Criticality in Formal Languages

Jonas Peters

July 28, 2025

Contents

1 Model Framework

Empirical analysis of *natural language* reveals an interesting relation between the distance of characters in a text and their mutual information.

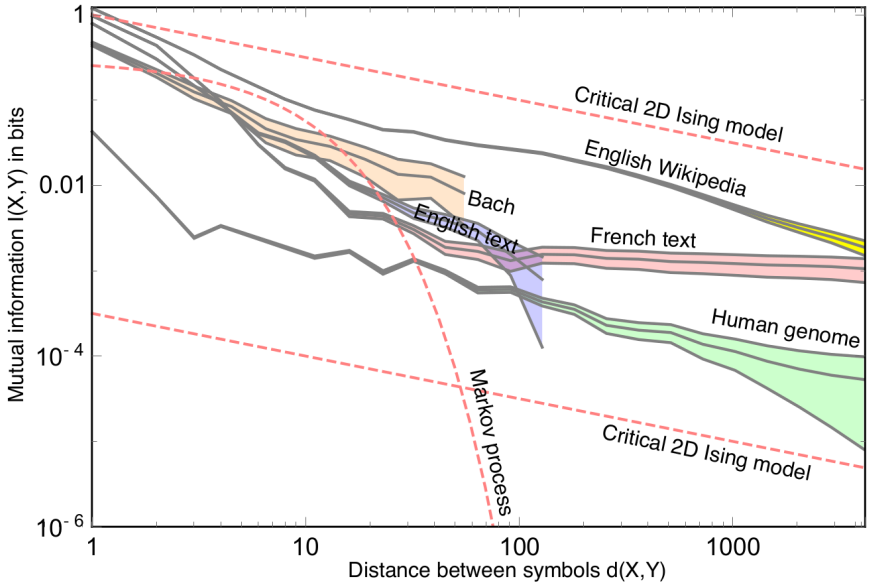

Figure 1: Decay of mutual information with separation. Source: [1]

The data show that mutual information tends to decay with distance, but it does so slowly. To be precise, *mutual information in natural language seems to follow a power-law.*

Thus, models of natural language should show a similar behavior. In order to filter potential models for natural language by their ability for *power-law behavior*, we need a precise definition.

1.1 The Framework

The tokens, represented by random variables X_t, are elements of a finite alphabet Σ. For every t and every separation $\tau > 0$, 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 ensures that the mutual information does not decay too quickly, while the latter ensures that $I(X_t; X_{t+\tau}) \xrightarrow{\tau \to \infty} 0$, just like the data show. We also may replace the latter condition by this implication.

The importance of the second condition is further emphasized when considering models like the following time-homogenous Markov chain consisting of the two states A and B: The probability to transition from A to A is one, and hence the transition probability from A to B is zero. Similarly, starting at state B, we can only remain at state B.

The pairwise mutual information of this simple model equals a constant. Thus, we can lower bound it by a power-law. However, the problem is that the mutual information does not *decay with distance*. Thus, we conclude that high mutual information alone is not a good indicator for model quality.

1.1.1 Model Definition

In sequence modeling, a model is typically built from a finite set of rules or parameters defined over an alphabet Σ. These rules allow us to assign a probability to any finite string. For instance, a time-homogeneous Markov chain uses a fixed transition matrix to define a probability measure over the set of all strings of a given length n (i.e., over Σ^n) for any $n \in \mathbb{N}$.

This central idea of a family of finite domains leads to our general definition:

Definition 1.1 (Model over Σ^*)**.** Let Σ be a finite alphabet. A model S over Σ^* is a function $S : \Sigma^* \mapsto [0, 1]$ that assigns a probability to each finite string $w \in \Sigma^*$, subject to the constraint that for any length $n \in \mathbb{N}$, the probabilities of all strings of that length sum to one:

$$\sum_{w \in \Sigma^n} S(w) = 1 \quad .$$

Furthermore, we denote the restriction of S to strings of length n as $S_n \coloneqq S|_{\Sigma^n}$. The function $S_n : \Sigma^n \to [0, 1]$ is thus a probability measure over Σ^n.

Another strength of this definition is that we don't constrain models by their parameterization. How the models are defined, and how they compute the probabilities is up to them.

Additionally, we might want to restrain S in order to have reasonable time and space complexity, and to ensure the model is *consistent*, which means that the language of S_n should look *similar* to S_{n+d}, whatever this might mean. We also write w_i for X_i. We can think of w as a 1-indexed String of random variables.

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 $n \in \mathbb{N},\ w \in \Sigma^n$ it holds true that

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

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 $w \in \Sigma^n$:*

$$\sum_{s \in \Sigma^d} S_{n+d}(ws) = S_n(w) \quad .$$

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

$$\begin{aligned}
\sum_{s \in \Sigma^{k+1}} S_{n+k+1}(ws) &= \sum_{v \in \Sigma^k} \sum_{c \in \Sigma} S_{n+k+1}(wvc) \\
&= \sum_{v \in \Sigma^k} \left(\sum_{c \in \Sigma} S_{n+k+1}((wv)c) \right) \\
&\underset{\text{bulk marginal property}}{=} \sum_{v \in \Sigma^k} S_{n+k}(wv) \\
&\underset{\text{inductive hypothesis}}{=} S_n(w) \quad ,
\end{aligned}$$

which concludes the induction. □

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

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

where $w \in \Sigma^n, c \in \Sigma$.

Remark 1.1. If $\sum_{c' \in \Sigma} S_{n+1}(wc') = 0$, we might set $S^*_{n+1}(wc) \coloneqq S^*_n(w) \dfrac{1}{|\Sigma|}$.

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

Proof. We have:

$$\begin{aligned}\sum_{c\in\Sigma} S^*_{n+1}(wc) &\underset{\text{def of } S^*}{=} \sum_{c\in\Sigma} S^*_n(w)\frac{S_{n+1}(wc)}{\sum_{c'\in\Sigma} S_{n+1}(wc')} \\ &= S^*_n(w)\cdot\frac{1}{\sum_{c'\in\Sigma} S_{n+1}(wc')}\sum_{c\in\Sigma} S_{n+1}(wc) \\ &= S^*_n(w)\cdot\frac{\sum_{c\in\Sigma} S_{n+1}(wc)}{\sum_{c'\in\Sigma} S_{n+1}(wc')} \\ &\overset{\checkmark}{=} S^*_n(w) \quad .\end{aligned}$$

□

1.1.2 Restricting the Model

We can restrict our general model S by specifying that for each length n, the probability distribution S_n is computed by a parameterized **inference function**, $f_{n,\boldsymbol{\theta}_n} : \Sigma^n \to [0,1]$. Each function is identified by a parameter vector $\boldsymbol{\theta}_n$ from a corresponding **parameter space** Θ_n.

For the framework to be valid, each function must define a proper probability distribution. We denote this distribution by $S_{n,\boldsymbol{\theta}_n}$ and require that:

$$S_{n,\boldsymbol{\theta}_n}(w) \coloneqq f_{n,\boldsymbol{\theta}_n}(w) \quad \text{and} \quad \sum_{w\in\Sigma^n} f_{n,\boldsymbol{\theta}_n}(w) = 1 \quad .$$

This approach defines a **model class** $\mathcal{S}_n$, which is the set of all distributions that can be generated by the family of inference functions with parameters in Θ_n:

$$\mathcal{S}_n \coloneqq \{S_{n,\boldsymbol{\theta}_n} \mid \boldsymbol{\theta}_n \in \Theta_n\} \quad .$$

A complete model $S \equiv (S_n)_{n\in\mathbb{N}}$ is thus specified by an inference function and a corresponding sequence of chosen parameters $(\boldsymbol{\theta}_n)_{n\in\mathbb{N}}$.

Remark 1.2. The distinction between the inference function f and the distribution S is crucial. Two different functions, $f_{n,\boldsymbol{\theta}}$ and $f_{n,\boldsymbol{\theta}'}$, might have vastly different time and space complexities even if they compute the exact same distribution (i.e., $S_{n,\boldsymbol{\theta}} = S_{n,\boldsymbol{\theta}'}$). The complexity is therefore a property of the specific algorithmic implementation and parameterization of $f_{n,\boldsymbol{\theta}_n}$.

1.1.3 Power-Law Behavior

Now, we formalize what it means for a model to exhibit power-law behavior. Specifically, we require power-law decay in mutual information with respect to τ between *any* two variables X_t and $X_{t+\tau}$. This condition must hold for all t and for samples from *any* distribution S_n where $t + \tau \leq n$.

Definition 1.4. 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 sampled from S_n. Then:

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

Definition 1.5 (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 1.3. For a model S^* with the bulk marginal property we can replace "for every $n \in \mathbb{N}$" in definition 1.5 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 1.6. 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 1.7 (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 (Power-Law Decay for All Tokens in Models with the Bulk Marginal Property)**.** *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 $I(X_t; X_{t+\tau}) \in \Omega(\tau^{-\alpha})$. □

In other words, mutual information decays polynomially with distance for all tokens, and the constant α is always the same. This sounds like strong power-law behavior, but the issue is that the scalar c depends on t and we cannot lower bound it by constant greater than zero. Hence, the *starting threshold* $I(X_t; X_{t+1})$ can decay to zero for $t \to \infty$.

Remark 1.4. 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 1.5. 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 of hidden Markov models (and hence also strong power-law behavior).

