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:

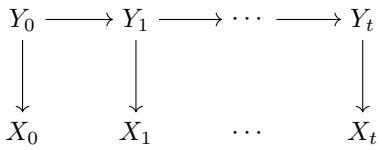

Figure 2: Bayesian network of a hidden Markov model.

These models are known as *hidden Markov models*. Unfortunately, there is no free lunch, as we will see in the following. But first, we will prove some lemmas.

Lemma 1.1. *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.*

Proof. A key property of an irreducible Markov chain with period p is that its state space S can be uniquely partitioned into p disjoint nonempty sets, called cyclic classes:

$$S = C_0 \cup C_1 \cup \cdots \cup C_{p-1} \quad .$$

These classes are defined such that a one-step transition from any state in class C_k can only lead to a state in the next class, $C_{(k+1) \pmod p}$.

Proving there are p Closed Classes

Let's consider the chain governed by the transition matrix $\boldsymbol{M}^p$, which represents taking steps of size p. If we start in any state $s \in C_k$, after one step of the new chain (which is p steps in the original chain), we will transition from $C_k \to C_{(k+1) \pmod p} \to \cdots \to C_{(k+p) \pmod p}$.

Since $(k+p) \pmod p = k$, any transition of length p starting in C_k must end in a state that is also in C_k. This means that for any state $s \in C_k$ and any state $t \notin C_k$, the transition probability is $(\boldsymbol{M}^p)_{st} = 0$.

Since transitions under $\boldsymbol{M}^p$ from C_k can only lead to states within C_k, each cyclic class is a closed communication class in the new chain. Furthermore, the partition gives us exactly p such classes: Because the original chain is irreducible, every state must belong to one of these classes, and each class must be able to reach the next, ensuring all p classes are part of the overall structure and are non-empty.

Proving Aperiodicity
Now we must show that each of these p classes is aperiodic in the $\boldsymbol{M}^p$ chain. Based on lemma **??**, we know that for every state i there exists an $\ell \in \mathbb{N}$ s.t. ℓp and $(\ell+1)p$ are both return times in $\boldsymbol{M}$. Hence, ℓ and $\ell+1$ are return times in $\boldsymbol{M}^p$, and thus state i is aperiodic in $\boldsymbol{M}^p$. □

Lemma 1.2. *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.*

Proof. Since $\boldsymbol{M}$ is irreducible and aperiodic, there exists an $m \in \mathbb{N}_{>0}$ s.t. $\boldsymbol{M}^m > \mathbf{0}$ based on theorem **??**. Hence, $\mathbf{0} < \boldsymbol{M}^{mn} = (\boldsymbol{M}^n)^m$. Finally, using corollary **??**, it follows that $\boldsymbol{M}^n$ is irreducible and aperiodic. □

Lemma 1.3 (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:

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

□

Lemma 1.4. *Let*

$$\boldsymbol{M} \coloneqq \begin{bmatrix} \boldsymbol{A} & \boldsymbol{B} \\ \boldsymbol{0} & \boldsymbol{C} \end{bmatrix}$$

be a matrix consisting of submatrices $\boldsymbol{A} \in \mathbb{R}^{k \times k}$, $\boldsymbol{B} \in \mathbb{R}^{k \times \ell}$, *and* $\boldsymbol{C} \in \mathbb{R}^{\ell \times \ell}$. *Let* $\boldsymbol{A}$ *be an irreducible aperiodic Markov transition matrix, and let* $\boldsymbol{C}^n \xrightarrow{n \to \infty} \boldsymbol{0}$ *with exponential decay. Then,* $\boldsymbol{M}^n$ *decays exponentially in* n *towards a matrix* $\boldsymbol{M}'$.

