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 which is also called a *hidden Markov model*:

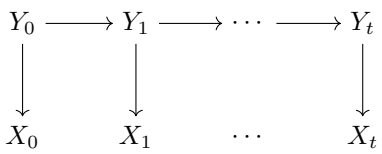

Figure 2: Bayesian network of a hidden Markov model.

Lemma 1.1 (Hidden Markov Models have the Bulk Marginal Property)**.** *Every hidden Markov model* $(\boldsymbol{M_Y}, \boldsymbol{M_X})$ *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 1.2. *Let $\boldsymbol{M}$ describe an irreducible aperiodic Markov chain. Then for every $n \in \mathbb{N}$, $\boldsymbol{M}^n$ is still 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, $\boldsymbol{M}^{mn} > \mathbf{0}$. Based on corollary **??** we know that $\boldsymbol{M}^n$ is irreducible and aperiodic. □

Lemma 1.3. *Let*

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

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 $\boldsymbol{C}^n \xrightarrow{n \to \infty} \mathbf{0}$ with exponential decay. Then, $\boldsymbol{M}^n \xrightarrow{n \to \infty} \boldsymbol{M'}$ with exponential decay for some matrix $\boldsymbol{M'}$.

Lemma 1.4. *Let $\boldsymbol{M}$ be a block matrix representing a Markov chain, defined as*

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

where $\boldsymbol{A} \in \mathbb{R}^{k \times k}$, $\boldsymbol{B} \in \mathbb{R}^{k \times \ell}$, and $\boldsymbol{C} \in \mathbb{R}^{\ell \times \ell}$. Assume the following conditions hold:

(i) *$\boldsymbol{A}$ is the transition matrix of an irreducible and aperiodic Markov chain.*

(ii) *$\lim_{n \to \infty} \boldsymbol{C}^n = \mathbf{0}$ with an exponential rate of decay.*

Then $\boldsymbol{M}^n$ converges to a matrix $\boldsymbol{M}_\infty$ exponentially fast.

Proof. By induction, the n-th power of the block matrix $\boldsymbol{M}$ is given by:

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

We will analyze the convergence of each block separately.

Convergence of $\boldsymbol{A}_n = \boldsymbol{A}^n$: Since $\boldsymbol{A}$ is the transition matrix of an irreducible, aperiodic finite Markov chain, the Perron-Frobenius theorem guarantees that its powers converge exponentially to a rank-one matrix $\boldsymbol{A}_\infty$. Each row of $\boldsymbol{A}_\infty$ is the unique stationary distribution $\boldsymbol{\pi_A}$. Thus, there exist constants $c_A > 0$ and $\rho_A \in (0, 1)$ such that for any compatible matrix norm $\|\cdot\|$:

$$\|\boldsymbol{A}^n - \boldsymbol{A}_\infty\| \leq c_A \rho_A^n \quad .$$

Convergence of $\boldsymbol{C}_n = \boldsymbol{C}^n$: By assumption (ii), $\boldsymbol{C}^n \to \mathbf{0}$ exponentially. This is equivalent to the condition that the spectral radius of $\boldsymbol{C}$, denoted $\rho(\boldsymbol{C})$, is

less than 1. This implies that there exist constants $c_C > 0$ and $\rho_C \in (\rho(\boldsymbol{C}), 1)$ such that:

$$\|\boldsymbol{C}^n\| \leq c_C \rho_C^n \quad .$$

Furthermore, since $\rho(\boldsymbol{C}) < 1$, the matrix $(\boldsymbol{I} - \boldsymbol{C})$ is invertible, and the geometric series of matrices converges: $\sum_{i=0}^{\infty} \boldsymbol{C}^i = (\boldsymbol{I} - \boldsymbol{C})^{-1}$.

Convergence of $\boldsymbol{D}_n = \sum_{i=0}^{n-1} \boldsymbol{A}^{n-1-i}\boldsymbol{B}\boldsymbol{C}^i$: Let us define the limit matrix $\boldsymbol{D}_\infty \coloneqq \sum_{i=0}^{\infty} \boldsymbol{A}_\infty \boldsymbol{B}\boldsymbol{C}^i = \boldsymbol{A}_\infty \boldsymbol{B}(\boldsymbol{I} - \boldsymbol{C})^{-1}$. The existence of $\boldsymbol{D}_\infty$ is guaranteed by the convergence of the geometric series of $\boldsymbol{C}$. We now show that $\|\boldsymbol{D}_n - \boldsymbol{D}_\infty\|$ decays exponentially.

$$\begin{aligned}
\boldsymbol{D}_\infty - \boldsymbol{D}_n &= \sum_{i=0}^{\infty} \boldsymbol{A}_\infty \boldsymbol{B}\boldsymbol{C}^i - \sum_{i=0}^{n-1} \boldsymbol{A}^{n-1-i}\boldsymbol{B}\boldsymbol{C}^i \\
&= \sum_{i=n}^{\infty} \boldsymbol{A}_\infty \boldsymbol{B}\boldsymbol{C}^i + \sum_{i=0}^{n-1} (\boldsymbol{A}_\infty - \boldsymbol{A}^{n-1-i})\boldsymbol{B}\boldsymbol{C}^i \quad .
\end{aligned}$$

Using the triangle inequality for the norm:

$$\|\boldsymbol{D}_\infty - \boldsymbol{D}_n\| \leq \underbrace{\left\| \sum_{i=n}^{\infty} \boldsymbol{A}_\infty \boldsymbol{B}\boldsymbol{C}^i \right\|}_{\text{Term 1}} + \underbrace{\left\| \sum_{i=0}^{n-1} (\boldsymbol{A}_\infty - \boldsymbol{A}^{n-1-i})\boldsymbol{B}\boldsymbol{C}^i \right\|}_{\text{Term 2}} \quad .$$