Remark 1.6. 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 1.7. 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 1.8. 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 $\tau \in \mathbb{N}$:

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

2 A Model with Power-Law Behavior

Based on our definitions it is very easy to define models without power-law behavior. For example, a single pair of random variables which are independent when marginalizing over the other ones violates every definition of power-law behavior, as it implies a mutual information of zero. Furthermore, hidden Markov models are also incapable of producing power-law behavior, see chapter 3.

This begs the question whether there actually exists a model that satisfies our strong power-law behavior definition.

2.1 A Continuous Model with Power-Law Behavior

Let's consider a multivariate continuous normal distribution. The following two properties of the normal distribution prove to be very helpful for our analysis of pairwise mutual information. They are standard results and thus are stated without proof.

Proposition 2.1 (Marginal Distributions of a Normal Distribution)**.** *Let the n-dimensional random vector $\boldsymbol{X}$ follow a multivariate normal distribution, $\boldsymbol{X} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. We partition $\boldsymbol{X}$ into two sub-vectors, $\boldsymbol{X}_1 \in \mathbb{R}^k$ and $\boldsymbol{X}_2 \in \mathbb{R}^{n-k}$, with the corresponding partitions of the mean vector $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$ as:*

$$\boldsymbol{X} = \begin{pmatrix} \boldsymbol{X}_1 \\ \boldsymbol{X}_2 \end{pmatrix} \quad , \quad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_1 \\ \boldsymbol{\mu}_2 \end{pmatrix} \quad , \quad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{11} & \boldsymbol{\Sigma}_{12} \\ \boldsymbol{\Sigma}_{21} & \boldsymbol{\Sigma}_{22} \end{pmatrix} \quad ,$$

where $\boldsymbol{\mu}_1 \in \mathbb{R}^k$ and $\boldsymbol{\Sigma}_{11}$ is a $k \times k$ matrix.

Then the marginal distribution of the sub-vector $\boldsymbol{X}_1$ is also a multivariate normal distribution given by:

$$\boldsymbol{X}_1 \sim \mathcal{N}(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_{11}) \quad .$$

Proposition 2.2 (Entropy of a Multivariate Normal Distribution)**.** *Let the random vector $\boldsymbol{X} \in \mathbb{R}^n$ follow a multivariate normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ with a positive definite covariance matrix $\boldsymbol{\Sigma}$. The entropy of $\boldsymbol{X}$ is given by*

$$H(\boldsymbol{X}) = \frac{1}{2} \log\left((2\pi e)^n \det(\boldsymbol{\Sigma})\right) \quad ,$$

where log *is the natural logarithm.*

2.1.1 A Formula for Mutual Information

With these two propositions at hand, we can derive a formula for mutual information based on the parameters of our normal distribution.

To this end, let $(Y_1, \ldots, Y_n)$ denote the random variables of the n-dimensional normal distribution

$$\mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho_1 & \rho_2 & \cdots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \cdots & \rho_{n-2} \\ \vdots & & \ddots & & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \cdots & 1 \end{pmatrix}\right) \quad .$$

Using proposition 2.1, we get the marginal distribution of Y_i, Y_j:

$$\begin{pmatrix} Y_i \\ Y_j \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho_\tau \\ \rho_\tau & 1 \end{pmatrix}\right) \quad ,$$

where we define $\tau \coloneqq |i - j|$.

Similarly, $Y_i \sim \mathcal{N}(0, 1)$ and $Y_j \sim \mathcal{N}(0, 1)$. Using proposition 2.2, we have:

$$\begin{aligned} I(Y_i; Y_j) &= H(Y_i) + H(Y_j) - H(Y_i, Y_j) \\ &= \left[\frac{1}{2}\log(2\pi e \cdot 1)\right] + \left[\frac{1}{2}\log(2\pi e \cdot 1)\right] - \left[\frac{1}{2}\log\left((2\pi e)^2 \det\begin{pmatrix} 1 & \rho_\tau \\ \rho_\tau & 1 \end{pmatrix}\right)\right] \\ &= \log(2\pi e) - \frac{1}{2}\log\left((2\pi e)^2(1 - \rho_\tau^2)\right) \\ &= \log(2\pi e) - \frac{1}{2}\left(\log((2\pi e)^2) + \log(1 - \rho_\tau^2)\right) \\ &= \log(2\pi e) - \frac{1}{2}\left(2\log(2\pi e) + \log(1 - \rho_\tau^2)\right) \\ &= \log(2\pi e) - \log(2\pi e) - \frac{1}{2}\log(1 - \rho_\tau^2) \\ &= -\frac{1}{2}\log(1 - \rho_\tau^2) \quad . \end{aligned} \tag{1}$$

Great, the mutual information $I(Y_i; Y_j)$ is fully specified by the parameter ρ_τ, and hence for every pair (Y_i, Y_j) we can fine tune $I(Y_i; Y_j)$ by only changing ρ_τ.

2.1.2 Initializing Parameters

We want the mutual information $I(Y_i; Y_j)$ to follow a power-law, i.e.

$$I(Y_i; Y_j) \stackrel{!}{=} c|i - j|^{-\alpha} \quad ,$$

for some $c, \alpha \in \mathbb{R}_{>0}$.

Hence, based on our previous result we have:

$$\begin{aligned}
& I(Y_i; Y_j) = c|i-j|^{-\alpha} \\
\iff \quad & -\frac{1}{2}\log(1-\rho_\tau^2) = c\tau^{-\alpha} \\
\iff \quad & \log(1-\rho_\tau^2) = -2c\tau^{-\alpha} \\
\iff \quad & 1-\rho_\tau^2 = e^{-2c\tau^{-\alpha}} \\
\iff \quad & \rho_\tau^2 = 1 - e^{-2c\tau^{-\alpha}} \\
\iff \quad & \rho_\tau = \pm\sqrt{1-e^{-2c\tau^{-\alpha}}} \quad .
\end{aligned}$$

Thus, upon defining the constants c and α, we can directly calculate the parameters ρ_τ, where we choose ρ_τ to be positive.

It seems like we are done, but not so fast! In order for our covariance matrix to define a valid normal distribution, we have to ensure its positive definiteness. But, it turns out that this is not an issue:

2.1.3 Ensure Positive Definiteness

Note that we have the freedom to define $c, \alpha \in \mathbb{R}_{>0}$ how we like. Thus, our choice of these constants should imply positive definiteness of the covariance matrix

$$\boldsymbol{\Sigma}_n = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{n-2} \\ \vdots & & \ddots & & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \dots & 1 \end{pmatrix} \quad .$$

Note that $\boldsymbol{\Sigma} \equiv \boldsymbol{\Sigma}_n$ is symmetric, and has positive entries along its diagonal. Thus, it is sufficient for positive definiteness to show that $\boldsymbol{\Sigma}$ is strictly diagonally dominant, i.e.

$$\begin{aligned}
& \forall i \in [n] : |\boldsymbol{\Sigma}_{ii}| > \sum_{j \neq i} |\boldsymbol{\Sigma}_{ij}| \\
\iff \quad & \forall i \in [n] : 1 > \sum_{j \neq i} \rho_{|i-j|} \quad . \qquad (2)
\end{aligned}$$

Note that a specific entry ρ_τ can occur two times in the same row. This happens especially in the middle rows of the matrix. However, for a fixed τ, ρ_τ can occur

at most two times in the same row. Hence, we derive the following bound:

$$\sum_{j \neq i} \rho_{|i-j|} \leq 2 \sum_{\tau=1}^{\infty} \rho_\tau \quad . \tag{3}$$

It seems like we need an upper bound for ρ_τ. Luckily, we can employ a clever bounding technique which relies on $x + 1 \leq e^x$:

Lemma 2.1. *Let* $\rho_\tau \coloneqq \sqrt{1 - e^{-2c\tau^{-\alpha}}}$. *Then, we have*

$$\rho_\tau \leq \sqrt{2c}\tau^{-0.5\alpha} \quad ,$$

where the inequality holds for all $\tau > (2c)^{\frac{1}{\alpha}}$.

Proof. Note the inequality $x + 1 \leq e^x$. For $x > -1$ it follows that

$$e^{-x} \leq \frac{1}{x+1} \quad . \tag{4}$$

Applying equation 4 to $e^{2c\tau^{-\alpha}}$ with $x \coloneqq -2c\tau^{-\alpha}$ yields

$$e^{2c\tau^{-\alpha}} \leq \frac{1}{-2c\tau^{-\alpha} + 1} \quad ,$$

where we have $x > -1 \iff -2c\tau^{-\alpha} > -1 \iff \tau > (2c)^{\frac{1}{\alpha}}$.

Under this assumption, both sides of the inequality are positive, and hence it follows that

$$e^{-2c\tau^{-\alpha}} \geq 1 - 2c\tau^{-\alpha} \quad .$$

Plugging this into the equation for ρ_τ gives

$$\begin{aligned} \rho_\tau &\leq \sqrt{1 - (1 - 2c\tau^{-\alpha})} \\ &= \sqrt{2c\tau^{-\alpha}} \\ &= \sqrt{2c}\tau^{-0.5\alpha} \quad . \end{aligned}$$

□

We can directly use our established inequality in equation (3). To this end, set $\alpha \coloneqq 4$, and $c < \frac{9}{2\pi^4}$, like $c \coloneqq 0.045$. For $\tau > (2c)^{\frac{1}{\alpha}}$ we can now use the upper

bound, which translates to all $\tau \in \mathbb{N}_{>0}$. Hence:

$$\begin{aligned}
\sum_{j \neq i} \rho_{|i-j|} &\leq 2 \sum_{\tau=1}^{\infty} \rho_\tau \\
&\leq 2 \sum_{\tau=1}^{\infty} \frac{\sqrt{2 \cdot 0.045}}{\tau^2} \\
&= 2 \cdot 0.3 \sum_{\tau=1}^{\infty} \frac{1}{\tau^2} \\
&= 2 \cdot 0.3 \cdot \frac{\pi^2}{6} \\
&= \frac{\pi^2}{10} \\
&< 1 \quad ,
\end{aligned}$$

which proves equation (2), and hence ultimately the positive definiteness of $\mathbf{\Sigma}$.

