Criticality in Formal Languages

Jonas Peters

May 26, 2025

Contents

1 Model Framework **3**

2 Mutual Information in Markov Chains **8**

2.1 Exponential Decay in Irreducible Aperiodic Markov Chains . . . 8

2.1.1 The Defective Case 13

2.2 No Markov Chain with Power-Law Decay 14

3 No Power-Law in Hidden Markov Models **17**

3.1 Conclusions for Model Selection 20

4 Tensor Networks **22**

5 Binary Tree Tensor Networks **28**

5.1 Bulk Marginal Property 28

5.2 Binary Tree Tensor Networks are Universal Approximators . . . 30

5.3 Restricting Parameters 31

A Markov Chains **34**

A.1 Properties 35

A.1.1 Irreducibility 36
A.1.2 Aperiodicity 38
A.2 Irreducible Aperiodic Markov Chains 39
A.3 Perron-Frobenius Theorem 41

B Information Theory **45**

B.1 Entropy 45
B.1.1 Joint, Conditional, and Cross Entropy 47
B.1.2 Properties of Entropy 48
B.2 Kullback-Leibler Divergence 50
B.3 Mutual Information 52
B.3.1 Data Processing Inequality 55
B.4 Bounding Mutual Information via Matrix Rank of the Joint Distribution 56

1 Model Framework

We are interested in models with asymptotically power-law decay of the mutual information measure with respect to the distance between the tokens in the sequence. So far so good. But what does it *actually* mean?

The tokens, represented by random variables X_t, are elements of a finite alphabet Σ. The distance between X_t and $X_{t+\tau}$ is τ, and for every t and every $\tau > 0$ we want to bound

$$I(X_t, X_{t+\tau}) \in \Omega(\tau^{-\alpha}),\ I(X_t, X_{t+\tau}) \in \mathcal{O}(\tau^{-\beta}) \quad ,$$

for some fixed $\alpha, \beta \in \mathbb{R}_{>0}$. The first condition is the important one, while the latter ensures that $I(X_t, X_{t+\tau}) \xrightarrow{\tau \to \infty} 0$. We also may replace the latter condition by this one.

This was straight forward. The challenging part is to define what a model is. In the case of Markov chains this seems trivial: We define a finite set of parameters (the transition probabilities), and we get a model over Σ^*, that is for every $n \in \mathbb{N}$ the model defines a probability measure over Σ^n. Thus:

Definition 1.1 (Model over Σ^*)**.** A model S over Σ^* is a function $S : \mathbb{N} \times \Sigma^* \mapsto [0,1],\ (n, w) \mapsto p$, for $n \in \mathbb{N},\ w \in \Sigma^n,\ p \in [0,1]$ s.t. $\sum_{w \in \Sigma^n} S(n, w) = 1$. S assigns the probability p to the word w of length n.

But really, we want to restrain S in order to have reasonable time and space complexity, and to ensure the model is *reasonable*, which means that the language of $S_n(w)$ should look *similar* to $S_{n+d}(w)$, whatever this might mean, where we used the notation $S_n(w) \equiv S(n, w)$. We also write w_i for X_i. Really, w is a 1-indexed String of X_i.

We present one strict definition for this *similarity* in the following definition:

Definition 1.2. We say S has the *bulk marginal property* iff for every $n \in \mathbb{N},\ w \in \Sigma^{n+1}$ it holds true that

$$\sum_{w_{n+1} \in \Sigma} S_{n+1}(w) = S_n(w_{-\{n+1\}}) \quad .$$

Remark 1.1. Markov chains have the bulk marginal property.

Lemma 1.1. *For every $d \in \mathbb{N}$, let $I \coloneqq [n+d] \setminus [n] = \{n+1, \ldots, n+d\}$. Then, if S has the bulk marginal property, we have for every $w \in \Sigma^{n+d}$:*

$$\sum_{w_I \in \Sigma^d} S_{n+d}(w) = S_n(w_{-I}) \quad .$$

Proof. We use induction over d. The base case follows directly from the definition of the bulk marginal property. Thus, assume the claim holds for some $d \coloneqq k$. Then we have

$$\begin{aligned}\sum_{w_I \in \Sigma^{k+1}} S_{n+k+1}(w) &= \sum_{w_{I\setminus\{k+1\}} \in \Sigma^k} \sum_{w_{k+1} \in \Sigma} S_{n+k+1}(w) \\ &\underset{\text{bulk marginal property}}{=} \sum_{w_{I\setminus\{k+1\}} \in \Sigma^k} S_{n+k}(w_{-\{k+1\}}) \\ &\underset{\text{inductive hypothesis}}{=} S_n(w_{-I}) \quad ,\end{aligned}$$

which concludes the induction step. □

Definition 1.3 (Induced Bulk Marginal Model)**.** Based on the model S, we can construct an *induced bulk marginal* model S^* by defining S^*_n recursively as

- $S^*_1 \coloneqq S_1$,
- $S^*_{n+1}(w) \coloneqq S^*_n(w_{-\{n+1\}}) \dfrac{S_{n+1}(w)}{\sum_{w_{n+1} \in \Sigma} S_{n+1}(w)}$.

Remark 1.2. If $\sum_{w_{n+1} \in \Sigma} S_{n+1}(w) = 0$, we might set $S^*_{n+1}(w) \coloneqq S^*_n(w_{-\{n+1\}}) \dfrac{1}{|\Sigma|}$.

Lemma 1.2. *The induced bulk marginal model S^* indeed has the bulk marginal property.*

Proof. We have:

$$\begin{aligned}\sum_{w_{n+1} \in \Sigma} S^*_{n+1}(w) &= \sum_{w_{n+1} \in \Sigma} S^*_n(w_{-\{n+1\}}) \frac{S_{n+1}(w)}{\sum_{w_{n+1} \in \Sigma} S_{n+1}(w)} \\ &= \frac{S^*_n(w_{-\{n+1\}})}{\sum_{w_{n+1} \in \Sigma} S_{n+1}(w)} \sum_{w_{n+1} \in \Sigma} S_{n+1}(w) \\ &\overset{\checkmark}{=} S^*_n(w_{-\{n+1\}}) \quad .\end{aligned}$$

□

Now, we want to look at how we might restrict our model $(S_n)_{n\in\mathbb{N}} \equiv S$. One approach might be to define a model structure for every $n \in \mathbb{N}$. To this end, we define S_n by some finite parameters $\boldsymbol{\theta}_n$ over the *model space* $\mathcal{S}(n) \equiv \mathcal{S}_n$, which specifies the structure of our models. Thus:

$$S_n \in \{S_n(\boldsymbol{\theta}_n) : \boldsymbol{\theta}_n \in \Theta_n\} =: \mathcal{S}_n \quad ,$$

where Θ_n is the set of all possible parameters of $\mathcal{S}_n$. We write $S_{n,\boldsymbol{\theta}_n}$ for S_n with parameters $\boldsymbol{\theta}_n$. Hence, $(S_n)_{n\in\mathbb{N}}$ is completely defined by $(\mathcal{S}_n, \boldsymbol{\theta}_n)_{n\in\mathbb{N}}$.

Remark 1.3. The parameter space Θ_n may consist of parameter vectors with varying lengths. The same model S_n may be defined by two parameter vectors with very different sizes over the same model space $\mathcal{S}_n$ or potentially two different model spaces. Thus, the parametrization complexity depends of the model space $\mathcal{S}$.

Definition 1.4 (Family of Models)**.** We say $(S_n)_{n\in\mathbb{N}}$ is a *family of models* over the model space $\mathcal{S}$ iff $S_n \in \mathcal{S}_n$ for every $n \in \mathbb{N}$. As a shorthand, we write $S \in \mathcal{S}$.

For our model S, we want power-law decay in the mutual information with respect to τ between *any* two variables X_t, $X_{t+\tau}$, i.e. it has to hold for every t and *every* S_n. But what does this actually mean?

Definition 1.5. We define $i_{S_n}(\tau)$ and $I_{S_n}(\tau)$ to be the minimal and maximal mutual information between any two variables of S_n with distance τ. Formally, let $X_t, X_{t+\tau}$ $(t + \tau \leq n)$ be random variables with distributions defined by S_n. Then:

- $i_{S_n}(\tau) \coloneqq \min_{t\in[n-\tau]} I(X_t, X_{t+\tau}) \quad ,$
- $I_{S_n}(\tau) \coloneqq \max_{t\in[n-\tau]} I(X_t, X_{t+\tau}) \quad .$

Definition 1.6 (Strong Power-Law Behavior)**.** A model S has *strong lower bound power-law behavior* iff there exist constants $c_\alpha, \alpha \in \mathbb{R}_{>0}$ s.t. for every $n \in \mathbb{N}$ it holds true that $i_{S_n}(\tau) \geq c_\alpha \tau^{-\alpha}$. Similarly, S has *upper bound power-law behavior* iff there exist constants $c_\beta, \beta \in \mathbb{R}_{>0}$ s.t. for every $n \in \mathbb{N}$ it holds true that $I_{S_n}(\tau) \leq c_\beta \tau^{-\beta}$. Furthermore, S has *decaying behavior* iff for every $n \in \mathbb{N}$ we have $I_{S_{n+\tau}}(\tau) \xrightarrow{\tau\to\infty} 0$. Lastly, S has *strong power-law behavior* iff it has strong lower bound and upper bound power-law behavior (alternatively decaying behavior instead of upper bound power-law behavior).

Remark 1.4. For a model S^* with the bulk marginal property we can replace "for every $n \in \mathbb{N}$" in definition 1.6 with "for $n \to \infty$" thanks to lemma 1.1.

Proposition 1.1. *Upper bound power-law behavior implies decaying behavior.*

Proof. Assume model S has upper bound power-law behavior. Then there exist constants $c_\beta, \beta \in \mathbb{R}_{>0}$ s.t. for every $n \in \mathbb{N}$ it holds true that $I_{S_n}(\tau) \leq c_\beta \tau^{-\beta}$, especially for $n \coloneqq n' + \tau$. Thus, for every $n' \in \mathbb{N}$:

$$I_{S_{n'+\tau}}(\tau) \leq c_\beta \tau^{-\beta} \xrightarrow{\tau \to \infty} 0 \quad .$$

□

Definition 1.7. We define $\overline{i_{S_n}}$ to be the minimal mutual information between any two variables over S_n with arbitrary distance τ. Formally, let X_i, X_j $(1 \leq i < j \leq n)$ be random variables with distributions defined by S_n. Then:

$$\overline{i_{S_n}} \coloneqq \min_{(i,j) \in [n]^2, i<j} I(X_i, X_j) = \min_{\tau \in [n-1]} i_{S_n}(\tau) \quad .$$

Definition 1.8 (Weak Power-Law Behavior)**.** A model S has *weak lower bound power-law behavior* iff $\overline{i_{S_n}} \in \Omega(n^{-\alpha})$ for some $\alpha \in \mathbb{R}_{>0}$. Additionally, S has *weak power-law behavior* iff it has weak lower bound and upper bound power-law behavior (alternatively decaying behavior instead of upper bound power-law behavior).