Proof. Using induction, it is easy to show that

$$\boldsymbol{M}^n = \begin{bmatrix} \boldsymbol{A}^n & \sum_{i=0}^{n-1} \boldsymbol{A}^{n-1-i} \boldsymbol{B} \boldsymbol{C}^i \\ \boldsymbol{0} & \boldsymbol{C}^n \end{bmatrix} \quad .$$

We aim to show that this sum converges to a finite matrix:

$$\lim_{n \to \infty} \sum_{i=0}^{n-1} \boldsymbol{A}^{n-1-i} \boldsymbol{B} \boldsymbol{C}^i =: \boldsymbol{D}_\infty \quad .$$

Let us denote:

$$\boldsymbol{D}_n := \sum_{i=0}^{n-1} \boldsymbol{A}^{n-1-i} \boldsymbol{B} \boldsymbol{C}^i \quad .$$

We want to show that $\|\boldsymbol{D}_{n+1} - \boldsymbol{D}_n\| \leq c e^{-\alpha n}$ for some $\alpha \in \mathbb{R}_{>0}$. Note that the exponentially fast convergence of $\boldsymbol{D}_n$ is equivalent to exponentially fast convergence of every entry in the matrix.

Since $\boldsymbol{A}$ is irreducible and aperiodic, it follows that $\lambda_1 = 1$ and $|\lambda_2| < 1$ based on theorem **??**. Hence, it converges exponentially fast with a basis of $|\lambda_2|$.

Furthermore, note that every entry α in every matrix for every n is bounded by $\alpha \in [0, 1]$. Hence, we can argue about their element-wise deviation in big $\mathcal{O}$ notation where the hidden constants remain bounded:

$$
\begin{aligned}
&\boldsymbol{D}_{n+1} - \boldsymbol{D}_n \\
&= \sum_{i=0}^{n} \boldsymbol{A}^{n-i}\boldsymbol{B}\boldsymbol{C}^i - \sum_{i=0}^{n-1} \boldsymbol{A}^{n-1-i}\boldsymbol{B}\boldsymbol{C}^i \\
&= \boldsymbol{B}\boldsymbol{C}^n + \sum_{i=0}^{n-1} \left[\boldsymbol{A}^{n-i}\boldsymbol{B}\boldsymbol{C}^i - \boldsymbol{A}^{n-1-i}\boldsymbol{B}\boldsymbol{C}^i\right] \\
&= \boldsymbol{B}\boldsymbol{C}^n + \sum_{i=0}^{\frac{n}{2}-1} \left[\left(\boldsymbol{A}^{n-i} - \boldsymbol{A}^{n-1-i}\right)\boldsymbol{B}\boldsymbol{C}^i\right] + \sum_{\frac{n}{2}}^{n-1} \left[\left(\boldsymbol{A}^{n-i} - \boldsymbol{A}^{n-1-i}\right)\boldsymbol{B}\boldsymbol{C}^i\right] \\
&= \pm\mathcal{O}(e^{-\alpha_1 n}) + \sum_{i=0}^{\frac{n}{2}-1} \left[\left(\boldsymbol{A}^{n-i} - \boldsymbol{A}^{n-1-i}\right)\boldsymbol{B}\boldsymbol{C}^i\right] + \sum_{\frac{n}{2}}^{n-1} \left[\left(\boldsymbol{A}^{n-i} - \boldsymbol{A}^{n-1-i}\right)\boldsymbol{B}\cdot\mathcal{O}(e^{-\alpha_1 \frac{n}{2}})\right] \\
&= \pm\mathcal{O}(e^{-\alpha_1 n}) + \sum_{i=0}^{\frac{n}{2}-1} \left[\left(\boldsymbol{A}^{n-i} - \boldsymbol{A}^{n-1-i}\right)\boldsymbol{B}\boldsymbol{C}^i\right] \pm \mathcal{O}(ne^{-\alpha_1 \frac{n}{2}}) \\
&= \pm\mathcal{O}(e^{-\alpha_1 n}) + \sum_{i=0}^{\frac{n}{2}-1} \left[\left(\boldsymbol{A}^{n-i} - \boldsymbol{A}^{\infty} + \boldsymbol{A}^{\infty} - \boldsymbol{A}^{n-1-i}\right)\boldsymbol{B}\boldsymbol{C}^i\right] \pm \mathcal{O}(ne^{-\alpha_1 \frac{n}{2}}) \\
&= \pm\mathcal{O}(e^{-\alpha_1 n}) + \sum_{i=0}^{\frac{n}{2}-1} \left[\left(\pm\mathcal{O}(e^{-\alpha_2(n-1-i)})\right)\boldsymbol{B}\boldsymbol{C}^i\right] \pm \mathcal{O}(ne^{-\alpha_1 \frac{n}{2}}) \\
&= \pm\mathcal{O}(e^{-\alpha_1 n}) \pm \mathcal{O}(ne^{-\alpha_2 \frac{n}{2}}) \pm \mathcal{O}(ne^{-\alpha_1 \frac{n}{2}}) \\
&= \pm\mathcal{O}(e^{-\alpha n}) \quad .
\end{aligned}
$$