With that we have successfully defined a model where the pairwise mutual information perfectly follows a power-law! Now, we would like to extend this family of normal distributions so it fits our model definition **??**. The only thing needed is a finite domain of our random variables.

2.2 The Discretized Model

The central idea is to discretize our probability space by integrating over the quadrants. Thus, we effectively created a probability measure S_n over $\{-1, 1\}^n$, where for example $S_2(11)$ is defined as the integral over the first quadrant, $S_2(-11)$ as the integral over the second quadrant, and so on. This leads to our formal definition:

Definition 2.1 (The Model)**.** Let $(\mathcal{N}(\mathbf{0}, \mathbf{\Sigma}_n))_{n=1}^{\infty}$ be a family of normal distributions with a positive definite parameter matrix $\mathbf{\Sigma}_n$ for all $n \in \mathbb{N}_{>0}$, and let $p_n(\boldsymbol{x})$ denote the associated probability density functions. We define the probability measure $S_{n,\mathbf{\Sigma}_n}$ over $\{-1, 1\}^n$ as:

$$S_{n,\mathbf{\Sigma}_n}(w) \coloneqq \int_{Q_w} p_n(\boldsymbol{x}) d\boldsymbol{x} \quad ,$$

where Q_w is the quadrant

$$Q_w \coloneqq \{\boldsymbol{x} \in \mathbb{R}^n \mid \forall i \in [n] : w_i \boldsymbol{x}_i \geq 0\} \quad .$$

By defining our model this way, we ensure that S_n is a valid probability measure over $\{-1,1\}^n$ for every $n \in \mathbb{N}$. Thus, we can dive right into the analysis of pairwise mutual information:

2.2.1 Strong Power-Law Behavior

In this section we will prove that our model has the desired property of strong power-law behavior according to definition 1.5.

Another way to think about our model is that the continuous normal distribution has an associated vector $(Y_1, \ldots, Y_n) \in \mathbb{R}^n$ of random variables, and in order to get the random variables $(X_1, \ldots, X_n) \in \{-1,1\}^n$ of our model, we *process* each Y_i in the following manner:

$$X_i = \begin{cases} 1 & \text{if } Y_i \geq 0 \\ -1 & \text{else} \end{cases} \quad .$$

We can make use of the data processing inequality to derive that

$$I(X_i; X_j) \leq I(X_i; Y_j) \leq I(Y_i; Y_j) = c \cdot |i-j|^{-\alpha} \quad ,$$

i.e. the model has upper bound power-law behavior.

The final step is to prove strong lower bound power-law behavior. This one is a bit trickier, as we will use multiple bounds in our calculation.

In order to calculate and bound $I(X_i; X_j)$, we need the joint distribution of these two variables when marginalizing over the other ones.

Here is the clever part: Instead of discretizing the continuous normal distribution first and then marginalizing our distribution, we can also marginalize the continuous distribution first and only then transition to our discretized model by integrating. Thus, we can use proposition 2.1 again in order to analyze

$$\begin{pmatrix} Y_i \\ Y_j \end{pmatrix} \sim \mathcal{N}\left(\mathbf{0}, \begin{pmatrix} 1 & \rho_\tau \\ \rho_\tau & 1 \end{pmatrix}\right) \quad .$$

Thus, in order to calculate the joint distribution of (X_i, X_j), we need a formula for integrating a two dimensional normal distribution over the quadrants. As it turns out, the solution is a neat formula in terms of ρ:

$$P(Y_i > 0, Y_j > 0) = \frac{1}{4} + \frac{1}{2\pi} \arcsin(\rho) \quad .$$

The derivation of this fact is rather technical and can be found in section A.1 of the appendix.

Now, set $\delta_\tau := \frac{\arcsin(\rho_\tau)}{2\pi}$ for the sake of clarity. Note that since $\rho_\tau \in (0,1)$, we have $\delta_\tau \in (0, \frac{1}{4})$. Based on our formula, and using the symmetry of the normal distribution, we derive the joint probability distribution of (X_i, X_j):

	$X_i = 1$	$X_i = -1$
$X_j = 1$	$\frac{1}{4} + \delta_\tau$	$\frac{1}{4} - \delta_\tau$
$X_j = -1$	$\frac{1}{4} - \delta_\tau$	$\frac{1}{4} + \delta_\tau$

Now that we know the joint probability distribution, we can finally compute the mutual information $I(X_i; X_j)$. First, note that the marginalized distribution of a single X_i is just the uniform distribution on $\{-1, 1\}$, and hence $H(X_i) = H(X_j) = \log 2$. Furthermore, let's substitute $f : (0, \frac{1}{2}) \to \mathbb{R},\ x \mapsto x \log x$, where log is the natural logarithm. Thus:

$$\begin{aligned} I(X_i; X_j) &= H(X_i) + H(X_j) - H(X_i, X_j) \\ &= \log 2 + \log 2 - \left[- \sum_{(x_i, x_j) \in \{-1,1\}^2} p(x_i, x_j) \log p(x_i, x_j) \right] \\ &= \log 4 + 2 \cdot f(\frac{1}{4} - \delta_\tau) + 2 \cdot f(\frac{1}{4} + \delta_\tau) \quad . \end{aligned}$$

Using that $\log 4 = -\log \frac{1}{4} = -4 \cdot f(\frac{1}{4})$, we arrive at

$$\begin{aligned} I(X_i; X_j) &= 2 \cdot f(\frac{1}{4} - \delta_\tau) + 2 \cdot f(\frac{1}{4} + \delta_\tau) - 4 \cdot f(\frac{1}{4}) \\ &= 4 \cdot \left[\frac{1}{2} \cdot f(\frac{1}{4} - \delta_\tau) + \frac{1}{2} \cdot f(\frac{1}{4} + \delta_\tau) - f(\frac{1}{4}) \right] \quad . \end{aligned} \tag{5}$$

This expression has a very peculiar form: It looks like it measures how *convex* the function f is at the point $\frac{1}{4}$. Hence, we might guess that we can lower bound the expression by substituting a less convex function g for f. This idea is formalized in the following lemma:

Lemma 2.2. *Let $I \subseteq \mathbb{R}$ be an interval, and let $f, g \in \mathcal{C}^2(I, \mathbb{R})$ be two functions that are twice differentiable s.t. $f''(x) \geq g''(x)$ for all $x \in I$. Then the following inequality holds for all $\delta \in \mathbb{R}_{>0}, x_0 \in \mathbb{R}$ s.t. $[x_0 - \delta, x_0 + \delta] \subseteq I$:*

$$\frac{1}{2} \cdot f(x_0 - \delta) + \frac{1}{2} \cdot f(x_0 + \delta) - f(x_0) \geq \frac{1}{2} \cdot g(x_0 - \delta) + \frac{1}{2} \cdot g(x_0 + \delta) - g(x_0) \quad .$$

Proof. First, note that $[x_0 - \delta, x_0 + \delta] \subseteq I$ implies $x_0 \in I$. Now, define a new function $h : I \to \mathbb{R}$ as

$$h(x) \coloneqq f(x) - g(x) \quad .$$

Compting

$$h''(x) = f''(x) - g''(x) \geq 0 \quad \forall x \in I$$

shows that h is convex on I. Hence, we conclude:

$$\frac{1}{2}h(x_0 - \delta) + \frac{1}{2}h(x_0 + \delta) \geq h(x_0) \quad .$$

Now, after substituting the definition of $h(x) = f(x) - g(x)$ back into this inequality

$$\frac{1}{2}[f(x_0 - \delta) - g(x_0 - \delta)] + \frac{1}{2}[f(x_0 + \delta) - g(x_0 + \delta)] \geq f(x_0) - g(x_0)$$

and rearranging the terms to separate the functions f and g, we arrive at the desired result:

$$\frac{1}{2}f(x_0 - \delta) + \frac{1}{2}f(x_0 + \delta) - f(x_0) \geq \frac{1}{2}g(x_0 - \delta) + \frac{1}{2}g(x_0 + \delta) - g(x_0) \quad .$$

