_0$. But actually, note that we still 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.

1.2.2 Case: P(b) Does Converge

So now we assume that $\boldsymbol{M}^\tau$ converges to a stationary matrix $\boldsymbol{M}_{\boldsymbol{\mu}}$. $\boldsymbol{M}_{\boldsymbol{\mu}}$ must contain 0-entries, as if it didn't, it would mean that $\boldsymbol{M}$ is irreducible and aperiodic based on remark **??**, and we already have covered that.

So this must mean that the set S_C of all the states defined by the rows of $\boldsymbol{M}_{\boldsymbol{\mu}}$ that are not 0 form the only closed communication class (and hence all other states are in an open communication class. All those states form the set S_O). Hence, after $n = |S|$ ticks we already have that all the rows associated with states in S_O are filled up with 0-entries.

So we look at $\boldsymbol{M}^n$. Imagine for simplicity that the top rows are zeroed out. But then we can *solely* focus on the sub-matrix $\boldsymbol{M}_{S_C}$ defined by the set of states S_C. But $\boldsymbol{M}_{S_C}$ converges to a positiv matrix, so $\boldsymbol{M}_{S_C}$ must be aperiodic based on remark **??** again. This means that the irreducible aperiodic matrix $\boldsymbol{M}_{S_C}$ converges with exponential decay, and hence so does $\boldsymbol{M}^n$. Furthermore, we also know that $\boldsymbol{M}$ itself converges, and hence it also must experience exponential decay towards $\boldsymbol{M}_{\boldsymbol{\mu}}$.