1 No Power-Law Behavior 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 time-homogenous Markov chains $P(X_{t+1} = a \mid X_t = b)$ is independent of t and hence is constant over time. Thus, all the arrows in figure 1 represent the same transition.

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 with implicitly infinite transitions). Furthermore, when making a prediction of the next token given a fixed context window, it seems reasonable to assume invariance in time, i.e. fixed transition probabilities similar to time-homogenous Markov chains.

Now, the question is, can we achieve power-law behavior with an arbitrarily large fixed-sized context window with fixed transition probabilities?

In order to answer this question, let us first reduce our setting to a simpler model, specifically to the already mentioned time-homogenous Markov chains:

The idea is to employ a hidden variable $Y \in \Sigma^{s+1}$, where Σ is the alphabet, and s is the size of the context window (for Markov chains $s = 1$). Clearly, Y captures the entire *state* at time t of our model, that is all the previous s tokens and the current one, and we can model the transitions $Y_t \to Y_{t+1}$ as simple time-homogenous Markov chain transitions (and hence invariant in time). And, of course, once we know Y_t, we also know X_t (which of course can be modelled with time-homogenous Markov chain transitions as well). These models are known as *hidden Markov models*.

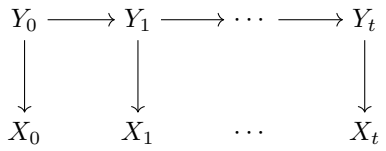

Figure 2: Bayesian network of a hidden Markov model.

Unfortunately, there is no free lunch, as we will see in the following. However, in order to use theorem **??** to disprove weak and thus strong power-law behavior, we need to prove the bulk marginal property of hidden Markov models first:

Lemma 1.1 (Hidden Markov Models have the Bulk Marginal Property)**.** *Every hidden Markov model with finite state spaces (S_Y, S_X) for its latent variable Y and observable variable X with transition matrices $(\boldsymbol{M_Y}, \boldsymbol{M_X})$ complies with the bulk marginal property.*

Proof. Let $w_i \coloneqq X_{i-1}$ for $n \in [n+1]$. Then we have:

$$
\begin{aligned}
&\sum_{w_{n+1} \in \Sigma} S_{n+1}(w) \\
&\underset{\text{Bayesian network}}{=} \sum_{w_{n+1} \in \Sigma} \sum_{q_1, \dots, q_{n+1} \in S_Y} \left[P(q_1) \prod_{i=2}^{n+1} P(q_i \mid q_{i-1}) \cdot \prod_{i=1}^{n+1} P(w_i \mid q_i) \right] \\
&= \sum_{q_1, \dots, q_{n+1} \in S_Y} \sum_{w_{n+1} \in \Sigma} \left[P(q_1) \prod_{i=2}^{n+1} P(q_i \mid q_{i-1}) \cdot \prod_{i=1}^{n+1} P(w_i \mid q_i) \right] \\
&= \sum_{q_1, \dots, q_{n+1} \in S_Y} \left[P(q_1) \prod_{i=2}^{n+1} P(q_i \mid q_{i-1}) \cdot \prod_{i=1}^{n} P(w_i \mid q_i) \sum_{w_{n+1} \in \Sigma} P(w_{n+1} \mid q_{n+1}) \right] \\
&= \sum_{q_1, \dots, q_{n+1} \in S_Y} \left[P(q_1) \prod_{i=2}^{n+1} P(q_i \mid q_{i-1}) \cdot \prod_{i=1}^{n} P(w_i \mid q_i) \right] \\
&= \sum_{q_1, \dots, q_n \in S_Y} \sum_{q_{n+1} \in S_Y} P(q_{n+1} \mid q_n) \left[P(q_1) \prod_{i=2}^{n} P(q_i \mid q_{i-1}) \cdot \prod_{i=1}^{n} P(w_i \mid q_i) \right] \\
&= \sum_{q_1, \dots, q_n \in S_Y} \left[P(q_1) \prod_{i=2}^{n} P(q_i \mid q_{i-1}) \cdot \prod_{i=1}^{n} P(w_i \mid q_i) \sum_{q_{n+1} \in S_Y} P(q_{n+1} \mid q_n) \right] \\
&= \sum_{q_1, \dots, q_n \in S_Y} \left[P(q_1) \prod_{i=2}^{n} P(q_i \mid q_{i-1}) \cdot \prod_{i=1}^{n} P(w_i \mid q_i) \right] \\
&\overset{\checkmark}{=} S_n(w_{-\{n+1\}}) \quad .
\end{aligned}
$$

□

We will also make use of the following well-established results regarding time-homogenous Markov chains. Hence, they are stated without proof:

Lemma 1.2. *Let $\boldsymbol{M}$ be the transition matrix for an irreducible Markov chain with period p. Then the chain described by $\boldsymbol{M}^p$ consists of exactly p aperiodic, closed communication classes.*

Lemma 1.3. *Let $\boldsymbol{M}$ describe an irreducible aperiodic Markov chain. Then, for every $n \in \mathbb{N}_{>0}$, the Markov chain described by $\boldsymbol{M}^n$ is also irreducible and aperiodic.*

Additional prerequisites can be found in the appendix, see section **??**.