Theorem 1.1 (Every Token has Power-Law Decay in Models with the Bulk Marginal Property and Weak Power-Law Behavior)**.** *Let S be a model that satisfies the bulk marginal property and exhibits weak lower bound power-law behavior. Then, there exists an $\alpha \in \mathbb{R}_{>0}$ s.t. for every X_t, $I(X_t, X_{t+\tau}) \in \Omega(\tau^{-\alpha})$ (where X_t and $X_{t+\tau}$ are sampled over $S_{t+\tau}$, or, equivalently, any $S_{t+\tau+k}$).*

Proof. Since S has weak lower bound power-law behavior, there exist $\alpha', c' \in \mathbb{R}_{>0}$ s.t. $\overline{i_{S_n}} \geq c' n^{-\alpha'}$. Then, for every $t \in \mathbb{N}$, we have for $n \coloneqq t + \tau$ by the definition of $\overline{i_{S_n}}$:

$$\begin{aligned} I(X_t, X_{t+\tau}) &\geq \overline{i_{S_{t+\tau}}} \\ &\geq c'(t+\tau)^{-\alpha'} \\ &= c'\tau^{-\alpha'}(\frac{t}{\tau} + 1)^{-\alpha'} \\ &\geq c'\tau^{-\alpha'}(t+1)^{-\alpha'} \quad . \end{aligned}$$

Since S is has the bulk marginal property, this inequality holds when sampling over any $S_{t+\tau+k},\ k \in \mathbb{N}$. Now, set $\alpha \coloneqq \alpha'$ and $c \coloneqq c'(t+1)^{-\alpha'}$. Note that α does not depend on t. Finally, we see that $I(X_t, X_{t+\tau}) \geq c\tau^{-\alpha}$. Thus, we get $I(X_t, X_{t+\tau}) \in \Omega(\tau^{-\alpha})$. □

Remark 1.5. If additionally S had decaying behavior, then of course we would also have $I(X_t, X_{t+\tau}) \xrightarrow{\tau \to \infty} 0$.

Remark 1.6. The importance of this implication might depend on the context. However, this theorem proves to be very useful when considering its contraposition. In fact, we will use this contraposition later to disprove weak power-law behavior for Markov chains (and hence also strong power-law behavior).

Remark 1.7. It is crucial for S to have the bulk marginal property in theorem 1.1, or else $I(X_t, X_{t+\tau})$ might depend on S_n, and we cannot exclude $I(X_t, X_{t+\tau}) \xrightarrow{n\to\infty} 0$.

Proposition 1.2. *Strong lower bound power-law behavior implies weak lower bound power-law behavior.*

Proof. Assume model S has strong lower bound power-law behavior. Thus, it follows that there exist $c_\alpha, \alpha \in \mathbb{R}_{>0}$ s.t. for all $n \in \mathbb{N}$ we have that $i_{S_n}(\tau) \geq c_\alpha \tau^{-\alpha}$. Hence:

$$\begin{aligned}
\overline{i_{S_n}} &= \min_{\tau\in[n-1]} i_{S_n}(\tau) \\
&\geq \min_{\tau\in[n-1]} c_\alpha \tau^{-\alpha} \\
&\geq c_\alpha (n-1)^{-\alpha} \\
&= c_\alpha n^{-\alpha} (1 - \frac{1}{n})^{-\alpha} \\
&\geq c_\alpha n^{-\alpha} 1^{-\alpha} \\
&= c_\alpha n^{-\alpha} \quad .
\end{aligned}$$

It follows that $\overline{i_{S_n}} \in \Omega(n^{-\alpha})$, and hence S has weak lower bound power-law behavior. □

Remark 1.8. Weak lower bound power-law behavior does *not* imply strong lower bound power-law behavior, not even for models with the bulk marginal property. To see this, note that we might have $i_{S_n}(1) \xrightarrow{n\to\infty} 0$ for some models with weak lower bound power-law behavior. (S_n may force $i_{S_n}(1)$ to decay to 0 for $n \to \infty$ because of weak correlations of consecutive tokens very late in the sequence.) The proof of theorem 1.1 fails when defining c, as it depends on t.

Remark 1.9. If S has decaying behavior, we cannot prove that S has strong lower bound power-law behavior by bounding $\overline{i_{S_{t+\tau}}}$ (using $\overline{i_{S_{t+\tau}}} \leq i_{S_{t+\tau}}(\tau)$), as we have for every $\tau \in \mathbb{N}$:

$$0 \leq \overline{i_{S_{t+\tau}}} \leq I_{S_{t+\tau}}(t) \xrightarrow{t\to\infty} 0 \quad .$$

Definition 1.9. A model S is called *large scale time invariant* iff there exists a vector $\boldsymbol{\mu} \in [0,1]^\Sigma$ s.t. for all $a \in \Sigma$ we have

$$P(X_t = a) \xrightarrow{t\to\infty} \boldsymbol{\mu}_a \quad ,$$

where X_t is sampled over any $S_{t+k}, k \in \mathbb{N}$.

2 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 A.

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 B.

2.1 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 2.1 (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 2.1. Based on theorem 1.1, 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.

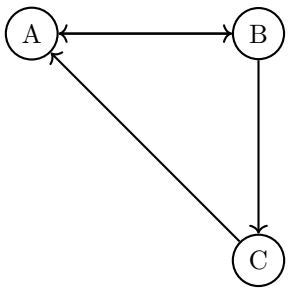

Figure 1: A simple irreducible aperiodic Markov chain. Note that if $X_{t_0} = C$, then we know that $X_{t_0+1} = A$.

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 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 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 2.2. 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 2.3. 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 A.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

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

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.

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 power-law behavior (see definition 1.6 or definition 1.8). 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 2.1 (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 2.2 (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 A.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 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 A.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 $\boldsymbol{v} \in \mathbb{C}^{|S_C|}$ 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 A.3 we can apply the Perron-Frobenius Theorem (with the exception that the eigenvector doesn't have to be positive) to $\boldsymbol{M}^m$ and $\boldsymbol{M}^{m+1}$, and thanks to lemma A.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 $\lim_{\tau \to \infty}(\boldsymbol{M}^\tau)_{ba} = \boldsymbol{\mu}_b$ ($\coloneqq \lim_{\tau \to \infty} P(b)$). But those were the only two assumptions made in section 2.1.1. Thus, by the same logic, we get $I(X_{t_0}; X_{t_0+\tau}) \in \mathcal{O}(|\lambda_2^+|^\tau)$. □

3 No Power-Law in Hidden Markov Models

When modelling *natural language*, we generate the next token(s) based on the previous tokens (and hence not based on future tokens; we generate text from left to right). Thus, when disregarding the future tokens, i.e. when marginalizing over them, we can determine the Markov blanket of the current token, i.e. determine the minimal set of tokens that infuluence the current one.

Naturally, we can think of natural language modelling as *Bayesian networks* over the characters in the text. Because we generate text from left to right, we naturally assume all arrows to go from previous tokens to future tokens (because this is also the modus operandi for token generation). For example, Markov chains up to character position t have the following simple representation:

$$X_0 \longrightarrow X_1 \longrightarrow \cdots \longrightarrow X_t$$

Figure 2: Bayesian network of Markov chains. All arrows go from previous tokens to future tokens.

But really, in Markov chains $P(X_{t+1} = a \mid X_t = b)$ is independent of t and hence is constant over time. So really, all the arrows in figure 2 represent the same transition, this is very important to note.

From a modelling perspective, it is very reasonable to reuse the same transitions over time, as we cannot have infinitely many "hard-coded" transitions (but we can use different models whith implicitely infitite transitions). Furthermore, for the same *mode of transition*, which we define as the "arrow structure" of all ingoing edges into the current node in the Bayesian network (in figure 2 the mode of transition would be from the current token to the next), it seems reasonable to assume invariance in time, i.e. fixed transition probabilities. We call such transitions to be *constant*.

Now, the question is, can we achieve power law decay with only one constant (hard-coded) mode of transition? Well, for Markov chains it did not work, so maybe we just have to augment the context window and create new modes of transition.

This is an interesting approach, which we will investigate on. Since we already established interesting results for Markov chains, we would like to reduce any constant mode of transition to a Markov chain. But how do we do this for a larger context window, where we have many random variables influencing the current one?

The idea is to employ a hidden variable $Y \in \Sigma^s$, where Σ is the alphabet, and s is the size of the context window, which we define as the length of the longest arrow in the mode of transition (for Markov chains $s = 1$). Clearly, Y captures the entire *state* at time t of our model, and we can model the transitions $Y_t \to Y_{t+1}$ as simple Markov chain transitions (and hence independent of time). And, of course, once we know Y_t, we also know X_t (which of course can be modelled with Markov chain transitions as well). Thus, we have the following Bayesian network:

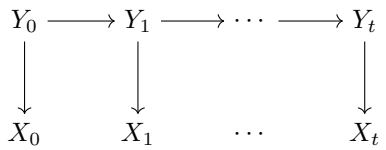

Figure 3: Bayesian network of a hidden Markov model.