□

Remember, we defined $f : (0, \frac{1}{2}) \to \mathbb{R},\ x \mapsto x \log x$. Hence:

$$f'(x) = \log x + 1$$
$$f''(x) = \frac{1}{x} \quad .$$

Note that $f''(x) \geq 2$ for all $x \in (0, \frac{1}{2})$. Thus, when we define $g : (0, \frac{1}{2}) \to \mathbb{R},\ x \mapsto x^2$, where $g''(x) \equiv 2$, we can use lemma 2.2 in equation (5):

$$
\begin{aligned}
I(X_i; X_j) &= 4 \cdot \left[\frac{1}{2} \cdot f(\frac{1}{4} - \delta_\tau) + \frac{1}{2} \cdot f(\frac{1}{4} + \delta_\tau) - f(\frac{1}{4}) \right] \\
&\geq 4 \cdot \left[\frac{1}{2} \cdot g(\frac{1}{4} - \delta_\tau) + \frac{1}{2} \cdot g(\frac{1}{4} + \delta_\tau) - g(\frac{1}{4}) \right] \\
&= 4 \cdot \left[\frac{1}{2} \left(\frac{1}{4} - \delta_\tau \right)^2 + \frac{1}{2} \left(\frac{1}{4} + \delta_\tau \right)^2 - \left(\frac{1}{4} \right)^2 \right] \\
&= 2 \left[\left(\frac{1}{16} - \frac{1}{2}\delta_\tau + \delta_\tau^2 \right) + \left(\frac{1}{16} + \frac{1}{2}\delta_\tau + \delta_\tau^2 \right) \right] - 4 \left(\frac{1}{16} \right) \\
&= 2 \left[\frac{2}{16} + 2\delta_\tau^2 \right] - \frac{4}{16} \\
&= 2 \left[\frac{1}{8} + 2\delta_\tau^2 \right] - \frac{1}{4} \\
&= \frac{1}{4} + 4\delta_\tau^2 - \frac{1}{4} \\
&= 4\delta_\tau^2 \quad .
\end{aligned}
$$

Since $\delta_\tau = \frac{\arcsin(\rho_\tau)}{2\pi}$, we conclude

$$
\begin{aligned}
I(X_i; X_j) &\geq 4 \frac{\arcsin(\rho_\tau)^2}{4\pi^2} \\
&\geq \frac{\rho_\tau^2}{\pi^2} \quad ,
\end{aligned}
$$

since $\arcsin(x) \geq x$ for $x \in (0, 1)$.

This is looking very promising! We only need to bound ρ_τ again, but this time from below:

Lemma 2.3. *Let $\rho_\tau \coloneqq \sqrt{1 - e^{-2c\tau^{-\alpha}}}$. Then, we have*

$$
\sqrt{\frac{2c}{2c+1}} \tau^{-0.5\alpha} \leq \rho_\tau
$$

for all $\tau \in \mathbb{N}_{>0}$.

Proof. Note the inequality $x + 1 \leq e^x$. For $x > -1$ it follows that

$$
e^{-x} \leq \frac{1}{x+1} \quad . \tag{6}
$$

Applying equation 6 to $e^{-2c\tau^{-\alpha}}$ with $x \coloneqq 2c\tau^{-\alpha}$ yields

$$
e^{-2c\tau^{-\alpha}} \leq \frac{1}{2c\tau^{-\alpha} + 1} \quad ,
$$

where we have $x > -1 \iff 2c\tau^{-\alpha} > -1 \impliedby \tau \in \mathbb{N}_{>0}$.

Plugging this into the equation for ρ_τ gives

$$\begin{aligned}
\rho_\tau &\geq \sqrt{1 - \frac{1}{2c\tau^{-\alpha} + 1}} \\
&= \sqrt{\frac{2c\tau^{-\alpha}}{2c\tau^{-\alpha} + 1}} \\
&= \sqrt{2c}\tau^{-0.5\alpha}\sqrt{\frac{1}{2c\tau^{-\alpha} + 1}} \\
&\geq \sqrt{2c}\tau^{-0.5\alpha}\sqrt{\frac{1}{2c1^{-\alpha} + 1}} \\
&= \sqrt{\frac{2c}{2c+1}}\tau^{-0.5\alpha} \quad .
\end{aligned}$$

□

Finally, we use the lower bound of ρ_τ provided by lemma 2.3 to arrive at

$$I(X_i; X_j) \geq \frac{2c}{(2c+1) \cdot \pi^2}\tau^{-\alpha} \quad .$$

This proves the strong lower bound power-law behavior property of our model.

2.3 Summary

Let's summarize our findings in a concise theorem:

Theorem 2.1 (A Model with Strong Power-Law Behavior)**.** *Define $\alpha \coloneqq 4$, and $c \coloneqq 0.045$. Furthermore, let $\rho_\tau \coloneqq \sqrt{1 - e^{-2c\tau^{-\alpha}}}$. We define the matrix*

$$\mathbf{\Sigma}_n = \begin{pmatrix} 1 & \rho_1 & \rho_2 & \dots & \rho_{n-1} \\ \rho_1 & 1 & \rho_1 & \dots & \rho_{n-2} \\ \vdots & & \ddots & & \vdots \\ \rho_{n-1} & \rho_{n-2} & \rho_{n-3} & \dots & 1 \end{pmatrix} \quad .$$

Define the model S according to definition 2.1 using the covariance matrix $\mathbf{\Sigma}_n$. It follows that $S_{n,\mathbf{\Sigma}_n}$ is a valid probability measure over $\{-1, 1\}^n$, since $\mathbf{\Sigma}_n$ is positive definite especially. Furthermore, S has strong power-law behavior according to definition 1.5. Specifically, for any random variables X_i, X_j sampled from S, we have:

$$\frac{2c}{(2c+1) \cdot \pi^2} |i-j|^{-\alpha} \leq I(X_i; X_j) \leq |i-j|^{-\alpha} \quad .$$

Proof. The proof directly follows from our preliminary considerations. □

3 No Power-Law Behavior 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 influence 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 time-homogenous Markov chains $P(X_{t+1} = a \mid X_t = b)$ is independent of t and hence is constant over time. Thus, all the arrows in figure 2 represent the same transition.

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 with implicitly infinite transitions). Furthermore, when making a prediction of the next token given a fixed context window, it seems reasonable to assume invariance in time, i.e. fixed transition probabilities similar to time-homogenous Markov chains.

Now, the question is, can we achieve power-law behavior with an arbitrarily large fixed-sized context window with fixed transition probabilities?

In order to answer this question, let us first reduce our setting to a simpler model, specifically to the already mentioned time-homogenous Markov chains:

The idea is to employ a hidden variable $Y \in \Sigma^{s+1}$, where Σ is the alphabet, and s is the size of the context window (for Markov chains $s = 1$). Clearly, Y captures the entire *state* at time t of our model, that is all the previous s tokens and the current one, and we can model the transitions $Y_t \to Y_{t+1}$ as simple time-homogenous Markov chain transitions (and hence invariant in time). And, of course, once we know Y_t, we also know X_t (which of course can be modelled with time-homogenous Markov chain transitions as well). These models are known as *hidden Markov models*.

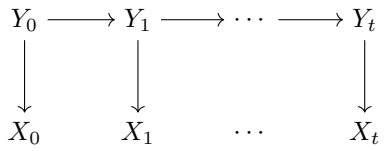

Figure 3: Bayesian network of a hidden Markov model.

Unfortunately, there is no free lunch, as we will see in the following. However, in order to use theorem 1.1 to disprove weak and thus strong power-law behavior, we need to prove the bulk marginal property of hidden Markov models first:

Lemma 3.1 (Hidden Markov Models have the Bulk Marginal Property)**.** *Every hidden Markov model with finite state spaces (S_Y, S_X) for its latent variable Y and observable variable X with transition matrices $(\boldsymbol{M_Y}, \boldsymbol{M_X})$ complies with the bulk marginal property.*

Proof. Let $w_i \coloneqq X_{i-1}$ for $n \in [n+1]$. Then we have:

$$
\begin{aligned}
&\sum_{w_{n+1}\in\Sigma} S_{n+1}(w) \\
&\underset{\text{Bayesian network}}{=} \sum_{w_{n+1}\in\Sigma} \sum_{q_1,\dots,q_{n+1}\in S_Y} \left[P(q_1) \prod_{i=2}^{n+1} P(q_i \mid q_{i-1}) \cdot \prod_{i=1}^{n+1} P(w_i \mid q_i) \right] \\
&= \sum_{q_1,\dots,q_{n+1}\in S_Y} \sum_{w_{n+1}\in\Sigma} \left[P(q_1) \prod_{i=2}^{n+1} P(q_i \mid q_{i-1}) \cdot \prod_{i=1}^{n+1} P(w_i \mid q_i) \right] \\
&= \sum_{q_1,\dots,q_{n+1}\in S_Y} \left[P(q_1) \prod_{i=2}^{n+1} P(q_i \mid q_{i-1}) \cdot \prod_{i=1}^{n} P(w_i \mid q_i) \sum_{w_{n+1}\in\Sigma} P(w_{n+1} \mid q_{n+1}) \right] \\
&= \sum_{q_1,\dots,q_{n+1}\in S_Y} \left[P(q_1) \prod_{i=2}^{n+1} P(q_i \mid q_{i-1}) \cdot \prod_{i=1}^{n} P(w_i \mid q_i) \right] \\
&= \sum_{q_1,\dots,q_n\in S_Y} \sum_{q_{n+1}\in S_Y} P(q_{n+1} \mid q_n) \left[P(q_1) \prod_{i=2}^{n} P(q_i \mid q_{i-1}) \cdot \prod_{i=1}^{n} P(w_i \mid q_i) \right] \\
&= \sum_{q_1,\dots,q_n\in S_Y} \left[P(q_1) \prod_{i=2}^{n} P(q_i \mid q_{i-1}) \cdot \prod_{i=1}^{n} P(w_i \mid q_i) \sum_{q_{n+1}\in S_Y} P(q_{n+1} \mid q_n) \right] \\
&= \sum_{q_1,\dots,q_n\in S_Y} \left[P(q_1) \prod_{i=2}^{n} P(q_i \mid q_{i-1}) \cdot \prod_{i=1}^{n} P(w_i \mid q_i) \right] \\
&\overset{\checkmark}{=} S_n(w_{-\{n+1\}}) \quad .
\end{aligned}
$$

□

We will also make use of the following well-established results regarding time-homogenous Markov chains. Hence, they are stated without proof:

Lemma 3.2. *Let $\boldsymbol{M}$ be the transition matrix for an irreducible Markov chain with period p. Then the chain described by $\boldsymbol{M}^p$ consists of exactly p aperiodic, closed communication classes.*

Lemma 3.3. *Let $\boldsymbol{M}$ describe an irreducible aperiodic Markov chain. Then, for every $n \in \mathbb{N}_{>0}$, the Markov chain described by $\boldsymbol{M}^n$ is also irreducible and aperiodic.*

Additional prerequisites can be found in the appendix, see section A.2.