With all that being covered, we can finally state and prove our main result:

Theorem 1.1 (No Hidden Markov Model with Power-Law Behavior)**.** *There is no hidden Markov model $(\boldsymbol{M_Y}, \boldsymbol{M_X})$ with with weak power-law behavior (and hence also strong power-law behavior).*

Proof. Since hidden Markov models satisfy the bulk marginal property, we can use the contraposition of theorem **??** to show that hidden Markov models are incapable of weak power-law behavior. Note that we can choose our starting referencing random variable freely. Hence, we may analyze $I(X_0; X_\tau)$.

First, note that we can construct the following Bayesian network with adjusted transitions depicted in figure 3.

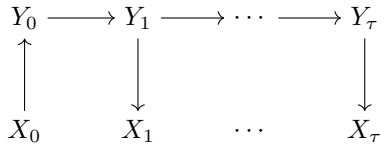

Figure 3: Adjusted Bayesian network of a hidden Markov model.

We see that $P(X_\tau = a \mid X_0 = b) = (\boldsymbol{M_X} \boldsymbol{M_Y}^\tau \boldsymbol{M_R})_{ab}$, where $\boldsymbol{M_R}$ is the transition matrix from X_0 to Y_0.

Now, for the sake of contradiction, assume that there exists a model $(\boldsymbol{M_Y}, \boldsymbol{M_X})$ with weak power-law behavior. It follows that $I(X_0; X_\tau) \xrightarrow{\tau \to \infty} 0$. We will show that for a certain $m \in \mathbb{N}$ we have $\boldsymbol{M_X} \boldsymbol{M_Y}^{m\tau} \boldsymbol{M_R} \xrightarrow{\tau \to \infty} \boldsymbol{M'}$ exponentially fast in τ, which implies exponentially fast convergence of $\boldsymbol{M_X} \boldsymbol{M_Y}^\tau \boldsymbol{M_R}$ for the subsequence $\tau := n \cdot m, n \in \mathbb{N}$. Now, either $\boldsymbol{M'}$ implies a mutual information greater than zero, but then we don't have decay towards zero and hence no power-law behavior, or we indeed have mutual information of zero, but since a subsequence converges exponentially fast to $\boldsymbol{M'}$, the mutual information cannot be lower bounded by a power-law (see corollary **??**).

Note that if $\boldsymbol{M_Y^{m\tau}}$ converges to any matrix exponentially fast for $\tau \to \infty$, then $\boldsymbol{M_X M_Y^{m\tau} M_R}$ will be forced to converge exponentially fast as well.

We differentiate the following cases based on the properties of $\boldsymbol{M_Y}$:

Case 1: Irreducible and Aperiodic
If $\boldsymbol{M_Y}$ is irreducible and aperiodic, then we can use the data processing inequality to argue that

$$I(X_0; X_\tau) \leq I(Y_0; Y_\tau) \quad .$$

Now, using the Perron-Frobenius theorem, we know that $\boldsymbol{M_Y^\tau}$ converges to a rank-one matrix exponentially fast in τ. Hence, $I(Y_0; Y_\tau)$ converges exponentially fast to zero.

Case 2: Irreducible and Periodic
Assume $\boldsymbol{M_Y}$ has periodicity p. Let's analyze $\boldsymbol{M_Y^p}$: Based on lemma 1.2, it must decompose into p aperiodic closed blocks (when ordering the states accordingly):

$$\boldsymbol{M_Y^p} = \begin{bmatrix} \boldsymbol{B_1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B_2} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{B_p} \end{bmatrix} \quad .$$

Since all blocks represent irreducible aperiodic Markov chains, $\boldsymbol{M_Y^{p\tau}}$ must converge exponentially fast. But this means that $I(X_0; X_\tau)$ converges exponentially fast for $\tau = n \cdot p,\ n \in \mathbb{N}$, and hence it cannot be lower bounded by a power-law assuming convergence to zero.