These models are known as *hidden Markov models*. Unfortunately, there is no free lunch, as we see in the following theorem:

Theorem 3.1 (No Hidden Markov Model with Power-Law decay)**.** *There is no hidden Markov model* $(\boldsymbol{M_Y}, \boldsymbol{M_X})$ *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. Set $t \coloneqq t_0$. First, note that $P(X_{t+\tau} = b \mid Y_t = c) = (\boldsymbol{M_X}\boldsymbol{M_Y}^\tau)_{bc}$. Furthermore,

$$\begin{aligned} P(Y_t = c \mid X_t = a) &= \frac{P(Y_t = c, X_t = a)}{P(X_t = a)} \\ &= \frac{P(X_t = a \mid Y_t = c)P(Y_t = c)}{P(X_t = a)} \\ &= (\boldsymbol{M_X})_{ac}\frac{P(Y_t = c)}{P(X_t = a)} \quad . \end{aligned}$$

We are interested in $P(X_{t+\tau} = b \mid X_t = a)$. Based on our observations, we have

$$\begin{aligned} P(X_{t+\tau} = b \mid X_t = a) &= \sum_{c \in S_Y} P(Y_t = c \mid X_t = a)P(X_{t+\tau} = b \mid Y_t = c) \\ &= \sum_{c \in S_Y} (\boldsymbol{M_X})_{ac}\frac{P(Y_t = c)}{P(X_t = a)}(\boldsymbol{M_X}\boldsymbol{M_Y}^\tau)_{bc} \\ &= \frac{1}{P(X_t = a)} \sum_{c \in S_Y} (\boldsymbol{M_X})_{ac}P(Y_t = c)(\boldsymbol{M_X}\boldsymbol{M_Y}^\tau)_{bc} \\ &= \frac{1}{P(X_t = a)} \sum_{c \in S_Y} (\boldsymbol{M_X})_{ac}P(Y_t = c) \sum_{d \in S_Y} (\boldsymbol{M_X})_{bd}(\boldsymbol{M_Y}^\tau)_{dc} \quad . \end{aligned}$$

For the case that $\boldsymbol{M}_{\boldsymbol{Y}}^{\tau}$ converges to $\boldsymbol{M}_{\boldsymbol{\mu_Y}}$, it must do so with exponential decay. So we get:

$$
\begin{aligned}
&= \frac{1}{P(X_t = a)} \sum_{c \in S_Y} (\boldsymbol{M_X})_{ac} P(Y_t = c) \sum_{d \in S_Y} (\boldsymbol{M_X})_{bd} ((\boldsymbol{\mu_Y})_d \pm \mathcal{O}(|\lambda_2^+|^\tau)) \\
&= \frac{1}{P(X_t = a)} \sum_{c \in S_Y} (\boldsymbol{M_X})_{ac} P(Y_t = c) \left[(\boldsymbol{M_X \mu_Y})_b \pm \mathcal{O}(|\lambda_2^+|^\tau) \right] \\
&= \frac{1}{P(X_t = a)} \sum_{c \in S_Y} \left[(\boldsymbol{M_X})_{ac} P(Y_t = c) (\boldsymbol{M_X \mu_Y})_b \pm \mathcal{O}(|\lambda_2^+|^\tau) \right] \\
&= \left[\frac{1}{P(X_t = a)} \sum_{c \in S_Y} (\boldsymbol{M_X})_{ac} P(Y_t = c) (\boldsymbol{M_X \mu_Y})_b \right] \pm \mathcal{O}(|\lambda_2^+|^\tau) \\
&= \left[\frac{(\boldsymbol{M_X \mu_Y})_b}{P(X_t = a)} \sum_{c \in S_Y} (\boldsymbol{M_X})_{ac} P(Y_t = c) \right] \pm \mathcal{O}(|\lambda_2^+|^\tau) \\
&= \left[\frac{(\boldsymbol{M_X \mu_Y})_b}{P(X_t = a)} P(X_t = a) \right] \pm \mathcal{O}(|\lambda_2^+|^\tau) \\
&= (\boldsymbol{M_X \mu_Y})_b \pm \mathcal{O}(|\lambda_2^+|^\tau) \quad .
\end{aligned}
$$

Thus, we get:

$$
\begin{aligned}
I_R(X_t, X_{t+\tau}) + 1 &= \sum_{(a,b) \in {S_X}^2} \frac{P(X_t = a)}{P(X_{t+\tau} = b)} P(X_{t+\tau} = b \mid X_t = a)^2 \\
&= \sum_{(a,b) \in {S_X}^2} \frac{P(X_t = a)}{P(X_{t+\tau} = b)} \left[(\boldsymbol{M_X \mu_Y})_b \pm \mathcal{O}(|\lambda_2^+|^\tau) \right]^2 \\
&= \sum_{(a,b) \in {S_X}^2} \frac{P(X_t = a)}{P(X_{t+\tau} = b)} \left[(\boldsymbol{M_X \mu_Y})_b^2 \pm \mathcal{O}(|\lambda_2^+|^\tau) \right] \\
&= \left[\sum_{(a,b) \in {S_X}^2} \frac{P(X_t = a)}{P(X_{t+\tau} = b)} (\boldsymbol{M_X \mu_Y})_b^2 \right] \pm \mathcal{O}(|\lambda_2^+|^\tau) \\
&= \left[\sum_{(a,b) \in {S_X}^2} \frac{P(X_t = a)}{(\boldsymbol{M_X \mu_Y})_b \pm \mathcal{O}(|\lambda_2^+|^{t+\tau})} (\boldsymbol{M_X \mu_Y})_b^2 \right] \pm \mathcal{O}(|\lambda_2^+|^\tau) \\
&= \left[\sum_{(a,b) \in {S_X}^2} P(X_t = a) (\boldsymbol{M_X \mu_Y})_b \pm \mathcal{O}(|\lambda_2^+|^{t+\tau}) \right] \pm \mathcal{O}(|\lambda_2^+|^\tau) \\
&= 1 \pm \mathcal{O}(|\lambda_2^+|^\tau) \quad ,
\end{aligned}
$$

from which $I(X_t, X_{t+\tau}) \in \mathcal{O}(|\lambda_2^+|^\tau)$ follows.

Now assume $\boldsymbol{M}_{\boldsymbol{Y}}^{\tau}$ does not converges to $\boldsymbol{M}_{\boldsymbol{\mu_Y}}$. This case is much harder to prove, and we will only provide an intuition.

Like before, we still must have $\lim_{\tau\to\infty} |P(X_{t+\tau} = b \mid X_t = a) - P(X_{t+\tau} = b)| = 0$ in order for $I(X_t, X_{t+\tau})$ to converge to 0. Again, this means that $P(X_{t+\tau} = b \mid X_t = a)$ should become independent of a. Let's analyze it further:

$$\begin{aligned} P(X_{t+\tau} = b \mid X_t = a) &= \sum_{c\in S_Y} P(Y_t = c \mid X_t = a) P(X_{t+\tau} = b \mid Y_t = c) \\ &= \sum_{c\in S_Y} P(Y_t = c \mid X_t = a)(\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{M}_{\boldsymbol{Y}}^{\tau})_{bc} \quad . \end{aligned}$$

If we assume $P(X_{t+\tau} = b)$ to converge, then our expression must also be independent of t! But the coefficients $P(Y_t = c \mid X_t = a)$ vary a lot with t and a, so $(\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{M}_{\boldsymbol{Y}}^{\tau})_{bc}$ should become independent of c (for every b). But since $\boldsymbol{M}_{\boldsymbol{Y}}^{\tau}$ does not converge to $\boldsymbol{M}_{\boldsymbol{\mu_Y}}$, $\boldsymbol{M}_{\boldsymbol{X}}$ must correct it. But $\boldsymbol{M}_{\boldsymbol{X}}\boldsymbol{M}_{\boldsymbol{Y}}^{\tau}$ must also converge with exponential decay (why?).

The case that $P(X_{t+\tau} = b)$ does not converge is also not easy. But for now we are satisfied with the fact that natural languages should have this property, so we may assume convergence of $P(X_{t+\tau} = b)$. □

3.1 Conclusions for Model Selection

Since we are interested in natural language modelling, we should choose a model with power-law decay in the mutual independence measure. And since a constant mode of transition is not sufficient for this purpose, we should instead look at alternatives.

1. Change Transition Tables over Time. This is a simple approach, but it assumes a prior about the character distribution based on their position, but this non-characteristic of natural language. Instead, we should change the transitions based on what we have seen (or generated) thus far, but this is equivalent of augmenting the context window at each step. Of course, we now must have a dynamic model capable of processing arbitrary large sentences.

2. Augmenting Context Window Dynamically. This is a very natural approach. We can choose whether we want to read in the entire previous text, or maybe every second character, or so on, but the context window should keep growing indefinitely (or else we would have the same mode of transition at two points, and we assume that the same mode of transition stays constant over time, and it would be strange to alternate between finite modes of transition, because this assumes a prior based on the character position again).

Intuitively, we should base our token guess based on the entire previous text, as we humans operate in similar manner. Or, alternatively, the context window should grow really large, until there will only be minor differences, at which point it may stay constant (in theory we might have some exponential decay, but this would only be noticeable over very large distances, where the mutual information naturally already decayed to almost zero).

Theoretically, however, we can always construct counter examples where the mutual information stays relatively high over large distances. Thus, such models may serve only as heuristics—albeit very capable ones.

In the next chapter, we analyze models with dynamically increasing context windows, and we will actually see some power-law behavior.

4 Tensor Networks

Our goal now is to focus on a subclass of models over Σ^*. To this end, we analyze *tensor networks*.

We denote a tensor T_v with k axes of sizes $D_v = \{d_1, \ldots, d_k\}$ as a function

$$T_v : [d_1] \times \cdots \times [d_k] \mapsto \mathbb{R} \quad .$$

As a shorthand, we write

$$[D_v] \coloneqq [d_1] \times \cdots \times [d_k] \quad .$$

Since indexing is usually clear from context, we treat D_v as a multiset of axis sizes.

Given two tensors T_u and T_v that share a common axis of size d_e, their contraction over this axis produces a new tensor T_C with dimension set

$$D_C = (D_u \setminus \{d_e\}) \cup (D_v \setminus \{d_e\}) \quad ,$$

defined as

$$T_C(i) = \sum_{i_e \in [d_e]} T_u(i_{D_u}, i_e) \cdot T_v(i_{D_v}, i_e) \quad ,$$

where $i \in [D_C]$. Note that $d_e \notin D_C$, which is why we explicitly included index i_e in the summation.

Definition 4.1 (Tensor Network over Σ^n)**.** A *tensor network* $\mathcal{T}$ over Σ^n is defined by a graph $G = (V, E)$ with the following structure:

- V is the set of vertices, where each vertex $v = (\text{layer}, \text{index}) \in V$ corresponds to a tensor T_v with axis sizes $D_v = \{d_1, \ldots, d_k\}$. Let $V_{\text{layer}} \subseteq V$ denote the set of all vertices at a given layer.
- The input set $I = (T_{0,1}, \ldots, T_{0,n}) \subset V$ consists of tensors each having a single axis of size $|\Sigma|$. These serve as the one-hot-encoded inputs corresponding to a string $w \in \Sigma^n$.
- $E \subseteq \{\{u, v\} \mid u \in V_l,\, v \in V_{l+1}\}$ is the set of edges. Each edge $e = \{u, v\}$ represents a shared index of size d_e between tensors T_u and T_v, which is summed over during contraction.
- The usual tensor network constrains: For each vertex $v \in V$, the degree $\deg(v)$ must match the number of axes $|D_v|$, and shared indices must correspond to same axis sizes.

