

1 Mutual Information

1.1 Exponential Decay in Markov Chains

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

$$\lim_{i \rightarrow \infty} \mathbf{M}^i = \mathbf{M}_\mu \quad ,$$

where \mathbf{M}_μ is the matrix whose columns all consist of the unique stationary probability distribution μ .

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 $\mathbf{M}^{t_0} \approx \mathbf{M}_\mu$. 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 $P'(X, Y) := P(X) \cdot P(Y)$. We use the Kullback-Leibler divergence, so our target expression becomes

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

Note that of course this divergence $I(X, Y) := D(P(X, Y) \parallel P'(X, Y))$ depends on the properties of \mathbf{M} , as well as on τ . Because \mathbf{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 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:

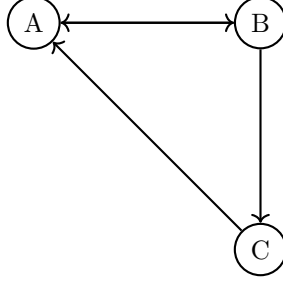


Figure 1: A simple irreducible aperiodic Markov chain. Note that if $X_{t_0} = C$, then we know that $X_{t_0+1} = A$.

- $A \rightarrow B \rightarrow C \rightarrow A \rightarrow B$
- $A \rightarrow B \rightarrow A \rightarrow B \rightarrow C$
- $A \rightarrow B \rightarrow A \rightarrow B \rightarrow A$

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 $P(X = a) \approx \mu_a$ because of this "washing out". Similarly, we have $P(Y = b) \approx \mu_b$, since the probability distribution will only get attracted more towards μ . As we now increase τ , $P(Y = b | X = a)$ itself will converge to μ_b exactly due to the same "washing out" reason. Note that $P(Y = b | X = a) = (\mathbf{M}^\tau)_{ba} \xrightarrow{\tau \rightarrow \infty} \mu_b$. And, of course, if $P(X = a, Y = b) = P(X = a) \cdot P(Y = b | X = a) = \mu_a \cdot \mu_b$, we have $I(X, Y) = 0$. Hence, in a sense the theorem describes how fast $\mathbf{M}^\tau \mathbf{p}_0$ converges to μ , or, equivalently, \mathbf{M}^τ towards \mathbf{M}_μ .

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 $B := e$. 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 $P(a) \approx \mu_a$ and $P(b) \approx \mu_b$, and also $P(b|a) = (\mathbf{M}^\tau)_{ba}$. Thus:

$$\begin{aligned}
I_R(X, Y) + 1 &= \sum_{(a,b) \in S^2} \frac{P(a, b)^2}{P(a)P(b)} \\
&= \sum_{(a,b) \in S^2} \frac{P(b|a)^2 P(a)^2}{P(a)P(b)} \\
&= \sum_{(a,b) \in S^2} \frac{\mu_a}{\mu_b} [(\mathbf{M}^\tau)_{ba}]^2 \quad .
\end{aligned}$$

Let us now focus on $(\mathbf{M}^\tau)_{ba}$. For simplicity, we consider the case that \mathbf{M} is diagonalizable. Note that since \mathbf{M} is irreducible and aperiodic, we have that $1 = \lambda_1 > |\lambda_2| \geq \dots \geq |\lambda_n|$. The authors provide proof for the other case as well. But for now, let

$$\mathbf{M} = \mathbf{BDB}^{-1}$$

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

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

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

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 $(\mathbf{M}^\tau)_{ba}$ to look like for $\tau \rightarrow \infty$? $\boldsymbol{\mu}_b$ of course. What does \mathbf{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 \mathbf{B}^{-1} . But we know

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

So we know that

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

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

This is looking promising, as this means that the discrepancy between $(\mathbf{M}^\tau)_{ba}$ 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:

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

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

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

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. Based on the proof, we see that if the distance between \boldsymbol{M}^τ and \boldsymbol{M}_μ 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 irreducible aperiodic Markov chains have this property in remark ??.

1.2 The Degenerate Case

Nonetheless, we will prove the case that \boldsymbol{M} is not diagonalizable separately and establish the connection to λ_2 .