Case 3: Multiple Closed Aperiodic Communication Classes
In this case, we can order the states such that $\boldsymbol{M_Y}$ is block diagonal, i.e.

$$\boldsymbol{M_Y} = \begin{bmatrix} \boldsymbol{B_1} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B_2} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{B_k} \end{bmatrix} \quad .$$

It follows that

$$\boldsymbol{M_Y^\tau} = \begin{bmatrix} \boldsymbol{B_1^\tau} & \boldsymbol{0} & \cdots & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{B_2^\tau} & \cdots & \boldsymbol{0} \\ \vdots & \vdots & \ddots & \vdots \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{B_k^\tau} \end{bmatrix} \quad .$$

Hence, $\boldsymbol{M_Y^\tau}$ converges to a certain block diagonal matrix exponentially fast, since all the blocks $\boldsymbol{B_i}$ are irreducible and aperiodic.

Case 4: Multiple Closed Communication Classes
Now assume $\boldsymbol{M_Y}$ consists of many closed communication classes that can be either periodic or aperiodic. But we know that all the aperiodic classes converge exponentially fast, and the periodic ones as well if we restrict $\tau \equiv_{m_i} 0$ for a specific m_i associated with block $\boldsymbol{B_i}$. By calculating the smallest common multiple of all m_i defined as m_I, we see that $\boldsymbol{M_Y^\tau}$ converges exponentially fast for the subsequence $\tau = n \cdot m_I,\ n \in \mathbb{N}$.

Case 5: The Generic Case
Finally, we allow $\boldsymbol{M_Y}$ to consist of multiple closed and open communication classes. Let S_C denote the set of all states that are in a closed communication class, and let S_O denote the set of states in open communication classes. We also use them to refer to certain submatrices (see below). After ordering states appropriately, we have:

$$\boldsymbol{M_Y} = \begin{bmatrix} \boldsymbol{S_C} & \boldsymbol{S'_O} \\ \boldsymbol{0} & \boldsymbol{S_O} \end{bmatrix} \quad ,$$

where the blocks $\boldsymbol{S_C}$ and $\boldsymbol{S_O}$ are square. Hence:

$$\boldsymbol{M_Y^\tau} = \begin{bmatrix} \boldsymbol{S_C^\tau} & \boldsymbol{S_O'^{(\tau)}} \\ \boldsymbol{0} & \boldsymbol{S_O^\tau} \end{bmatrix} \quad .$$

Thus, the block described by $\boldsymbol{S_C}$ will converge exponentially fast for $\tau = n \cdot m,\ n \in \mathbb{N}$ for some $m \in \mathbb{N}$ based on Case 4, and $\boldsymbol{S_O^\tau}$ decays to $\boldsymbol{0}$ exponential fast as well, since we are guaranteed to leave the open communication classes at some point to stay in a closed one. Note that convergence to $\boldsymbol{0}$ is always exponentially fast.

But what about the states in $\boldsymbol{S'_O}$? Well, based on lemma 1.3 and the previous discussion, we know there exists an $m \in \mathbb{N}$ s.t. $\boldsymbol{S_C^m}$ is block diagonal with every block being irreducible and aperiodic:

$$\boldsymbol{M_Y^m} = \begin{bmatrix} \boldsymbol{B_1} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \uparrow \\ \boldsymbol{0} & \boldsymbol{B_2} & \cdots & \boldsymbol{0} & \boldsymbol{S_O'^{(m)}} \\ \vdots & \vdots & \ddots & \vdots & | \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{B_k} & \downarrow \\ \boldsymbol{0} & \boldsymbol{0} & \cdots & \boldsymbol{0} & \boldsymbol{S_O^m} \end{bmatrix} \quad .$$

Let's consider the submatrix $\boldsymbol{M_i}$ consisting of the states in $\boldsymbol{B_i}$ and S_O:

$$\boldsymbol{M_i} = \begin{bmatrix} \boldsymbol{B_i} & (\boldsymbol{S_O'^{(m)}})_i \\ \boldsymbol{0} & \boldsymbol{S_O^m} \end{bmatrix} \quad .$$

We see that the columns of the states in S_O match for $\boldsymbol{M_i^l}$ and $\boldsymbol{M_Y^{ml}}$ in the associated rows. Hence, we may focus on analyzing $\boldsymbol{M_i^l}$.

Since $\boldsymbol{B_i}$ is irreducible and aperiodic and $(\boldsymbol{S_O^m})^\tau \xrightarrow{\tau \to \infty} \boldsymbol{0}$ exponentially fast, we can apply lemma **??**, and see that $\boldsymbol{M_i^\tau}$ converges exponentially fast, and hence so must all entries in $\boldsymbol{M_Y^{m\tau}}$. □

1.1 Conclusions for Model Selection

Since we are interested in natural language modelling, a model with strong power-law behavior seems very desirable. However, as we just saw, a fixed-sized context window is not sufficient for power-law behavior, hence we should look out for alternatives instead.

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 is 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.

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.