Once the input tensors are initialized with one-hot encodings derived from a string $w \in \Sigma^n$, the network computes a scalar output $\mathcal{T}(w)$. This induces a probability distribution over Σ^n defined by:

$$S_{n,\mathcal{T}}(w) \coloneqq \frac{f(\mathcal{T}(w))}{\sum_{w' \in \Sigma^n} f(\mathcal{T}(w'))} \quad ,$$

where $f : \mathbb{R} \mapsto \mathbb{R}_{\geq 0}$ is any arbitrary function like $f \equiv \exp$.

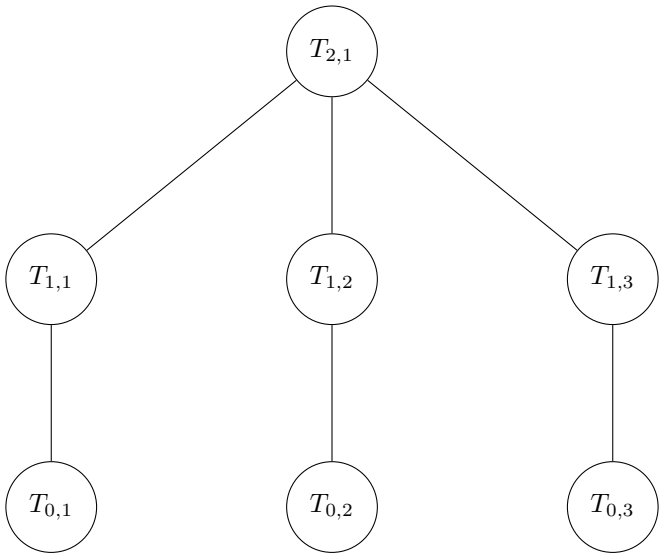

Figure 4: A basic tensor network over Σ^3.

Definition 4.2 (Normalized and Non-Negative Tensor Networks)**.** Let $\mathcal{T}$ be a tensor network over Σ^n with scalar output $\mathcal{T}(w)$ for each $w \in \Sigma^n$. Define the total mass of the network as

$$|\mathcal{T}| \coloneqq \sum_{w \in \Sigma^n} \mathcal{T}(w) \quad .$$

We say $\mathcal{T}$ is *normalized* iff $|\mathcal{T}| = 1$.

Furthermore, a tensor network is said to be *non-negative* iff for all $w \in \Sigma^n$ we have $\mathcal{T}(w) \geq 0$.

Remark 4.1. We can enforce all tensor networks of our model space to be non-negative by only allowing for non-negative tensors in the networks.

Definition 4.3 (Normalization of Tensor Networks)**.** Let $\mathcal{T}$ be a tensor network over Σ^n, and let $H \coloneqq V \setminus I$ be the set of non-input tensors, and define $|H|$ as its cardinality. The *induced normalized tensor network* $\frac{\mathcal{T}}{|\mathcal{T}|}$ is the same network as $\mathcal{T}$, but each entry of each tensor in H is scaled by the factor $\frac{1}{\sqrt[|H|]{|\mathcal{T}|}}$.

Lemma 4.1. *Let $J \subseteq [n]$ and let $\mathcal{T}$ be a tensor network over Σ^n. Define a modified network $\mathcal{T}_J$ where for all $j \in J$, the input tensor $T_{0,j}$ is initialized to the all-ones vector (i.e., $\mathbf{1} \in \mathbb{R}^{|\Sigma|}$). Then for any $w \in \Sigma^{[n]\setminus J}$:*

$$\sum_{w_J \in \Sigma^{|J|}} \mathcal{T}(w_J, w) = \mathcal{T}_J(w) \quad .$$

Proof. We proceed by induction on the size of the subset $J \subseteq [n]$.

Base case: $|J| = 0$. Then $J = \emptyset$, so $\mathcal{T}_J = \mathcal{T}$, and the sum over $w_J \in \Sigma^{|J|}$ is a sum over a singleton (the empty word), yielding:

$$\sum_{w_J \in \Sigma^0} \mathcal{T}(w_J, w) = \mathcal{T}(w),$$

and since $\mathcal{T}_J(w) = \mathcal{T}(w)$, the base case holds.

Inductive step: Assume the lemma holds for all subsets of size k, and let $J \subseteq [n]$ with $|J| = k + 1$. Pick any $j_0 \in J$, and define $J' = J \setminus \{j_0\}$, which has size k. By the inductive hypothesis, for any $w \in \Sigma^{[n]\setminus J}$, we have:

$$\sum_{w_{J'} \in \Sigma^k} \mathcal{T}(w_{J'}, w, w_{j_0}) = \mathcal{T}_{J'}(w, w_{j_0}) \quad ,$$

where $w_{j_0} \in \Sigma$ varies over its values.

Now consider the sum over all $w_J \in \Sigma^{k+1}$, which we write as:

$$\sum_{w_{J'} \in \Sigma^k} \sum_{w_{j_0} \in \Sigma} \mathcal{T}(w_{J'}, w_{j_0}, w) \quad .$$

By the inductive hypothesis, this equals:

$$\sum_{w_{j_0} \in \Sigma} \mathcal{T}_{J'}(w_{j_0}, w) \quad .$$

Observe that in $\mathcal{T}_{J'}$, the input tensor at position j_0 is still initialized to a one-hot vector, while the inputs at J' have been replaced with the all-ones vector.

Now, note that the inner sum over w_{j_0} is equivalent to replacing the input at j_0 with the all-ones vector, since the sum represents a sum over vector dot products of vector $\boldsymbol{v}_{w_{j_0}} \coloneqq \mathcal{T}_{J'}(w_{j_0}, w)$ with one-hot encoded vectors. It follows from linearity that we can factor out $\boldsymbol{v}$, and the sum of the one-hot encoded vector yields the all-ones vector. Thus:

$$\sum_{w_{j_0} \in \Sigma} \mathcal{T}_{J'}(w_{j_0}, w) = \mathcal{T}_J(w) \quad .$$

Hence, by induction, the lemma holds for all subsets $J \subseteq [n]$.

□

Corollary 4.1. *Let $\mathcal{T}$ be a tensor network over Σ^n, and let $\mathcal{T}_{[n]}$ be the network where all input tensors are initialized to the all-ones vector. Then:*

$$\mathcal{T} \text{ is normalized} \iff \mathcal{T}_{[n]} = 1 \quad ,$$

i.e., the total contraction of the network with all-one input tensors equals 1.

Lemma 4.2. *Let $\mathcal{T}$ be a tensor network over Σ^n. The induced normalized tensor network $\frac{\mathcal{T}}{|\mathcal{T}|}$ is indeed normalized, and if additionally $\mathcal{T}$ is non-negative and $f \equiv id$, we have for all $w \in \Sigma^n$:*

$$S_{n,\mathcal{T}}(w) = S_{n,\frac{\mathcal{T}}{|\mathcal{T}|}}(w) \quad .$$

Proof. Let H be the set of non-input tensors in $\mathcal{T}$, and let $|H| = m$. In the induced normalized network, every tensor in H is scaled by a factor $\alpha = \frac{1}{\sqrt[m]{|\mathcal{T}|}}$. Since the final output $\mathcal{T}(w)$ is a multilinear contraction over the tensors, this means the scalar output for any $w \in \Sigma^n$ becomes:

$$\left(\prod_{v \in H} \alpha \right) \cdot \mathcal{T}(w) = \alpha^m \cdot \mathcal{T}(w) = \frac{1}{|\mathcal{T}|} \cdot \mathcal{T}(w) \quad .$$

Hence,

$$\left(\frac{\mathcal{T}}{|\mathcal{T}|} \right)(w) = \frac{\mathcal{T}(w)}{|\mathcal{T}|} \quad .$$

Summing over all $w \in \Sigma^n$,

$$\left| \frac{\mathcal{T}}{|\mathcal{T}|} \right| = \sum_{w \in \Sigma^n} \frac{\mathcal{T}(w)}{|\mathcal{T}|} = \frac{1}{|\mathcal{T}|} \sum_{w \in \Sigma^n} \mathcal{T}(w) = \frac{|\mathcal{T}|}{|\mathcal{T}|} = 1 \quad .$$

Moreover, since the normalization rescales all outputs by the same constant, the ratio of the terms to the total sum remains unchanged:

$$S_{n,\frac{\mathcal{T}}{|\mathcal{T}|}}(w) = \frac{\left(\frac{\mathcal{T}(w)}{|\mathcal{T}|}\right)}{\sum_{w' \in \Sigma^n}\left(\frac{\mathcal{T}(w')}{|\mathcal{T}|}\right)} = \frac{\mathcal{T}(w)}{|\mathcal{T}|} \cdot \frac{1}{1} = S_{n,\mathcal{T}}(w) \quad .$$

This completes the proof. □

One might ask whether our definition for tensor networks is bit restrictive, as it only allows for contraction over *pairs* of tensors. But what if we wanted to contract, say, three tensors at once over a common index?

Proposition 4.1. *Let $V' \subseteq V$ be a set of tensors in a tensor network, each containing an axis of dimension d labeled by a shared index i. Contracting all tensors in V' over the shared index i is equivalent to contracting each tensor individually with a single tensor*

$$\delta_{|V'|} : [d]^{|V'|} \mapsto \mathbb{R}_{\geq 0}$$

defined by

$$\delta_{|V'|}(i_1, \ldots, i_{|V'|}) = \begin{cases} 1 & \text{if } i_1 = \cdots = i_{|V'|} \ , \\ 0 & \text{otherwise.} \end{cases} \quad .$$

That is, a full contraction over a shared index can be implemented by introducing a single copy tensor connected to each tensor in V'.

Proof. Each tensor T_v for $v \in V'$ has an index $i \in [d]$ corresponding to the shared axis. The contraction over this index is defined by summing over the common value of i across all tensors:

$$\sum_{i=1}^{d} \prod_{v \in V'} T_v(\ldots, i, \ldots) \quad .$$

Now consider a new tensor $\delta_{|V'|}$ of order $|V'|$, defined as 1 if all indices are equal and 0 otherwise. Let each tensor T_v maintain its original indices, but connect to $\delta_{|V'|}$ via the position corresponding to v.