With all that being covered, we can finally state and prove our main result:

Theorem 3.1 (No Hidden Markov Model with Power-Law Behavior)**.** *There is no hidden Markov model $(\boldsymbol{M_Y}, \boldsymbol{M_X})$ with with weak power-law behavior (and hence also strong power-law behavior).*

Proof. Since hidden Markov models satisfy the bulk marginal property, we can use the contraposition of theorem 1.1 to show that hidden Markov models are incapable of weak power-law behavior. Note that we can choose our starting referencing random variable freely. Hence, we may analyze $I(X_0; X_\tau)$.

First, note that we can construct the following Bayesian network with adjusted transitions depicted in figure 4.

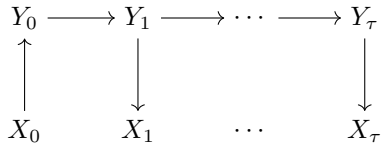

Figure 4: Adjusted Bayesian network of a hidden Markov model.

We see that $P(X_\tau = a \mid X_0 = b) = (\boldsymbol{M_X} \boldsymbol{M_Y}^\tau \boldsymbol{M_R})_{ab}$, where $\boldsymbol{M_R}$ is the transition matrix from X_0 to Y_0.

Now, for the sake of contradiction, assume that there exists a model $(\boldsymbol{M_Y}, \boldsymbol{M_X})$ with weak power-law behavior. It follows that $I(X_0; X_\tau) \xrightarrow{\tau \to \infty} 0$. We will show that for a certain $m \in \mathbb{N}$ we have $\boldsymbol{M_X} \boldsymbol{M_Y}^{m\tau} \boldsymbol{M_R} \xrightarrow{\tau \to \infty} \boldsymbol{M'}$ exponentially fast in τ, which implies exponentially fast convergence of $\boldsymbol{M_X} \boldsymbol{M_Y}^\tau \boldsymbol{M_R}$ for the subsequence $\tau := n \cdot m, n \in \mathbb{N}$. Now, either $\boldsymbol{M'}$ implies a mutual information greater than zero, but then we don't have decay towards zero and hence no power-law behavior, or we indeed have mutual information of zero, but since a subsequence converges exponentially fast to $\boldsymbol{M'}$, the mutual information cannot be lower bounded by a power-law (see corollary A.2).

Note that if $\boldsymbol{M_Y^{m\tau}}$ converges to any matrix exponentially fast for $\tau \to \infty$, then $\boldsymbol{M_X M_Y^{m\tau} M_R}$ will be forced to converge exponentially fast as well.

We differentiate the following cases based on the properties of $\boldsymbol{M_Y}$:

Case 1: Irreducible and Aperiodic
If $\boldsymbol{M_Y}$ is irreducible and aperiodic, then we can use the data processing inequality to argue that

$$I(X_0; X_\tau) \leq I(Y_0; Y_\tau) \quad .$$

Now, using the Perron-Frobenius theorem, we know that $\boldsymbol{M_Y^\tau}$ converges to a rank-one matrix exponentially fast in τ. Hence, $I(Y_0; Y_\tau)$ converges exponentially fast to zero.

Case 2: Irreducible and Periodic
Assume $\boldsymbol{M_Y}$ has periodicity p. Let's analyze $\boldsymbol{M_Y^p}$: Based on lemma 3.2, it must decompose into p aperiodic closed blocks (when ordering the states accordingly):

$$\boldsymbol{M_Y^p} = \begin{bmatrix} \boldsymbol{B_1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B_2} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{B_p} \end{bmatrix} \quad .$$

Since all blocks represent irreducible aperiodic Markov chains, $\boldsymbol{M_Y^{p\tau}}$ must converge exponentially fast. But this means that $I(X_0; X_\tau)$ converges exponentially fast for $\tau = n \cdot p,\ n \in \mathbb{N}$, and hence it cannot be lower bounded by a power-law assuming convergence to zero.

Case 3: Multiple Closed Aperiodic Communication Classes
In this case, we can order the states such that $\boldsymbol{M_Y}$ is block diagonal, i.e.

$$\boldsymbol{M_Y} = \begin{bmatrix} \boldsymbol{B_1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B_2} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{B_k} \end{bmatrix} \quad .$$

It follows that

$$\boldsymbol{M_Y^\tau} = \begin{bmatrix} \boldsymbol{B_1^\tau} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B_2^\tau} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{B_k^\tau} \end{bmatrix} \quad .$$

Hence, $\boldsymbol{M_Y^\tau}$ converges to a certain block diagonal matrix exponentially fast, since all the blocks $\boldsymbol{B_i}$ are irreducible and aperiodic.

Case 4: Multiple Closed Communication Classes
Now assume $\boldsymbol{M_Y}$ consists of many closed communication classes that can be either periodic or aperiodic. But we know that all the aperiodic classes converge exponentially fast, and the periodic ones as well if we restrict $\tau \equiv_{m_i} 0$ for a specific m_i associated with block $\boldsymbol{B_i}$. By calculating the smallest common multiple of all m_i defined as m_I, we see that $\boldsymbol{M_Y^\tau}$ converges exponentially fast for the subsequence $\tau = n \cdot m_I,\ n \in \mathbb{N}$.

Case 5: The Generic Case
Finally, we allow $\boldsymbol{M_Y}$ to consist of multiple closed and open communication classes. Let S_C denote the set of all states that are in a closed communication class, and let S_O denote the set of states in open communication classes. We also use them to refer to certain submatrices (see below). After ordering states appropriately, we have:

$$\boldsymbol{M_Y} = \begin{bmatrix} \boldsymbol{S_C} & \boldsymbol{S'_O} \\ \boldsymbol{0} & \boldsymbol{S_O} \end{bmatrix} \quad ,$$

where the blocks $\boldsymbol{S_C}$ and $\boldsymbol{S_O}$ are square. Hence:

$$\boldsymbol{M_Y^\tau} = \begin{bmatrix} \boldsymbol{S_C^\tau} & \boldsymbol{S_O'^{(\tau)}} \\ \boldsymbol{0} & \boldsymbol{S_O^\tau} \end{bmatrix} \quad .$$

Thus, the block described by $\boldsymbol{S_C}$ will converge exponentially fast for $\tau = n \cdot m,\ n \in \mathbb{N}$ for some $m \in \mathbb{N}$ based on Case 4, and $\boldsymbol{S_O^\tau}$ decays to $\boldsymbol{0}$ exponential fast as well, since we are guaranteed to leave the open communication classes at some point to stay in a closed one. Note that convergence to $\boldsymbol{0}$ is always exponentially fast.

But what about the states in $\boldsymbol{S'_O}$? Well, based on lemma 3.3 and the previous discussion, we know there exists an $m \in \mathbb{N}$ s.t. $\boldsymbol{S_C^m}$ is block diagonal with every block being irreducible and aperiodic:

$$\boldsymbol{M_Y^m} = \begin{bmatrix} \boldsymbol{B_1} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \uparrow \\ \boldsymbol{0} & \boldsymbol{B_2} & \cdots & \boldsymbol{0} & \boldsymbol{S_O'^{(m)}} \\ \vdots & \vdots & \ddots & \vdots & | \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{B_k} & \downarrow \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{S_O^m} \end{bmatrix} \quad .$$

Let's consider the submatrix $\boldsymbol{M_i}$ consisting of the states in $\boldsymbol{B_i}$ and S_O:

$$\boldsymbol{M_i} = \begin{bmatrix} \boldsymbol{B_i} & (\boldsymbol{S_O'^{(m)}})_i \\ \boldsymbol{0} & \boldsymbol{S_O^m} \end{bmatrix} \quad .$$

We see that the columns of the states in S_O match for $\boldsymbol{M_i^l}$ and $\boldsymbol{M_Y^{ml}}$ in the associated rows. Hence, we may focus on analyzing $\boldsymbol{M_i^l}$.

Since $\boldsymbol{B_i}$ is irreducible and aperiodic and $(\boldsymbol{S_O^m})^\tau \xrightarrow{\tau \to \infty} \boldsymbol{0}$ exponentially fast, we can apply lemma A.2, and see that $\boldsymbol{M_i^\tau}$ converges exponentially fast, and hence so must all entries in $\boldsymbol{M_Y^{m\tau}}$. □

3.1 Conclusions for Model Selection

Since we are interested in natural language modelling, a model with strong power-law behavior seems very desirable. However, as we just saw, a fixed-sized context window is not sufficient for power-law behavior, hence we should look out for alternatives instead.

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 is 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.

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.

A Technical Details

In this appendix we provide some further details to our arguments.

