1 Model Framework

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

The tokens, represented by random variables X_t, are elements of a finite alphabet Σ. The distance between X_t and $X_{t+\tau}$ is τ, and for every t and every 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, the first conclusion is every model S must define a probability measure over Σ^n for every $n \in \mathbb{N}$.

As a first formalization, S is a function $S : (n, w) \mapsto [0, 1]$, for $n \in \mathbb{N},\ w \in \Sigma^n$ s.t. $\sum_{w\in\Sigma^n} S(n, w) = 1$.

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 1.1. We say S is *well behaved* iff for every $n \in \mathbb{N},\ w \in \Sigma^{n+1}$ it holds true that

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

Lemma 1.1. *For every $d \in \mathbb{N}$, let $I := [n + d] \setminus [n] = \{n + 1, \dots, n + d\}$. Then, if S is well behaved, 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 S being well behaved. Thus, assume the claim holds for some $d := k$.

Then we have

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

which concludes the induction step. ☐

Definition 1.2. From any model S, we can construct the *induced well behaved* model S^* by defining S_n^* recursively as

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

Lemma 1.2. *The induced well behaved model S^* is indeed well behaved.*

Proof. We have:

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

☐

Now, we want to look at how we might restrict our model $(S_n)_{n\in\mathbb{N}} \equiv S$. One approach might be to define a model structure for every $n \in \mathbb{N}$ with finite parameters $\boldsymbol{\theta}_n$, thus $S_n \in \{S_n(\boldsymbol{\theta}_n) : \boldsymbol{\theta}_n \in \Theta_n\} =: \mathcal{S}_n$. 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 $(\mathcal{S}_n, \boldsymbol{\theta}_n)_{n\in\mathbb{N}}$. We call $\mathcal{S}(n) \equiv \mathcal{S}_n$ the *model space.*

Definition 1.3. 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 the mutual information between *any* two variables X_t, $X_{t+\tau}$ to be $I(X_t, X_{t+\tau}) \in \Omega(\tau^{-\alpha})$, i.e. it has to hold for every t and *every* S_n. Thus, let us define the following:

Definition 1.4. We define $I_S(n)$ te be the minimal mutual information between any two variables of S_n. Formally, let X_i, X_j be random variables with distributions defined by S_n. Then:

$$I_S(n) \coloneqq \min_{(i,j)\in[n]^2} I(X_i, X_j) \quad .$$

Theorem 1.1. *Let S be a model. If $I_S(n) \geq cn^{-\alpha}$ for some $c, \alpha \in \mathbb{R}_{>0}$, then for every $X_t, X_{t+\tau}$ over any S_n ($n \geq t+\tau$), it follows that $I(X_t, X_{t+\tau}) \geq c\tau^{-\alpha}$.*

Proof. Assume there exists t, τ and S_n s.t. $I(X_t, X_{t+\tau}) < c\tau^{-\alpha}$. But then, by the definition of $I_S(n)$, we must have $I_S(\tau) \leq I(X_t, X_{t+\tau}) < c\tau^{-\alpha}$, a contradiction. □

Based on this result, we define:

Definition 1.5. A model space $\mathcal{S}$ has power-law capacity iff there exists an $S \in \mathcal{S}$ s.t. $I_S(n) \in \Omega(n^{-\alpha})$ for some $\alpha \in \mathbb{R}_{>0}$ and $I_S(n) \xrightarrow{n\to\infty} 0$.