The contraction over this shared structure gives:

$$\sum_{i_1, \ldots, i_{|V'|}} \left(\prod_{v \in V'} T_v(\ldots, i_v, \ldots) \right) \delta_{|V'|}(i_1, \ldots, i_{|V'|}) \quad .$$

By definition of $\delta_{|V'|}$, this enforces $i_1 = \cdots = i_{|V'|}$, reducing the above to:

$$\sum_{i=1}^{d} \prod_{v \in V'} T_v(\ldots, i, \ldots) \quad ,$$

which is exactly the original contraction. Hence, the two constructions are equivalent. □

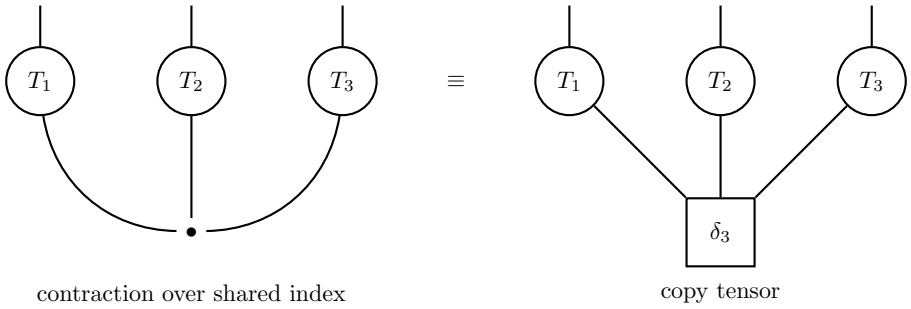

Figure 5: Contracting multiple tensors over one shared index is equivalent to contracting them individually with a single copy tensor.

5 Binary Tree Tensor Networks

We are interested in the model space of binary tree tensor networks as shown in figure 6. We assume that every network is non-negative and normalized, and we set $f \equiv \mathrm{id}$.

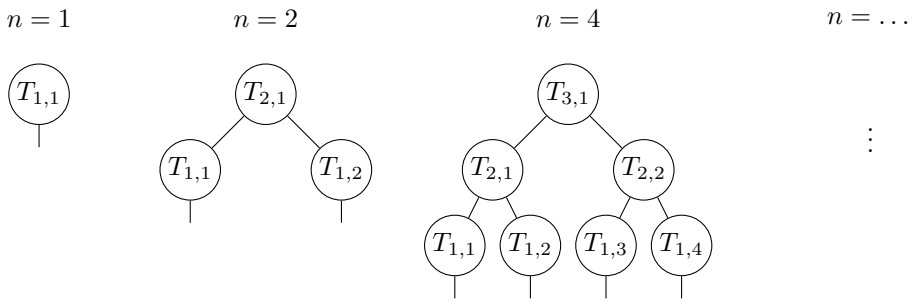

Figure 6: Binary tree model space for sequences of length $n = 2^k$.

5.1 Bulk Marginal Property

In definition 1.3 we saw how to construct a model with the desired bulk marginal property based on the base model. However, we might not always have a base model for every $n \in \mathbb{N}$ like discussed. Luckily, it turns out that this is not an issue, as there are many ways we can build a new model with the bulk marginal property from a base model even if it is only defined on a subset of $\mathbb{N}$. Without a proof, we might do the same procedure as in definition 1.3 but with bigger steps (instead of taking always the consecutive model), and induce the in-between models by marginalizing the bigger ones.

Alternatively, if we wanted a model with bulk marginal property that itself is also an element of our specified model space, we might ask ourselves, how we can construct a bigger tensor network while preserving the distribution in its leading random variables.

Let's analyze the following example: Say we wanted to integrate the tree tensor network for $n = 2$ in figure 6 into a bigger tree tensor network with $n = 4$. Note that by assumptions the tensor networks are non-negative and normalized, and $f \equiv \mathrm{id}$. Thus, in order for our new tensor network to have the bulk marginal property, contracting the smaller network must be equivalent to contracting the bigger one, where the new nodes (in this case $T'_{1,3}$ and $T'_{1,4}$) are contracted with all-ones vectors, see figure 7.

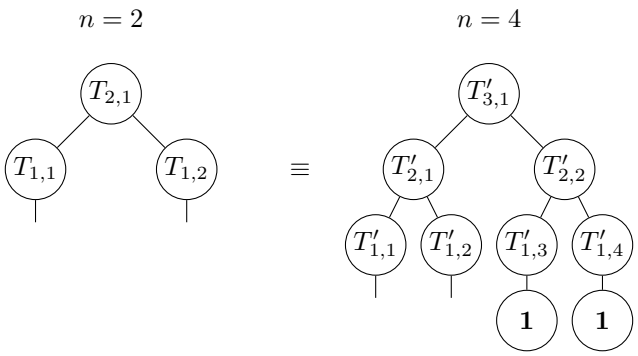

Figure 7: Bulk marginal property enforces the equivalence of these models.

Note that if we indeed had equivalence of these models, this would imply that the bigger model is now normalized as well based on lemma 4.1.

Let's now analyze the vector $\boldsymbol{v}$ depicted in figure 8.

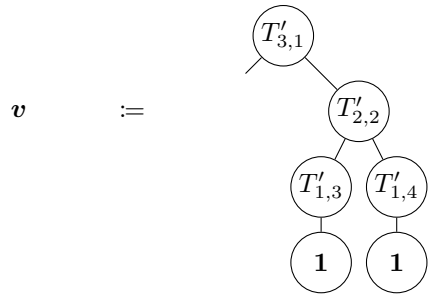

Figure 8: Contracting this sub-network with all-ones vectors yields vector $\boldsymbol{v}$.

What can we say about $\boldsymbol{v}$? Well, we can assume that it has at least one non-zero entry. This is always possible, and if $\boldsymbol{v}$ consisted of only zeros, then the network wouldn't be normalized.

But now we can initialize the remaining tensors: The leaf tensors $T'_{1,1}$ and $T'_{2,1}$ can be taking over from the smaller model, and the new tensor $T'_{2,1}$, which is now a vector of matrices like the old $T_{2,1}$, can initialized as the matrix $T_{2,1}$ divided by the non-zero entry of $\boldsymbol{v}$ at the same position. All other matrices in this vector will just be set to zero-matrices.

It is apparent that this initialization ensures the equivalence of the models as depicted in figure 7. It is also clear that this method works when transitioning from any $n = 2^k$ to $n' = 2^{k+1}$. Note that we also didn't increase the sizes of the tensors (except for $T'_{2,1}$, as it got another axis). Furthermore, when assuming all tensor entries are non-negative, then the new tensor network $\mathcal{T}'$ is also non-negative. This leads to the following observation:

Corollary 5.1. *Let $\mathcal{T}$ be a binary tree tensor network over $\Sigma^n, n = 2^k$. Then, there exists a binary tree tensor network $\mathcal{T}'$ over $\Sigma^{n'}, n' = 2^{k+1}$ s.t. the transition from $\mathcal{T}$ to $\mathcal{T}'$ complies with the bulk marginal property. Furthermore, $\mathcal{T}'$ complies with axes-sizes constrains (under the assumption of a maximum axis-size constrain which is increasing in n).*

5.2 Binary Tree Tensor Networks are Universal Approximators

Now, we want to analyze the properties of these binary tree tensor networks further. It may not bother us how we construct increasingly bigger models that satisfy the bulk marginal property, we know that the model space of binary tree tensor networks is capable of producing such families.

One question we might ask is whether such a model space restricts the space of possible probability distributions, and if so by how much. As it turns out, in the most general case when allowing very large tensors in the networks, we can model *any* probability distribution:

Proposition 5.1. *Given any probability distribution $p : \Sigma^{2^k} \mapsto [0, 1]$, we can always construct a binary tree tensor network $\mathcal{T}$ over Σ^{2^k} s.t. $p \equiv S_{2^k,\mathcal{T}}$. (Where $\mathcal{T}$ has the properties mentioned in the beginning and has no constrains on the tensor sizes.)*

Proof. For clarity reasons, we only show how to construct $\mathcal{T}$ for $n = 2^k = 4$. The procedure can easily be extended to the general case.

Our model structure is depicted in figure 9.

Now, we initialize the leaf matrices as identity matrices $\delta_2 \in \mathbb{R}^{|\Sigma|\times|\Sigma|}$. Thus, when contracting a leaf tensor with a one-hot encoded input vector at position i, we get the vector $\boldsymbol{v}^{(i)}$ with $\boldsymbol{v}^{(i)}_j = \mathbf{1}[X_i = c_j], c_j \in \Sigma$.

Now, the tensors in layer two are of the following form:

$$T_{2,j} : |\Sigma| \times |\Sigma| \mapsto \mathbb{R}^{|\Sigma|^2} \quad .$$

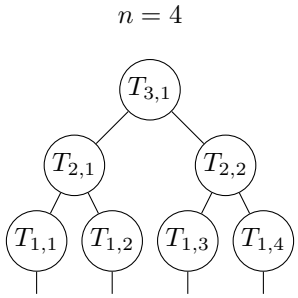

Figure 9: Model structure of binary tree tensor networks for $n = 2^k = 4$.

The outgoing axis may be index by $(X'_i, X'_{i+1}) \in \Sigma^2$. The map is then defined by

$$T_{2,j}(X_i, X_{i+1}) = \mathbf{1}[(X'_i, X'_{i+1}) = (X_i, X_{i+1})] \quad ,$$

i.e. $T_{2,j}$ is a three dimensional tensor with $|\Sigma| \times |\Sigma|$ many vectors of size $|\Sigma|^2$ which are one-hot encoded vectors of 2-tuples of Σ^2.

Finally, $T_{3,1}$ stores the entire probability distribution:

$$T_{3,1} : |\Sigma|^2 \times |\Sigma|^2 \mapsto [0,1], ((X'_1, X'_2), (X'_3, X'_4)) \mapsto p(X'_1, X'_2, X'_3, X'_4) \quad .$$

Thus, based on the construction we see that upon contracting the network with an initialization defined by $w \in \Sigma^4$, we get $S_{4,\mathcal{T}}(w) = p(w)$ as desired.

Note that this construction can easily be extended to arbitrary $n = 2^k$. □

As one might expect, we see that our general construction needs $\Omega(|\Sigma|^n)$ many parameters, as the root tensor stores all the $|\Sigma|^n$ many probabilities for $w \in \Sigma^n$.

5.3 Restricting Parameters

One natural question is what happens if we restrict the number of parameters. Obviously, if we don't have exponentially many parameters with respect to the word length, we won't be able to construct *every* probability distribution.

However, when modelling natural language for example, we really aren't interested in the most general case of probability distributions. For example, for a fixed word length n, we might want behavior similar to large scale time invariance (see definition 1.9). Note that for a fixed n, the constrain of the bulk marginal property has no restrictive effect on the possible distributions.

Most decisively, we are interested in models capable of power-law behavior. Our goal is to show that binary tree tensor networks are incapable of this when we cap the number of parameters (i.e. the tensor sizes).