A.1 Integrating over the Quadrants of a Normal Distribution

In order to derive the formula, we first have to prove an auxiliary lemma:

Lemma A.1. *Let X and Y have a bivariate normal distribution where X and Y are standard normal variables, $X, Y \sim \mathcal{N}(0,1)$, with correlation ρ. The variable $Z = \frac{Y-\rho X}{\sqrt{1-\rho^2}}$ is a standard normal variable, and X and Z are independent.*

Proof. First, we show that Z is a standard normal variable. Since Z is a linear combination of the jointly normal variables X and Y, Z is also a normal variable. We compute its mean and variance.

The mean of Z is:

$$E[Z] = E\left[\frac{Y-\rho X}{\sqrt{1-\rho^2}}\right] = \frac{E[Y]-\rho E[X]}{\sqrt{1-\rho^2}} = \frac{0-\rho\cdot 0}{\sqrt{1-\rho^2}} = 0 \quad .$$

The variance of Z is:

$$\begin{aligned}\mathrm{Var}(Z) &= \mathrm{Var}\left(\frac{Y-\rho X}{\sqrt{1-\rho^2}}\right) = \frac{1}{1-\rho^2}\mathrm{Var}(Y-\rho X) \\ &= \frac{1}{1-\rho^2}\left(\mathrm{Var}(Y)+\rho^2\mathrm{Var}(X)-2\rho\mathrm{Cov}(X,Y)\right) \quad .\end{aligned}$$

Since X and Y are standard normal variables, $\mathrm{Var}(X) = 1$, $\mathrm{Var}(Y) = 1$, and their covariance $\mathrm{Cov}(X,Y)$ is equal to their correlation ρ. Hence:

$$\mathrm{Var}(Z) = \frac{1}{1-\rho^2}(1+\rho^2(1)-2\rho(\rho)) = \frac{1-\rho^2}{1-\rho^2} = 1 \quad .$$

Thus, Z is a standard normal variable, $Z \sim \mathcal{N}(0,1)$.

To show that X and Z are independent, we compute their covariance. Since

they are jointly normal, zero covariance implies independence.

$$\begin{aligned}\mathrm{Cov}(X,Z) &= \mathrm{Cov}\left(X, \frac{Y-\rho X}{\sqrt{1-\rho^2}}\right) = \frac{1}{\sqrt{1-\rho^2}}\mathrm{Cov}(X, Y-\rho X) \\ &= \frac{1}{\sqrt{1-\rho^2}}\left(\mathrm{Cov}(X,Y) - \rho\mathrm{Cov}(X,X)\right) \\ &= \frac{1}{\sqrt{1-\rho^2}}\left(\rho - \rho\mathrm{Var}(X)\right) = \frac{1}{\sqrt{1-\rho^2}}(\rho - \rho\cdot 1) = 0 \quad .\end{aligned}$$

Since $\mathrm{Cov}(X,Z) = 0$ and they are jointly normal, X and Z are independent. □

Now we can prove our proposition of interest:

Proposition A.1. *For bivariate standard normal variables X and Y with correlation ρ,*

$$P(X>0, Y>0) = \frac{1}{4} + \frac{1}{2\pi}\arcsin(\rho) \quad .$$

Proof. Define the random variable Z like in the previous lemma. Then, the event $\{X>0, Y>0\}$ is the same as the event $\{X>0, Z > \frac{-\rho}{\sqrt{1-\rho^2}}X\}$, where X and Z are independent standard normal variables as shown above. Writing $a := \frac{-\rho}{\sqrt{1-\rho^2}}$ for brevity, the desired probability is expressible as a double integral involving the joint density of (X,Z):

$$\begin{aligned}P(X>0, Y>0) &= P(X>0, Z>aX) \\ &= \int_{x=0}^{\infty}\int_{z=ax}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-x^2/2}\frac{1}{\sqrt{2\pi}}e^{-z^2/2}\,dz\,dx \quad .\end{aligned}$$

Switching to polar coordinates ($x = r\cos\theta, z = r\sin\theta$), the integral becomes:

$$\int_{\theta=\arctan(a)}^{\pi/2}\int_{r=0}^{\infty}\frac{1}{2\pi}e^{-r^2/2}r\,dr\,d\theta = \int_{\theta=\arctan(a)}^{\pi/2}\frac{1}{2\pi}\left[-e^{-r^2/2}\right]_0^{\infty}\,d\theta \quad .$$

This equals:

$$\int_{\theta=\arctan(a)}^{\pi/2}\frac{1}{2\pi}\,d\theta = \frac{1}{2\pi}\left(\frac{\pi}{2} - \arctan(a)\right) = \frac{1}{4} - \frac{1}{2\pi}\arctan\left(\frac{-\rho}{\sqrt{1-\rho^2}}\right) \quad .$$

Using the fact that the arctan function is odd, i.e. $\arctan(-u) = -\arctan(u)$, we get:

$$\frac{1}{4} + \frac{1}{2\pi}\arctan\left(\frac{\rho}{\sqrt{1-\rho^2}}\right) \quad .$$

To finish, we use the identity $\arcsin(\rho) = \arctan\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)$. To see this, let $\phi = \arcsin(\rho)$ for $\phi \in [-\pi/2, \pi/2]$. Then $\sin(\phi) = \rho$ and $\cos(\phi) = \sqrt{1-\rho^2}$. Thus, $\tan(\phi) = \frac{\sin(\phi)}{\cos(\phi)} = \frac{\rho}{\sqrt{1-\rho^2}}$, which implies $\phi = \arctan\left(\frac{\rho}{\sqrt{1-\rho^2}}\right)$. Substituting this into our expression gives the final result:

$$P(X > 0, Y > 0) = \frac{1}{4} + \frac{1}{2\pi} \arcsin(\rho) \quad .$$

□

A.2 Prerequisites for Theorem 3.1

Lemma A.2. *Let*

$$\boldsymbol{M} \coloneqq \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{C} \end{bmatrix}$$

be a matrix consisting of submatrices $\boldsymbol{A} \in \mathbb{R}^{k \times k}$, $\boldsymbol{B} \in \mathbb{R}^{k \times \ell}$, and $\boldsymbol{C} \in \mathbb{R}^{\ell \times \ell}$. Let $\boldsymbol{A}$ be an irreducible aperiodic Markov transition matrix, and let $\boldsymbol{C}^n \xrightarrow{n \to \infty} \boldsymbol{0}$ with exponential decay. Then, $\boldsymbol{M}^n$ decays exponentially in n towards a matrix $\boldsymbol{M}'$.

Proof. The proof proceeds in three steps. First, we establish a closed-form expression for $\boldsymbol{M}^n$. Second, we determine the limit matrix $\boldsymbol{M}'$. Third, we prove that the convergence to this limit is exponential.

1. The Form of $\boldsymbol{M}^n$ Since $\boldsymbol{M}$ is a block upper triangular matrix, its powers take a specific form. By induction, we can show that:

$$\boldsymbol{M}^n = \begin{bmatrix} \boldsymbol{A}^n & \boldsymbol{X}_n \\ \boldsymbol{0} & \boldsymbol{C}^n \end{bmatrix} \quad \text{where} \quad \boldsymbol{X}_n = \sum_{j=0}^{n-1} \boldsymbol{A}^{n-1-j} \boldsymbol{B} \boldsymbol{C}^j \quad .$$

2. The Limit Matrix $\boldsymbol{M}'$ We analyze the limit of each block of $\boldsymbol{M}^n$ as $n \to \infty$.

- **Block $\boldsymbol{A}^n$:** Since $\boldsymbol{A}$ is an irreducible aperiodic Markov transition matrix, the Perron-Frobenius theorem for stochastic matrices guarantees that $\boldsymbol{A}^n$ converges to a rank-one matrix $\boldsymbol{A}' = \boldsymbol{\pi}\boldsymbol{1}^T$, where $\boldsymbol{\pi}$ is the unique stationary distribution. The convergence is exponential, so there exist constants $K_A > 0$ and $0 \leq \lambda < 1$ such that $\|\boldsymbol{A}^n - \boldsymbol{A}'\| \leq K_A \lambda^n$.

- **Block $\boldsymbol{C}^n$:** By hypothesis, $\boldsymbol{C}^n \to \mathbf{0}$ with exponential decay. This is equivalent to its spectral radius being less than one, $\rho(\boldsymbol{C}) < 1$. Thus, there exist constants $K_C > 0$ and $0 \leq \gamma < 1$ such that $\|\boldsymbol{C}^n\| \leq K_C \gamma^n$.
- **Block $\boldsymbol{X}_n$:** Let $\boldsymbol{E}_k = \boldsymbol{A}^k - \boldsymbol{A}'$. We have $\|\boldsymbol{E}_k\| \leq K_A \lambda^k$. We can rewrite $\boldsymbol{X}_n$ as:

$$\boldsymbol{X}_n = \sum_{j=0}^{n-1} (\boldsymbol{A}' + \boldsymbol{E}_{n-1-j}) \boldsymbol{B} \boldsymbol{C}^j = \boldsymbol{A}' \boldsymbol{B} \left(\sum_{j=0}^{n-1} \boldsymbol{C}^j \right) + \sum_{j=0}^{n-1} \boldsymbol{E}_{n-1-j} \boldsymbol{B} \boldsymbol{C}^j \quad .$$

  As $n \to \infty$, the first term converges to $\boldsymbol{A}'\boldsymbol{B}(\boldsymbol{I} - \boldsymbol{C})^{-1}$, since the series $\sum_{j=0}^{\infty} \boldsymbol{C}^j$ converges to $(\boldsymbol{I} - \boldsymbol{C})^{-1}$. The second term converges to $\mathbf{0}$ because its norm is bounded by a vanishing convolution sum. Thus, the limit of $\boldsymbol{X}_n$ is $\boldsymbol{X}' = \boldsymbol{A}'\boldsymbol{B}(\boldsymbol{I} - \boldsymbol{C})^{-1}$.

