1 Model Framework

We no longer focus on Markov chains, so the associated symbols like S and n no longer carry the same meaning. We will redefine them shortly. Also, in order for the polynomials to be well defined later, we will constrain $\tau > 0$.

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 τ 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 and hidden Markov models have the bulk marginal property.

Lemma 1.1. *For every $d \in \mathbb{N}$, let $I \coloneqq [n+d] \setminus [n] = \{n+1, \dots, n+d\}$. 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

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

which concludes the induction step. $\square$

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:

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

$\square$

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 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. The decay of $I(X_t, X_{t+\tau})$ follows from S having upper bound power-law behavior (or decaying 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 potentially decays to 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 $\tau \in \mathbb{N}$:

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