To formalize this, we first specify what it means to cap the parameters. There are two approaches that come to my mind: Either cap the total number of parameters (the entries of *all* tensors), or cap the axes-sizes and hence the size of each tensor individually. Of course, the latter approach implies that the total number of parameters are capped by $2n$ times the maximal number of parameters per tensor, as there are $2n-1$ many tensors in the network. Thus, it seems natural to cap the individual tensor sizes to ensure a *good* distribution of parameters over the tensors (which means that there shouldn't be one very large tensor and many small tensors). We might do this by capping the axes-sizes, as this allows the tensors to be more "cube-like" and to not have one big axis and two smaller ones for example (note that most tensors have three axes).

Thus, let us assume we cap the axes-sizes. We could define an upper bound for every tensor individually in the network, or, for every layer, but for simplicity's sake we define an upper bound on the axes-sizes in the entire network. As it turns out, it also doesn't really matter which approach we choose when arguing with complexity bounds.

So, we want to cap the axes-sizes. To this end, let d denote the biggest axis-size allowed. Now, there are multiple options again: First, d could stay constant for every network of size $n = 2^k$. In this case, the parameters grow linearly in n (since the number of tensors grows linearly). This, however, is probably too restrictive. The second approach is to let d grow with n. Of course, if $d \coloneqq |\Sigma|^n$ (or even $d \coloneqq |\Sigma|^{\frac{n}{2}} = \left(\sqrt{|\Sigma|}\right)^n$), we won't have any restrictions and way too many parameters. Alternatively, we could try to find a smaller base b for $d = b^n$. Another approach is to define $d(n) \in \mathcal{O}(n^p)$ for some $p \in \mathbb{N}$. This means that axes-sizes grow polynomially with respect to the word length n. Of course, this implies that every tensor grows polynomial in n with $\mathcal{O}((n^p)^3) = \mathcal{O}(n^{3p})$, and hence the parameter complexity of the entire network grows with $\mathcal{O}(n^{3p+1})$.

We see that polynomially growing parameters of the entire network with respect to n es equivalent to polynomially growing axes-sizes. Good. Let us now turn to the power-law constrain.

As we now know, there are different definitions for power-law behavior. The strongest proof would include to disprove weak power-law behavior (see definition 1.8), as this would also disprove strong power law behavior (see definition 1.6, proposition 1.2). Alternatively, under the assumption that our model complies with the bulk marginal property, we can use the contraposition of theorem 1.1 to show the weaker fact that there is no model satisfying the bulk marginal property and having weak power-law behavior (and hence also strong power-law behavior).

Let us look at the latter approach of employing the contraposition of theorem 1.1. What we would need to show is that there exists a character at position t, and no matter what tensor network we apply, we can bound the mutual information by $I(X_t, X_{t+\tau}) \in \mathcal{O}(e^{-\lambda\tau})$ for some fixed $\lambda \in \mathbb{R}_{>0}$, as this of course implies that there is no $\alpha \in \mathbb{R}_{>0}$ s.t. $I(X_t, X_{t+\tau}) \in \Omega(\tau^{-\alpha})$. Note that there obviously exist families satisfying the bulk marginal property. Thus, a potential proof of this claim has meaning concerning power-law behavior of binary tree tensor networks.

Theorem 5.1 (No Power-Law in BTTN with BMP for Polynomially Capped Parameters)**.** *Let $d(n)$ denote the maximum axis-size in a binary tree tensor network over Σ^n. If $d(n) \in \mathcal{O}(n^p)$ for some $p \in \mathbb{N}$, then there exists no family of binary tree tensor networks $(\mathcal{T}_{2^k})_{k\in\mathbb{N}}$ with associated model $S_n(w)$ s.t. S complies with the bulk marginal property and has weak power-law behavior.*

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

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

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

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

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

Note that each column sums to 1.

Remark A.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 A.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 $n \geq 1$. □

Lemma A.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.

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

□

A.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 A.1 can be visualized as in figure 10.

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.

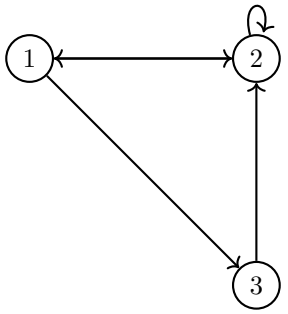

Figure 10: Graph representation of the Markov chain defined in example A.1.

A.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 11.

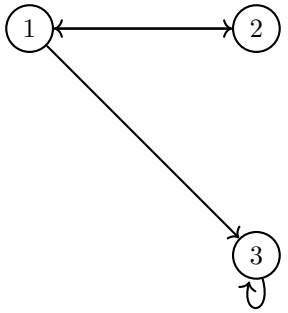

Figure 11: Graph of a reducible Markov chain. Note that once the chain transitions from state 1 to state 3, it will stay at state 3 indefinitely.

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

Definition A.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 A.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 A.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 A.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 A.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 A.2. The partition of figure 10 is $\{\{1, 2, 3\}\}$ and of figure 11 we have $\{\{1, 2\}, \{3\}\}$.

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

Definition A.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 A.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 A.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 11.

Theorem A.2 (Existence of Closed Communication Class)**.** *There is always at least one closed communication class.*

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

A.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:

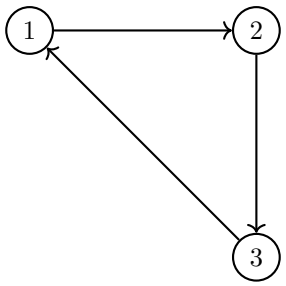

Figure 12: Graph of a periodic Markov chain. Note that once once the starting position is determined, then we also know the state after t steps.

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 A.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 A.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 A.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. □

This result leads to the following definitions:

Definition A.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.

A.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 12 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:

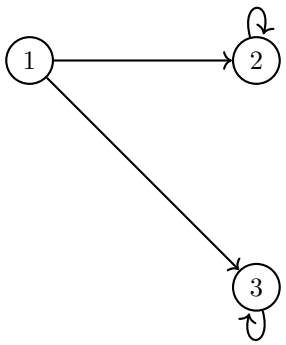

Figure 13: Graph of a reducible Markov chain with two closed communication classes. Note that we might end up stuck at either state 2 or state 3.

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 A.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 A.4. *Let* $i \in S$ *be an aperiodic state. Then there exists an* $L \in \mathbb{N}$ *s.t. for all* $n > L : (\boldsymbol{P}^n)_{ii} > 0$*.*

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

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

□

Remark A.5. The converse of lemma A.4 is also true for obvious reasons.

Proof of Theorem A.3. Let L' be defined as the maximum of all L defined like in Lemma A.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. □

Corollary A.1. *Once $\boldsymbol{P}^m > \boldsymbol{0}$, then for all $\tau \in \mathbb{N} : \boldsymbol{P}^{m+\tau} > \boldsymbol{0}$.*

Corollary A.2. *The Converse of Theorem A.3 is also true, that is if $\boldsymbol{P}^m > \boldsymbol{0}$ for some $m \in \mathbb{N}$, then P is irreducible and aperiodic. Irreducibility directly follows from $\boldsymbol{P}^m > \boldsymbol{0}$, and since $\boldsymbol{P}^{m+1} > \boldsymbol{0}$ as well, $\boldsymbol{P}$ must also be aperiodic.*

The reader might question the importance of Theorem A.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.

A.3 Perron-Frobenius Theorem

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

Theorem A.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 $T(\boldsymbol{v}) \coloneqq \boldsymbol{P}\boldsymbol{v}$. 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 $\boldsymbol{P}\boldsymbol{w} = \boldsymbol{P}c|\boldsymbol{w}| = c\boldsymbol{P}|\boldsymbol{w}| \stackrel{!}{=} \mu\boldsymbol{w}$. 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 $k \coloneqq \frac{\|\Delta_{\boldsymbol{P}}\|_1}{2}$, 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 A.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 doesn't have to be positive anymore, since that was the second application of $\boldsymbol{P} > \boldsymbol{0}$.*

Corollary A.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 $\boldsymbol{v} \coloneqq \boldsymbol{e_i}$. 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 A.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 $\lambda_1, \dots, \lambda_n$ 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 $\lambda_1^{(m)}, \dots, \lambda_n^{(m)}$ s.t. $\lambda_1^{(m)} > |\lambda_2^{(m)}| \geq \cdots \geq |\lambda_n^{(m)}|$. Now, assume that we have ordered $\lambda_1, \dots, \lambda_n$ 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 A.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 $\boldsymbol{v}^* \coloneqq \boldsymbol{v}^{(m)}$ itself is positive, which was the last claim (3). □

Corollary A.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 A.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 A.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

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

B Information Theory

B.1 Entropy

Definition B.1 (Entropy)**.** Let X be a discrete random variable taking values in a finite set $\mathcal{X}$ with probability mass function $p(x) = P(X = x)$. The *entropy* of X, denoted $H(X)$, is defined as:

$$H(X) \coloneqq -\sum_{x \in \mathcal{X}} p(x) \log p(x),$$

where the logarithm is typically taken base 2 (bits) or base e (nats).

Remark B.1. If $p(x) = 0$, we set $p(x) \log p(x) \coloneqq 0$. This ensures that $p(x) \log p(x)$ is continuous on $[0, 1]$.

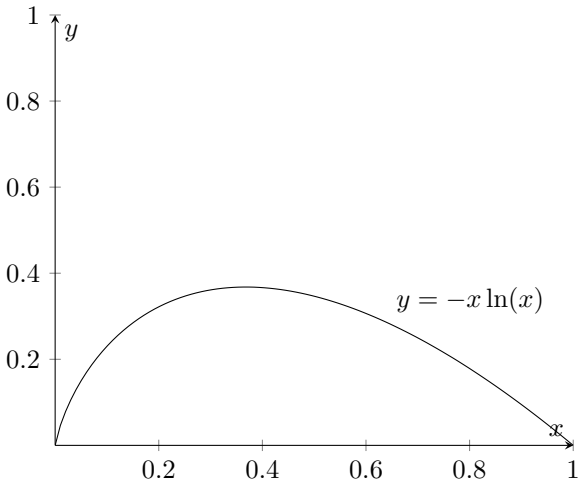

Figure 14: Plot of the function $y = -x \ln(x)$.

Remark B.2. Entropy measures the uncertainty or information content of a random variable. Higher entropy indicates more unpredictability.