It is easy to see that $(\boldsymbol{D}_n)_{n=1}^{\infty}$ is a Cauchy sequence and hence $\boldsymbol{D}_\infty$ exists. And of course

$$
\begin{aligned}
\|\boldsymbol{D}_\infty - \boldsymbol{D}_n\| &\leq \sum_{i=n}^{\infty} ce^{-\alpha i} \\
&= c\sum_{i=n}^{\infty} (e^{-\alpha})^i \\
&= c\left(\frac{1}{1-e^{-\alpha}} - \frac{1-e^{-\alpha n}}{1-e^{-\alpha}}\right) \\
&= c\frac{e^{-\alpha n}}{1-e^{-\alpha}} \in \mathcal{O}(e^{-\alpha n}) \quad .
\end{aligned}
$$

From here, the claim follows trivially. □

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 chose 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.

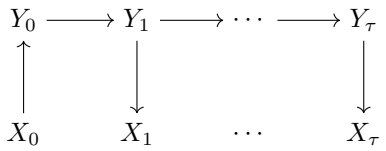

Figure 3: Adjusted Bayesian network of a hidden Markov model.

We see that $P(X_\tau = a \mid X_0 = b) = (\boldsymbol{M_X M_Y^\tau M_R})_{ab}$.

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 certain $m \in \mathbb{N}$ we have $\boldsymbol{M_X M_Y^{m\tau} M_R} \xrightarrow{\tau\to\infty} \boldsymbol{M'}$ with exponential decay. 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 we converge with exponential decay, 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 with exponential decay for $\tau \to \infty$, then $\boldsymbol{M_X M_Y^{m\tau} M_R}$ will be forced to converge with exponential decay 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 have based on theorem **??** that

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

But we have already proven that $I(Y_0; Y_\tau)$ decays exponentially in theorem **??**.

Case 2: Irreducible and Periodic
Assume $\boldsymbol{M_Y}$ has periodicity p. Let's analyze $\boldsymbol{M_Y^p}$: Based on lemma 1.1, 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 0.

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.

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

It follows that

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

Hence, $\boldsymbol{M_Y^\tau}$ converges to a certain block diagonal matrix with exponential decay 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 with exponential decay, 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 with exponential decay for $\tau = n \cdot m_I,\ n \in \mathbb{N}$.

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:

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

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

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

Thus, the block described by $\boldsymbol{S_C}$ will converge with exponential decay 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 $\boldsymbol{0}$ with exponential decay as well.

But what about the states in $\boldsymbol{S'_O}$? Well, based on lemma 1.2 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:

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

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

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

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 $(\boldsymbol{S_O^m})^\tau \xrightarrow{\tau\to\infty} \boldsymbol{0}$ with exponential decay, we can apply lemma 1.4, and see that $\boldsymbol{M_i^\tau}$ converges with exponential decay, 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, 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.