For **Term 1**, the tail of a convergent geometric series also decays geometrically:

$$\left\| \sum_{i=n}^{\infty} \boldsymbol{A}_\infty \boldsymbol{B}\boldsymbol{C}^i \right\| \leq \|\boldsymbol{A}_\infty\| \|\boldsymbol{B}\| \sum_{i=n}^{\infty} \|\boldsymbol{C}^i\| \leq \|\boldsymbol{A}_\infty\| \|\boldsymbol{B}\| \sum_{i=n}^{\infty} c_C \rho_C^i = \|\boldsymbol{A}_\infty\| \|\boldsymbol{B}\| c_C \frac{\rho_C^n}{1 - \rho_C} \quad .$$

This term decays exponentially with rate ρ_C.

For **Term 2**, we bound the sum, which is a convolution of two exponentially decaying sequences:

$$\begin{aligned}
\left\| \sum_{i=0}^{n-1} (\boldsymbol{A}_\infty - \boldsymbol{A}^{n-1-i})\boldsymbol{B}\boldsymbol{C}^i \right\| &\leq \sum_{i=0}^{n-1} \|\boldsymbol{A}_\infty - \boldsymbol{A}^{n-1-i}\| \|\boldsymbol{B}\| \|\boldsymbol{C}^i\| \\
&\leq \sum_{i=0}^{n-1} (c_A \rho_A^{n-1-i}) \|\boldsymbol{B}\| (c_C \rho_C^i) \\
&= c_A c_C \|\boldsymbol{B}\| \sum_{i=0}^{n-1} \rho_A^{n-1-i} \rho_C^i \\
&= c_A c_C \|\boldsymbol{B}\| \rho_A^{n-1} \sum_{i=0}^{n-1} \left(\frac{\rho_C}{\rho_A} \right)^i \quad .
\end{aligned}$$

If $\rho_A \neq \rho_C$, the geometric sum evaluates to $\frac{(\rho_C/\rho_A)^n - 1}{(\rho_C/\rho_A) - 1}$. The overall term is bounded by $k_1\rho_A^n + k_2\rho_C^n$ for some constants k_1, k_2. In either case (whether $\rho_A = \rho_C$ or not), the sum is bounded by a term that decays exponentially with a rate of $\max(\rho_A, \rho_C)$.

Let $\rho = \max(\rho_A, \rho_C) \in (0, 1)$. Both Term 1 and Term 2 are bounded by expressions of the form $K \cdot n^p \cdot \rho^n$ for $p \in \{0, 1\}$, which decays exponentially. Therefore, $\|\boldsymbol{D}_n - \boldsymbol{D}_\infty\|$ decays exponentially.

Conclusion: We have shown that each block of $\boldsymbol{M}^n$ converges exponentially to the corresponding block in the limit matrix

$$\boldsymbol{M}_\infty \coloneqq \begin{bmatrix} \boldsymbol{A}_\infty & \boldsymbol{D}_\infty \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \quad .$$

The rate of convergence for the entire matrix $\boldsymbol{M}^n$ is governed by the slowest decaying term, so $\|\boldsymbol{M}^n - \boldsymbol{M}_\infty\|$ decays exponentially, which completes the proof. □

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.

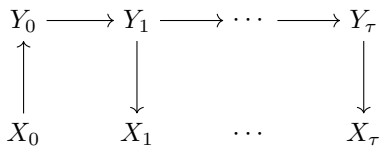

Figure 3: Adjusted Bayesian network of a hidden Markov model.

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

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 $\boldsymbol{M_X}\boldsymbol{M_Y^{m\tau}}\boldsymbol{M_R} \xrightarrow{\tau\to\infty} \boldsymbol{M'}$ 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. (Really true?)

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.

Case 2: 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 with exponential decay since all the blocks $\boldsymbol{B_i}$ are irreducible and aperiodic.

Case 3: Irreducible and Periodic
Assume $\boldsymbol{M_Y}$ has periodicity p. Let's analyze $\boldsymbol{M_Y^p}$: Because of periodicity, it must decompose into p independent blocks (when ordering the states accordingly):

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

Because $\boldsymbol{M_Y}$ is irreducible, so must all blocks $\boldsymbol{B_i}$. Now there are two cases: Either all blocks are aperiodic, but then we are in case 2, hence $\boldsymbol{M_Y^{p\tau}}$ converges with exponential decay. But this means that $I(X_0; X_\tau)$ has exponential decay for $\tau = n \cdot p,\ n \in \mathbb{N}$, and hence it cannot be lower bounded by a power-law.

On the other hand, if there are blocks that are still are periodic, we recursively apply this procedure to arrive at a matrix $\boldsymbol{M_Y^m}$ for some $m \in \mathbb{N}$ with irreducible aperiodic blocks after finitely many iterations (note lemma 1.2). Hence, by the same logic we arrive at a contradiction.

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:

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

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 with exponential decay for $\tau = n \cdot m,\ n \in \mathbb{N}$ for some $m \in \mathbb{N}$ based on Case 4. Furthermore, $\boldsymbol{S_O^\tau}$ decays to $\boldsymbol{0}$ with exponential decay. (One can show that $|\lambda_1| \leq 1$ like in the proof of theorem **??**. Also, since $\boldsymbol{S_O}$ describes a collection of states in open communication classes, we have $|\lambda_1| \neq 1$. (Why?))

But what about the states in $\boldsymbol{S'_O}$? Well, based on 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 $(\boldsymbol{S_O^m})^\tau \xrightarrow{\tau \to \infty} \boldsymbol{0}$ 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.