Proposition B.1 (Non-Negativity of Entropy)**.** *For any discrete random variable X, we have $H(X) \geq 0$.*

Proof. Since $0 \leq p(x) \leq 1$ and $-\log p(x) \geq 0$, each term in the sum is non-negative, so their total sum is non-negative. □

Lemma B.1 (Jensen's Inequality)**.** *Let $X \in \mathcal{X}$ be a random variable over a finite set $\mathcal{X}$, and let ϕ be a convex function defined for all X. Then:*

$$\phi(E[X]) \leq E[\phi(X)] \quad .$$

Proof. We use induction over $n = |\mathcal{X}|$. The base case $n = 1$ is trivial. Hence, assume that the claim holds for some n. We now prove the claim for $n + 1$. Clearly, for $n > 1$, we must have $P(X = x_k) < 1$ for some $x_k \in \mathcal{X}$. Without loss of generality, we assume $k = n + 1$. Hence:

$$\begin{aligned}
\phi(E[X]) &= \phi\left(\sum_{i=1}^{n+1} p(x_i)x_i\right) \\
&= \phi\left(\left[(1 - p(x_{n+1}))\sum_{i=1}^{n} \frac{p(x_i)}{1 - p(x_{n+1})}x_i\right] + p(x_{n+1})x_{n+1}\right) \\
&\underset{\text{convexity}}{\leq} (1 - p(x_{n+1}))\phi\left(\sum_{i=1}^{n} \frac{p(x_i)}{1 - p(x_{n+1})}x_i\right) + p(x_{n+1})\phi(x_{n+1}) \\
&\underset{\text{inductive hypothesis}}{\leq} (1 - p(x_{n+1}))\sum_{i=1}^{n} \frac{p(x_i)}{1 - p(x_{n+1})}\phi(x_i) + p(x_{n+1})\phi(x_{n+1}) \\
&= \sum_{i=1}^{n+1} p(x_i)\phi(x_i) = E[\phi(X)] \quad .
\end{aligned}$$

□

Remark B.3. For strictly convex ϕ, it can be shown that

$$\phi(E[X]) = E[\phi(X)] \text{ is maximized } \iff X \text{ is sampled from a uniform distribution} \quad .$$

Proposition B.2 (Maximum Entropy)**.** *For a discrete random variable X over n outcomes, entropy is maximized when X is uniform:*

$$H(X) \leq \log n \quad .$$

Proof. We have:

$$\begin{aligned} -H(X) &= -E[-\log(p(X))] \\ &= E\left[-\log\left(\frac{1}{p(X)}\right)\right] \\ &\underset{\text{Jensen's Inequality}}{\geq} -\log\left(E\left[\frac{1}{p(X)}\right]\right) \\ &= -\log n \quad , \end{aligned}$$

where we assumed $p(X) > 0$. Of course, the cases where $p(X) = 0$ follow directly, since $p(X)\log p(X) = 0$.

$H(X) \leq \log n$ follows directly. Note that we have equality iff X has uniform distribution (since $-\log(x)$ is strictly convex). □

B.1.1 Joint, Conditional, and Cross Entropy

Definition B.2 (Joint Entropy)**.** For a pair of discrete random variables X and Y, the joint entropy is:

$$H(X,Y) \coloneqq -\sum_{x,y} p(x,y)\log p(x,y) \quad .$$

Definition B.3 (Conditional Entropy)**.** The conditional entropy of Y given X is defined as:

$$H(Y \mid X) \coloneqq \sum_{x} p(x)H(Y \mid X = x) = -\sum_{x,y} p(x,y)\log p(y \mid x).$$

Corollary B.1. *We immediately see from the first equation that $H(Y \mid X) \geq 0$.*

Theorem B.1 (Chain Rule for Entropy)**.**

$$H(X,Y) = H(X) + H(Y \mid X) \quad .$$

Proof. We have:

$$\begin{aligned} H(X,Y) &= -\sum_{x,y} p(x,y) \log p(x,y) \\ &= -\sum_{x,y} p(x,y) \log \left(p(x)p(y \mid x)\right) \\ &= -\sum_{x,y} p(x,y) \log p(x) - \sum_{x,y} p(x,y) \log p(y \mid x) \\ &= H(X) + H(Y \mid X) \quad . \end{aligned}$$

□

Corollary B.2. $H(X,Y) \geq 0$ *follows directly.*

Definition B.4 (Cross-Entropy)**.** Let p and q be two probability distributions over a finite set $\mathcal{X}$, with $p(x) > 0 \Rightarrow q(x) > 0$. The *cross-entropy* of p relative to q is defined as:

$$H_q(p) \coloneqq -\sum_{x \in \mathcal{X}} p(x) \log q(x) \quad .$$

Remark B.4. Cross-entropy measures the expected number of bits required to encode samples from p using a code optimized for the distribution q.

Remark B.5. Cross-entropy is non-negative (see section B.2).

B.1.2 Properties of Entropy

Proposition B.3. *Conditional entropy satisfies:*

$$H(Y \mid X) \leq H(Y) \quad ,$$

with equality if and only if X and Y are independent.

Proof. From the chain rule:

$$H(X,Y) = H(Y) + H(X \mid Y) = H(X) + H(Y \mid X) \quad ,$$

which implies:

$$H(Y \mid X) = H(Y) + H(X \mid Y) - H(X) = H(Y) - I(X;Y) \quad ,$$

with mutual information $I(X;Y) \geq 0$ (see section B.3). Equality holds if and only if $I(X;Y) = 0$, i.e., X and Y are independent. □

Corollary B.3 (Subadditivity of Entropy)**.** *For any two random variables X and Y,*

$$H(X,Y) \leq H(X) + H(Y) \quad ,$$

with equality if and only if X and Y are independent.

Proof. From the chain rule:

$$H(X,Y) = H(X) + H(Y \mid X) \leq H(X) + H(Y) \quad ,$$

since $H(Y \mid X) \leq H(Y)$ based on proposition B.3. Equality holds if and only if $H(Y \mid X) = H(Y)$, i.e., X and Y are independent. □

Theorem B.2 (Concavity of Entropy)**.** *The entropy function $H(p)$, where $p \in \Delta$ is a probability vector, is concave on the probability simplex Δ.*

Proof. This follows from the fact that $f(x) = -x \log x$ is concave for $x \in [0,1]$, and entropy is the sum of such terms. Therefore, for every convex combination $p = \lambda p_1 + (1-\lambda)p_2$:

$$H(p) \geq \lambda H(p_1) + (1-\lambda)H(p_2) \quad .$$

□

Summary of Key Properties

- Non-negativity: $H(X) \geq 0$
- Maximum entropy: $H(X) \leq \log |\mathcal{X}|$
- Chain rule: $H(X,Y) = H(X) + H(Y \mid X)$
- Subadditivity: $H(X,Y) \leq H(X) + H(Y)$
- Conditioning reduces entropy: $H(Y \mid X) \leq H(Y)$
- Concavity: $H(p)$ is concave in the distribution p

B.2 Kullback-Leibler Divergence

Definition B.5 (KL Divergence)**.** Let P and Q be two discrete probability distributions over the same finite set $\mathcal{X}$, with $P(x) > 0 \Rightarrow Q(x) > 0$. The Kullback-Leibler divergence (or relative entropy) from P to Q is defined as:

$$\begin{aligned} D_{\mathrm{KL}}(P\|Q) &\coloneqq \sum_{x\in\mathcal{X}} P(x)\log\frac{P(x)}{Q(x)} \\ &= -\sum_{x} P(x)\log Q(x) + \sum_{x} P(x)\log P(x) \\ &= H_Q(P) - H(P) \quad . \end{aligned}$$

Remark B.6. If $P(x) = Q(x) = 0$, we set $P(x)\log\frac{P(x)}{Q(x)} \coloneqq 0$.

Remark B.7. KL divergence measures the inefficiency of assuming that the distribution is Q when the true distribution is P. It is not a metric: it is not symmetric and does not satisfy the triangle inequality.

Lemma B.2 (Gibb's Inequality)**.** *Suppose that $P = \{p_1, \ldots, p_n\}$ and $Q = \{q_1, \ldots, q_n\}$ are discrete probability distributions. Then:*

$$-\sum_{i=1}^{n} p_i \log p_i \leq -\sum_{i=1}^{n} p_i \log q_i \quad .$$

Proof. The claim is equivalent to $\sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} p_i \log q_i \geq 0$. We have:

$$\begin{aligned} \sum_{i=1}^{n} p_i \log p_i - \sum_{i=1}^{n} p_i \log q_i &= \sum_{i=1}^{n} p_i \log\frac{p_i}{q_i} \\ &= \sum_{i=1}^{n} p_i \left(-\log\frac{q_i}{p_i}\right) \\ &\underset{\text{Jensen's Inequality}}{\geq} -\log\left(\sum_{i=1}^{n} p_i \frac{q_i}{p_i}\right) \\ &= -\log(1) = 0 \quad . \end{aligned}$$

□

Corollary B.4. *It directly follows from the proof that $D_{\mathrm{KL}}(P\|Q) \geq 0$ and $0 \leq H(P) \leq H_Q(P)$.*

Proposition B.4 (Additivity)**.** *Let $P = P_1 \times P_2$, $Q = Q_1 \times Q_2$. Then:*

$$D_{\mathrm{KL}}(P\|Q) = D_{\mathrm{KL}}(P_1\|Q_1) + D_{\mathrm{KL}}(P_2\|Q_2) \quad .$$

Proof.

$$\begin{aligned} D_{\mathrm{KL}}(P_1 \times P_2\|Q_1 \times Q_2) &= \sum_{x,y} P_1(x)P_2(y) \log \frac{P_1(x)P_2(y)}{Q_1(x)Q_2(y)} \\ &= \sum_{x,y} P_1(x)P_2(y) \left(\log \frac{P_1(x)}{Q_1(x)} + \log \frac{P_2(y)}{Q_2(y)} \right) \\ &= \sum_{x} P_1(x) \log \frac{P_1(x)}{Q_1(x)} + \sum_{y} P_2(y) \log \frac{P_2(y)}{Q_2(y)} \\ &= D_{\mathrm{KL}}(P_1\|Q_1) + D_{\mathrm{KL}}(P_2\|Q_2) \quad . \end{aligned}$$

□

Proposition B.5 (Entropy Representation via KL Divergence)**.** *Let U be the uniform distribution over $\mathcal{X}$, where $|\mathcal{X}| = n$. Then for any distribution P,*

$$H(P) = \log n - D_{\mathrm{KL}}(P\|U) \quad .$$

Proof.

$$\begin{aligned} D_{\mathrm{KL}}(P\|U) &= \sum_{x} P(x) \log \frac{P(x)}{1/n} = \sum_{x} P(x) \log P(x) + \sum_{x} P(x) \log n \\ &= -H(P) + \log n \quad . \end{aligned}$$

□

Summary of Key Properties

- $D_{\mathrm{KL}}(P\|Q) \geq 0$
- $D_{\mathrm{KL}}(P\|Q) = 0 \iff P = Q$
- Asymmetric: $D_{\mathrm{KL}}(P\|Q) \neq D_{\mathrm{KL}}(Q\|P)$
- Additive over independent distributions
- Connection to entropy: $H(P) = \log n - D_{\mathrm{KL}}(P\|U)$

B.3 Mutual Information

Definition B.6 (Mutual Information)**.** Let X and Y be discrete random variables with joint distribution $p(x, y)$ and marginals $p(x)$, $p(y)$. The *mutual information* between X and Y is defined as:

$$I(X;Y) \coloneqq \sum_{x,y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \quad .$$

Remark B.8. Mutual information quantifies how much knowing X reduces uncertainty about Y, and vice versa. Per definition, it is symmetric: $I(X;Y) = I(Y;X)$.

