1 No Power-Law in Hidden Markov Models

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

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

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

Figure 1: Bayesian network of Markov chains. All arrows go from previous tokens to future tokens.

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

From a modelling perspective, it is very reasonable to reuse the same transitions over time, as we cannot have infinitely many ”hard-coded” transitions (but we can use different models whith implicitely infitite transitions). Furthermore, for the same *mode of transition*, which we define as the ”arrow structure” of all ingoing edges into the current node in the Bayesian network (in figure 1 the mode of transition would be from the current token to the next), it seems reasonable to assume invariance in time, i.e. fixed transition probabilities. We call such transitions to be *constant*.

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

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

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

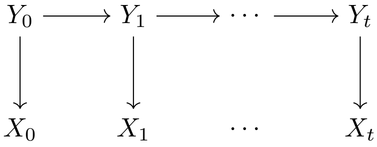

Figure 2: Bayesian network of a hidden Markov model.

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

Theorem 1.1 (No Hidden Markov Model with Power-Law decay)**.** *There is no hidden Markov model* $(\boldsymbol{M_Y}, \boldsymbol{M_X})$ *with* $I(X_{t_0}, X_{t_0+\tau}) \in \mathcal{O}(\tau^{-\alpha})$ *and* $I(X_{t_0}, X_{t_0+\tau}) \in \Omega(\tau^{-\beta})$ *for some* $\alpha, \beta \in \mathbb{R}_{>0}$*.*

Proof. Set $t := t_0$. First, note that $P(X_{t+\tau} = b \mid Y_t = c) = (\boldsymbol{M_X}\boldsymbol{M_Y^\tau})_{bc}$. Furthermore,

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

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

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

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

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

Thus, we get:

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

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

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

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

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

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

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

1.1 Conclusions for Model Selection

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

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

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

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

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

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