Combining these limits, the limit matrix is $\boldsymbol{M}' = \left[\begin{smallmatrix} \boldsymbol{A}' & \boldsymbol{X}' \\ \mathbf{0} & \mathbf{0} \end{smallmatrix} \right]$.

3. Exponential Rate of Convergence We analyze the norm of the difference matrix $\boldsymbol{M}^n - \boldsymbol{M}'$:

$$\boldsymbol{M}^n - \boldsymbol{M}' = \begin{bmatrix} \boldsymbol{A}^n - \boldsymbol{A}' & \boldsymbol{X}_n - \boldsymbol{X}' \\ \mathbf{0} & \boldsymbol{C}^n \end{bmatrix} \quad .$$

The blocks $\|\boldsymbol{A}^n - \boldsymbol{A}'\|$ and $\|\boldsymbol{C}^n\|$ decay exponentially by definition. We examine the convergence of the off-diagonal block:

$$\boldsymbol{X}_n - \boldsymbol{X}' = -\boldsymbol{A}'\boldsymbol{B} \left(\sum_{j=n}^{\infty} \boldsymbol{C}^j \right) + \sum_{j=0}^{n-1} \boldsymbol{E}_{n-1-j} \boldsymbol{B} \boldsymbol{C}^j \quad .$$

The norm of each part is bounded by an exponentially decaying function:

- $\left\| -\boldsymbol{A}'\boldsymbol{B} \left(\sum_{j=n}^{\infty} \boldsymbol{C}^j \right) \right\| \leq \|\boldsymbol{A}'\| \|\boldsymbol{B}\| \sum_{j=n}^{\infty} \|\boldsymbol{C}^j\| \leq \|\boldsymbol{A}'\| \|\boldsymbol{B}\| K_C \frac{\gamma^n}{1-\gamma}$. This decays with rate γ.
- $\left\| \sum_{j=0}^{n-1} \boldsymbol{E}_{n-1-j} \boldsymbol{B} \boldsymbol{C}^j \right\| \leq \sum_{j=0}^{n-1} K_A \lambda^{n-1-j} \|\boldsymbol{B}\| K_C \gamma^j = K_A \|\boldsymbol{B}\| K_C \sum_{j=0}^{n-1} \lambda^{n-1-j} \gamma^j$. This convolution sum is bounded by $K' n \mu^n$ where $\mu = \max(\lambda, \gamma)$, which decays exponentially.

Since $\|\boldsymbol{X}_n - \boldsymbol{X}'\|$ is bounded by a sum of exponentially decaying terms, it also decays exponentially. As all blocks of $\boldsymbol{M}^n - \boldsymbol{M}'$ converge to zero exponentially, the norm $\|\boldsymbol{M}^n - \boldsymbol{M}'\|$ does as well. This completes the proof. □

A.2.1 Convergence of Mutual Information

Theorem A.1 (Element-Wise Exponential Convergence Implies Exponential Convergence)**.** *Let $f : \mathcal{D} \to \mathbb{R}^m$ be a function defined on a convex domain $\mathcal{D} \subseteq \mathbb{R}^n$ that is a Cartesian product of real intervals, i.e., $\mathcal{D} = \mathcal{D}_1 \times \mathcal{D}_2 \times \cdots \times \mathcal{D}_n$ where each $\mathcal{D}_i \subseteq \mathbb{R}$ is an interval. Let $\{\boldsymbol{x}_k\}_{k=1}^{\infty} \subset \mathcal{D}$ be a sequence converging exponentially fast to $\boldsymbol{x}_0 \in \mathcal{D}$.*

Let $\boldsymbol{e}_j$ denote the j-th standard basis vector in $\mathbb{R}^n$. Suppose that for each input coordinate $j \in \{1, 2, \ldots, n\}$ there exist functions $K_j(C', \rho)$, $C_j(C', \rho)$, $P_j(C', \rho)$ s.t. for every sequence $\{\boldsymbol{u}_k\}_{k=1}^{\infty} \subset \mathcal{D}$ converging to $\boldsymbol{u}_0$ where the difference $\boldsymbol{u}_\ell - \boldsymbol{u}_{\ell'}$ is parallel to $\boldsymbol{e}_j$ (i.e., they only differ in the j-th coordinate) that satisfies $|\boldsymbol{u}_0 - \boldsymbol{u}_k| \leq C'\rho^k$ for all k and some $\rho \in [0, 1)$, $C' > 0$, we have for all $k \geq K_j(C', \rho)$:

$$\|f(\boldsymbol{u}_0) - f(\boldsymbol{u}_k)\| \leq C_j(C', \rho)\rho^k k^{P_j(C', \rho)} \quad .$$

Then, there exist constants $C > 0$ and $\sigma \in [0, 1)$ such that for all sufficiently large k:

$$\|f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k)\| \leq C\sigma^k \quad .$$

Proof. Let the sequence $\{\boldsymbol{x}_k\}_{k=1}^{\infty} \subset \mathcal{D}$ converge exponentially to $\boldsymbol{x}_0 \in \mathcal{D}$. By definition, there exist constants $C_x > 0$ and $\rho \in [0, 1)$ such that for all k,

$$\|\boldsymbol{x}_k - \boldsymbol{x}_0\| \leq C_x\rho^k \quad .$$

Let $\boldsymbol{x}_k = (x_{k,1}, \ldots, x_{k,n})^T$ and $\boldsymbol{x}_0 = (x_{0,1}, \ldots, x_{0,n})^T$. An immediate consequence is that each coordinate also converges exponentially, i.e., for each $j \in \{1, \ldots, n\}$:

$$|x_{k,j} - x_{0,j}| \leq \|\boldsymbol{x}_k - \boldsymbol{x}_0\|_\infty \leq \|\boldsymbol{x}_k - \boldsymbol{x}_0\| \leq C_x\rho^k \quad ,$$

where we use the equivalence of norms in $\mathbb{R}^n$.

To bound $\|f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k)\|$, we define a sequence of $n + 1$ intermediate points that form a path from $\boldsymbol{x}_k$ to $\boldsymbol{x}_0$ by changing one coordinate at a time. For each k, let:

$$\begin{aligned}
\boldsymbol{z}_{k,0} &:= \boldsymbol{x}_k = (x_{k,1}, x_{k,2}, \ldots, x_{k,n}) \\
\boldsymbol{z}_{k,1} &:= (x_{0,1}, x_{k,2}, \ldots, x_{k,n}) \\
&\vdots \\
\boldsymbol{z}_{k,j} &:= (x_{0,1}, \ldots, x_{0,j}, x_{k,j+1}, \ldots, x_{k,n}) \\
&\vdots \\
\boldsymbol{z}_{k,n} &:= (x_{0,1}, \ldots, x_{0,n}) = \boldsymbol{x}_0 \quad .
\end{aligned}$$

Since $\mathcal{D}$ is a cartesian product intervals and both $\boldsymbol{x}_k$ and $\boldsymbol{x}_0$ are in $\mathcal{D}$, all intermediate points $\boldsymbol{z}_{k,j}$ are also contained in $\mathcal{D}$. We can express the total difference $f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k)$ as a telescoping sum:

$$f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k) = f(\boldsymbol{z}_{k,n}) - f(\boldsymbol{z}_{k,0}) = \sum_{j=1}^{n} \left(f(\boldsymbol{z}_{k,j}) - f(\boldsymbol{z}_{k,j-1}) \right) \quad .$$

By the triangle inequality, we have:

$$\|f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k)\| \leq \sum_{j=1}^{n} \|f(\boldsymbol{z}_{k,j}) - f(\boldsymbol{z}_{k,j-1})\| \quad .$$

Now, we analyze each term $\|f(\boldsymbol{z}_{k,j}) - f(\boldsymbol{z}_{k,j-1})\|$ for a fixed $j \in \{1, \ldots, n\}$. The points $\boldsymbol{z}_{k,j}$ and $\boldsymbol{z}_{k,j-1}$ differ only in their j-th coordinate.