Proposition B.6 (Equivalent Expressions)**.** *Mutual information can also be expressed as:*

$$\begin{aligned} I(X;Y) &= D_{\mathrm{KL}}(p(x,y) \,\|\, p(x)p(y)) \\ &= H_{p(x)p(y)}(p(x,y)) - H(X,Y) \\ &= \left[-\sum_{x,y} p(x,y) \log(p(x)p(y)) \right] - H(X,Y) \\ &= H(X) + H(Y) - H(X,Y) \\ &= H(X) - H(X \mid Y) \\ &= H(Y) - H(Y \mid X) \end{aligned}$$

Proof. Each follows from basic entropy identities and the definition of KL divergence. □

Corollary B.5. *$I(X;Y) \geq 0$, since $I(X;Y) = D_{\mathrm{KL}}(p(x,y) \| p(x)p(y))$ and KL divergence is always non-negative.*

Definition B.7 (Conditional Mutual Information)**.** Let X, Y, Z be discrete random variables. The *conditional mutual information* of X and Y given Z is defined as:

$$I(X;Y \mid Z) \coloneqq \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z)p(y \mid z)} \quad .$$

Equivalently, in terms of entropy:

$$I(X;Y \mid Z) = H(X \mid Z) - H(X \mid (Y,Z)) \quad .$$

Proof.

$$
\begin{aligned}
&H(X \mid Z) - H(X \mid (Y,Z)) \\
&= \sum_z p(z) H(X \mid Z=z) - \sum_{y,z} p(y,z) H(X \mid Y=y, Z=z) \\
&= -\sum_z p(z) \sum_x p(x \mid z) \log p(x \mid z) \;+\; \sum_{y,z} p(y,z) \sum_x p(x \mid y,z) \log p(x \mid y,z) \\
&= \sum_{x,y,z} p(x,y,z) \log \frac{p(x \mid y,z)}{p(x \mid z)} \\
&= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z) p(y \mid z)} \\
&= I(X;Y \mid Z) \quad .
\end{aligned}
$$

□

Remark B.9. Conditional mutual information measures how much knowing Y reduces the uncertainty of X, *given* that we already know Z.

Proposition B.7 (Chain Rule for Mutual Information)**.** *Let X, Y, and Z be random variables. Then:*

$$I(X;Y,Z) = I(X;Z) + I(X;Y \mid Z) \quad .$$

Proof. We use entropy-based expressions for mutual information:

$$
\begin{aligned}
I(X;Y,Z) &= H(X) - H(X \mid (Y,Z)) \\
&= I(X;Z) + H(X \mid Z) - H(X \mid (Y,Z)) \\
&= I(X;Z) + H(X \mid Z) - (H(X \mid Z) - I(X;Y \mid Z)) \\
&= I(X;Z) + I(X;Y \mid Z) \quad .
\end{aligned}
$$

□

Proposition B.8 (Non-Negativity of Conditional Mutual Information)**.** *It holds true that*

$$I(X;Y \mid Z) \geq 0 \quad .$$

Proof. We have:

$$
\begin{aligned}
I(X;Y \mid Z) &= \sum_{x,y,z} p(x,y,z) \log \frac{p(x,y \mid z)}{p(x \mid z) p(y \mid z)} \\
&= \sum_z p(z) \sum_{x,y} p(x,y \mid z) \log \frac{p(x,y \mid z)}{p(x \mid z) p(y \mid z)} \\
&= \sum_z p(z) D_{\mathrm{KL}} \left(p(x,y \mid z) \,\|\, p(x \mid z) p(y \mid z) \right) \geq 0 \quad .
\end{aligned}
$$

□

Corollary B.6. *As a direct consequence, we have*

$$I(X;Z) \leq I(X;Y,Z) \quad .$$

Definition B.8 (Conditional Independence)**.** Let X, Y, Z be discrete random variables. We say that X is *conditionally independent* of Z given Y, and write:

$$X \perp Z \mid Y$$

if and only if

$$p(z \mid x, y) = p(z \mid y) \quad \text{for all } x, y, z \quad .$$

Equivalently:

$$p(x, z \mid y) = p(x \mid y)p(z \mid y) \quad .$$

Proposition B.9. *If $X \perp Z \mid Y$, then the conditional mutual information between X and Z given Y is zero:*

$$I(X;Z \mid Y) = 0 \quad .$$

Proof. By definition of conditional mutual information:

$$I(X;Z \mid Y) = \sum_{x,z,y} p(x,z,y) \log \frac{p(x,z \mid y)}{p(x \mid y)p(z \mid y)} \quad .$$

If $X \perp Z \mid Y$, then:

$$p(x, z \mid y) = p(x \mid y)p(z \mid y) \quad ,$$

so the logarithm becomes:

$$\log \frac{p(x \mid y)p(z \mid y)}{p(x \mid y)p(z \mid y)} = \log 1 = 0 \quad .$$

Hence, each term in the sum is zero, and:

$$I(X;Z \mid Y) = 0 \quad .$$

□

B.3.1 Data Processing Inequality

Lemma B.3 (Markov Chain)**.** *Let X, Y, Z be random discrete random variables forming the Markov chain $X \to Y \to Z$. Then:*

$$X \perp Z \mid Y \quad .$$

Proof. Per definition from Markov chains, we have:

$$p(z \mid x, y) = p(z \mid y) \quad ,$$

and hence $X \perp Z \mid Y$. □

Theorem B.3 (Data Processing Inequality)**.** *If $X \to Y \to Z$ is a Markov chain, then:*

$$I(X; Z) \leq I(X; Y) \quad .$$

Proof. We use the chain rule and conditional independence:

$$\begin{aligned} I(X; Z) &= I(X; Z, Y) - I(X; Y \mid Z) \\ &= I(X; Y) + I(X; Z \mid Y) - I(X; Y \mid Z) \quad . \end{aligned}$$

Since $X \to Y \to Z$, we have $I(X; Z \mid Y) = 0$. Thus:

$$I(X; Z) = I(X; Y) - I(X; Y \mid Z) \leq I(X; Y) \quad ,$$

because $I(X; Y \mid Z) \geq 0$. □

Corollary B.7 (No Gain in Processing)**.** *Any function $f(Y)$ of Y cannot increase information about X:*

$$I(X; f(Y)) \leq I(X; Y) \quad .$$

Proof. This follows by applying the DPI to the chain $X \to Y \to f(Y)$. □

Summary of Key Properties

- $I(X; Y) \geq 0$
- $I(X; Y) = 0$ if and only if $X \perp Y$
- $I(X; Y) = D_{\mathrm{KL}}(p(x, y) \| p(x)p(y))$
- Chain rule: $I(X; Y, Z) = I(X; Z) + I(X; Y \mid Z)$
- Data Processing Inequality: $X \to Y \to Z \Rightarrow I(X; Z) \leq I(X; Y)$

B.4 Bounding Mutual Information via Matrix Rank of the Joint Distribution

Theorem B.4. *Let X, Y be random variables from finite sets $\mathcal{X}, \mathcal{Y}$, and let matrix $\boldsymbol{P}$ denote their joint probability distribution, i.e. $\boldsymbol{P}_{ij} = p(x_i, y_j)$. Let $r \coloneqq \operatorname{rank} \boldsymbol{P}$ denote the rank of matrix $\boldsymbol{P}$. Then we have*

$$I(X;Y) \leq \log r \quad .$$

Proof. Let $n \coloneqq |\mathcal{X}|$ and $m \coloneqq |\mathcal{Y}|$. If $\boldsymbol{P}$ has rank r, then so must the transition matrix $\boldsymbol{P}_{Y|X} \in \mathbb{R}^{m \times n}$ defined as $(\boldsymbol{P}_{Y|X})_{ij} \coloneqq p(y_i \mid x_j) = \frac{p(x_j, y_i)}{\sum_k p(x_k, y_i)}$, since $\boldsymbol{P}_{Y|X}$ is created from $\boldsymbol{P}$ by transposing and column scaling. If one column consisted of only zeros, i.e. $\sum_k p(x_k, y_i) = 0$, we may just copy a different scaled column vector to this column.

Now, let's analyze matrix $\boldsymbol{P}_{Y|X}$. First, note that it is a Markov chain transition matrix, and hence all its columns lie in the m-dimensional unit simplex. Consider the convex hull of the column vectors, it is a r-dimensional convex polytope in the m-dimensional unit simplex. Thus, we can find a r-dimensional simplex with corners collected by matrix $\boldsymbol{U}$ s.t. it is a superset of this polytope and still a subset of the (potentially) higher dimensional unit simplex.

Thus, every column vector in $\boldsymbol{P}_{Y|X}$ can be written as a convex combination of the column vectors in $\boldsymbol{U}$. It follows that $\boldsymbol{P}_{Y|X}$ can be decomposed as

$$\boldsymbol{P}_{Y|X} = \boldsymbol{U}\boldsymbol{V} \ , \quad \boldsymbol{U} \in \mathbb{R}^{m \times r}, \boldsymbol{V} \in \mathbb{R}^{r \times n} \quad ,$$

where both $\boldsymbol{U}$ and $\boldsymbol{V}$ are Markov chain transition matrices as well.

Hence, we can introduce a latent variable $Z \in \{1, \ldots, r\}$, which forms the Markov chain

$$X \underset{\boldsymbol{V}}{\to} Z \underset{\boldsymbol{U}}{\to} Y \quad .$$

Finally, based on theorem B.3 it follows that

$$I(X;Y) \leq I(X;Z) = H(Z) - H(Z \mid X) \leq H(Z) \leq \log r \quad .$$

□