

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 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 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 which implicitly infitite transitions). Furthermore, for the same *mode of transition*, which we define as the "arrow structure" 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 for a single constant mode of transition based on the characteristics of natural language.

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 $Y_t \rightarrow Y_{t+1}$ 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:

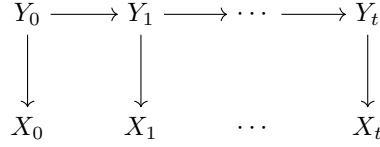


Figure 2: Bayesian network of a hidden Markov model.

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 $(\mathbf{M}_Y, \mathbf{M}_X)$ 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. First, note that $P(X_{t+\tau} = b \mid Y_t = c) = (\mathbf{M}_X \mathbf{M}_Y^\tau)_{bc}$. Furthermore,

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

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

$$\begin{aligned}
 P(X_{t+\tau} = b \mid X_t = a) &= \sum_{c \in S_Y} P(Y_t = c \mid X_t = a) P(X_{t+\tau} = b \mid Y_t = c) \\
 &= \sum_{c \in S_Y} (\mathbf{M}_X)_{ac} \frac{P(Y_t = c)}{P(X_t = a)} (\mathbf{M}_X \mathbf{M}_Y^\tau)_{bc}
 \end{aligned}$$

□