Let us define a sequence $\{\boldsymbol{u}_m\}_{m=1}^{\infty}$ and a limit point $\boldsymbol{u}_0$ that fit the condition in the theorem's hypothesis. For the given j and k, let

$$\begin{aligned} \boldsymbol{u}_m &:= (x_{0,1}, \ldots, x_{0,j-1}, x_{m,j}, x_{k,j+1}, \ldots, x_{k,n}) \\ \boldsymbol{u}_0 &:= (x_{0,1}, \ldots, x_{0,j-1}, x_{0,j}, x_{k,j+1}, \ldots, x_{k,n}) \quad . \end{aligned}$$

Note that $\boldsymbol{u}_0 = \boldsymbol{z}_{k,j}$ and by setting $m = k$, we get $\boldsymbol{u}_k = \boldsymbol{z}_{k,j-1}$. The sequence $\{\boldsymbol{u}_m\}$ lies on a line parallel to the j-th coordinate axis. As $m \to \infty$, $\boldsymbol{u}_m \to \boldsymbol{u}_0$ because $x_{m,j} \to x_{0,j}$. The convergence is exponential:

$$\|\boldsymbol{u}_m - \boldsymbol{u}_0\| = |x_{m,j} - x_{0,j}| \leq C_x \rho^m \quad .$$

The hypothesis states that for any such sequence, there exist constants $K_j(C_x, \rho)$, $C_j(C_x, \rho)$, $P_j(C_x, \rho)$ which are independent of the specific line, such that for all $k \geq K_j(C_x, \rho)$ we have $\|f(\boldsymbol{u}_0) - f(\boldsymbol{u}_m)\| \leq C_j(C_x, \rho)\rho^m m^{P_j(C_x,\rho)}$. Applying this for $m = k$:

$$\|f(\boldsymbol{z}_{k,j}) - f(\boldsymbol{z}_{k,j-1})\| = \|f(\boldsymbol{u}_0) - f(\boldsymbol{u}_k)\| \leq C_j(C_x, \rho)\rho^k k^{P_j(C_x,\rho)} \quad .$$

This inequality holds for each $j = 1, \ldots, n$. Substituting these bounds back into the sum:

$$\|f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k)\| \leq \sum_{j=1}^{n} C_j(C_x, \rho)\rho^k k^{P_j(C_x,\rho)} \quad .$$

Let $K := \max_{j \in \{1,\ldots,n\}} \{K_j(C_x, \rho)\}$, $C := \sum_{j=1}^{n} C_j(C_x, \rho)$ and $P := \max_{j \in \{1,\ldots,n\}} \{P_j(C_x, \rho)\}$. Hence, for all $k \geq K$:

$$\|f(\boldsymbol{x}_0) - f(\boldsymbol{x}_k)\| \leq \sum_{j=1}^{n} C_j(C_x, \rho)\rho^k k^P = \left(\sum_{j=1}^{n} C_j(C_x, \rho) \right) \rho^k k^P = C\rho^k k^P \quad .$$

This shows that $\{f(\boldsymbol{x}_k)\}$ converges exponentially to $f(\boldsymbol{x}_0)$ with a rate of σ s.t. $\rho < \sigma < 1$ (like $\sigma := \sqrt{\rho}$). This completes the proof. □

Lemma A.3. *Let the function $f : [0, \infty) \to \mathbb{R}$ be defined as $f(x) = x \log x$ for $x \in (0, 1)$, and $f(x) = 0$ everywhere else. If a sequence $\{x_k\}_{k=1}^{\infty} \subset [0, 1]$ converging to a limit $x_\infty \in [0, 1]$ satisfies $|x_k - x_\infty| \leq C\rho^k$ for some $C \in \mathbb{R}_{>0}$, $\rho \in [0, 1)$, then the sequence $\{f(x_k)\}$ converges to $f(x_\infty)$ with $|f(x_k) - f(x_\infty)| \leq C\rho^k \ |\log C + k \log \rho|$ for $k \geq \log_\rho \left(\frac{e^{-1}}{C} \right) =: K$.*

Proof. We consider two cases for the limit x_∞:

Case 1: $x_\infty = 0$
In this case, $|x_k - 0| = x_k \leq C\rho^k$. We want to bound the difference $|f(x_k) - f(0)| = |f(x_k)|$. Note that $|f(x)|$ is monotonically increasing on $[0, e^{-1}]$. For $k \geq K$ we have $x_k \leq C\rho^k \leq e^{-1}$. Thus, for all $k \geq K$:

$$|f(x_k)| \leq |f(C\rho^k)| = |C\rho^k \log(C\rho^k)| = C\rho^k \ |\log C + k \log \rho| \quad .$$

Case 2: $x_\infty > 0$
Similarly, for $k \geq K$ we have $|x_k - x_\infty| \leq C\rho^k \leq e^{-1}$. Based on the function graph, it follows that for $k \geq K$ we have:

$$|f(x_k) - f(x_\infty)| \leq |f(C\rho^k) - f(0)| = |f(C\rho^k)| = C\rho^k \ |\log C + k \log \rho| \quad .$$

□

Theorem A.2 (Element-Wise Exponential Convergence Property of Mutual Information)**.** *Let the function $f : [0, \infty) \to \mathbb{R}$ be defined as $f(x) = x \log x$ for $x \in (0, 1)$, and $f(x) = 0$ everywhere else. Define the function $I : [0, 1]^{m \times n} \to \mathbb{R}$ for a matrix $\boldsymbol{M} \in [0, 1]^{m \times n}$ as:*

$$I(\boldsymbol{M}) = \sum_{i=1}^{m} \sum_{j=1}^{n} f(M_{ij}) - \sum_{i=1}^{m} f\left(\sum_{j=1}^{n} M_{ij} \right) - \sum_{j=1}^{n} f\left(\sum_{i=1}^{m} M_{ij} \right) \quad .$$

This function exhibits element-wise exponential convergence. That is, for any single component (i_0, j_0), if a sequence of matrices $\{\boldsymbol{U}_k\}_{k=1}^{\infty} \subset [0, 1]^{m \times n}$ converges to a limit $\boldsymbol{U}_\infty$, varies only in the (i_0, j_0)-th component and satisfies $\|\boldsymbol{U}_k - \boldsymbol{U}_\infty\| \leq C\rho^k$ for some $C > 0$, $\rho \in [0, 1)$ and all k, then the sequence of values $\{I(\boldsymbol{U}_k)\}$ converges to $I(\boldsymbol{U}_\infty)$ with $|I(\boldsymbol{U}_k) - I(\boldsymbol{U}_\infty)| \leq C'(C, \rho)\rho^k k^{P(C,\rho)}$ for all $k \geq K(C, \rho)$.

Proof. The function $I(\boldsymbol{M})$ is a sum of terms involving f applied to the matrix entries and their row and column sums. Let $m_i(\boldsymbol{M}) = \sum_j M_{ij}$ and $m'_j(\boldsymbol{M}) = \sum_i M_{ij}$. We have:

$$I(\boldsymbol{M}) = \sum_{i,j} f(M_{ij}) - \sum_i f(m_i(\boldsymbol{M})) - \sum_j f(m'_j(\boldsymbol{M})) \quad .$$

We are given a sequence $\{\boldsymbol{U}_k\}$ that varies only in the (i_0, j_0)-th component, $u_k = \boldsymbol{U}_k(i_0, j_0)$. All other components are constant. The exponential convergence of $\{\boldsymbol{U}_k\}$ means $|u_k - u_\infty| \leq C\rho^k$.

The difference $I(\boldsymbol{U}_k) - I(\boldsymbol{U}_\infty)$ consists only of terms whose arguments change with k. These are:

1. The entry term: $f(u_k)$.
2. The row-sum term: $f(m_{i_0}(\boldsymbol{U}_k))$, where $m_{i_0}(\boldsymbol{U}_k) = u_k + \text{const}$.
3. The column-sum term: $f(m'_{j_0}(\boldsymbol{U}_k))$, where $m'_{j_0}(\boldsymbol{U}_k) = u_k + \text{const}$.

By the triangle inequality, the total error is bounded by the sum of the absolute errors of these three terms:

$$\begin{aligned} |I(\boldsymbol{U}_k) - I(\boldsymbol{U}_\infty)| \leq\ & |f(u_k) - f(u_\infty)| \\ & + |f(m_{i_0}(\boldsymbol{U}_k)) - f(m_{i_0}(\boldsymbol{U}_\infty))| \\ & + |f(m'_{j_0}(\boldsymbol{U}_k)) - f(m'_{j_0}(\boldsymbol{U}_\infty))| \quad . \end{aligned}$$

The arguments to the function f in each of these three terms converge exponentially to their limits with rate ρ and constant C, since $|m_{i_0}(\boldsymbol{U}_k) - m_{i_0}(\boldsymbol{U}_\infty)| = |u_k - u_\infty|$ and $|m'_{j_0}(\boldsymbol{U}_k) - m'_{j_0}(\boldsymbol{U}_\infty)| = |u_k - u_\infty|$.

Hence, by lemma A.3, we have:

$$\begin{aligned} |I(\boldsymbol{U}_k) - I(\boldsymbol{U}_\infty)| &\leq 3C\rho^k \ |\log C + k \log \rho| \\ &\leq 3C\rho^k k \ (|\log C| + |\log \rho|) \\ &= C' \rho^k k \quad , \end{aligned}$$

with $C' \coloneqq 3C(|\log C| + |\log \rho|)$ and for all $k \geq K(C, \rho)$. Note that K, C' and P only depend on C and ρ. □

Corollary A.1. *Using theorem A.1, we see that if a sequence $\{\boldsymbol{P}_k\}$ of joint probability distributions converges exponentially fast, then $\{I(\boldsymbol{P}_k)\}$ converges exponentially fast as well.*

Corollary A.2. *A joint probability matrix $\boldsymbol{P}_{X,Y}$ can be calculated from the conditional probability matrix $\boldsymbol{P}_{Y|X}$ and the diagonal matrix $\boldsymbol{P}_X$ with the probabilities for X on its diagonal using $\boldsymbol{P}_{X,Y} = \boldsymbol{P}_{Y|X}\boldsymbol{P}_X$. Hence, if $\boldsymbol{P}_{Y|X}$ converges exponentially fast while $\boldsymbol{P}_X$ stays constant, the mutual information $I(X;Y)$ will converge exponentially fast as well.*

References

[1] Henry W. Lin and Max Tegmark. Critical behavior in physics and probabilistic formal languages. *Entropy*, 19(7), 2017.