1 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 **??**.

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

1.1 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 1.1 (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 1.2 (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.*

1.1.1 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 1.1, 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 1.2, 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.

1.1.2 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:

1.1.3 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 1.1. *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 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 $x > -1 \iff -2c\tau^{-\alpha} > -1 \iff \tau > (2c)^{\frac{1}{\alpha}}$.

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.

1.2 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 1.1 (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 a model $S_{n,\mathbf{\Sigma}_n}$ over $\{-1, 1\}$ with

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

where Q_w is the quadrant

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

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:

1.2.1 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 **??**.

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 1.1 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 neat formula:

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

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

Now, set $\delta_\tau \coloneqq \frac{\arcsin(\rho_\tau)}{2\pi}$ 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 1.2. *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 1.2 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 1.3. *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 1.3 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.

1.3 Summary

Let's summarize our findings in a concise theorem:

Theorem 1.1 (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*

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

We use $\mathbf{\Sigma}_n$ *as the parameter covariance matrix of the model defined in definition 1.1. It follows that* $S_{n,\mathbf{\Sigma}_n}$ *is a valid model over* $\{-1, 1\}$*, since* $\mathbf{\Sigma}_n$ *is positive definite especially. Furthermore,* S *has strong power-law behavior according to definition* **??***. 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. □