A Uniform Convergence Result for Learning Text Data

Szymon Snoeck

March 2025

1 Abstract

With the advent of large language models, the forefront of machine learning research has diverged from classical learning theory's assumption that data is sampled independently. In a small step towards remedying this difference, a new model for learning *sequential* data is introduced alongside an agnostic uniform convergence result which aims to close the gap between traditional i.i.d.-based learning theory and the present field of NLP.

2 Introduction

While classical learning theory has been tremendously successful at quantifying the difficulty of learning in the regime of independent data, to the best of the author's knowledge, there exist few examples of extensions to the non-i.i.d. regime, and of the examples that do exist, they fail to capture the reality of learning sequential data such as text data. For example, Hao et al. [2018] studied the task of estimating the transition matrix of a Markov chain from samples with respect to f-divergences. While Markov chain's can some what approximate text data, Hao et al. [2018] does not provide a true uniform convergence result of the learning theory style as they assume the hypothesis class $\mathcal{G}$ is unconstrained. Yu [1994] improves on this by providing uniform convergence results for learning Markov chains based on the covering number of $\mathcal{G}$ and the β mixing time (a slightly weaker condition then bound mixing time). However, Yu [1994] has two blind spots. First, it is assumed that the Markov chain starts from the stationary distribution which is unrealistic of real data and, secondly, Markov chains do not inherently capture next token generation models where the next output has dependencies on the last *several* outputs. The main result of this manuscript improves in both these regards. First, it makes no assumption on the initial distribution, and second, it proves an agnostic uniform convergence result with respect to the task of learning next-token prediction for hypothesis classes of bound pseudo dimension where the context length is arbitrary.

3 Preliminaries

Let $\mathcal{V}$ be the vocabulary—the list of tokens that are possible to output—and assume it is finite in size. For an ordered sequence of words $(v_1, \dots, v_m) \in \mathcal{V}^m$, we use the simplified notation $v_{1:m}$. Moreover, for $u_{1:n} \in \mathcal{V}^n$, $v_{1:m} \cup u_{1:n} = (v_1, \dots, v_m, u_1, \dots, u_n)$. Let $\mathcal{V}^* = \bigcup_{n \in \mathbb{N}_0} \mathcal{V}^n$. In this report, we will assume that $\mathcal{V} = \{0, 1\}$ for the sake of analysis but we present the learning model in full generality.

Let $\Delta^{\mathcal{V}} = \{(a_v)_{v\in\mathcal{V}} \in [0,1]^{|\mathcal{V}|} | \sum_{v\in\mathcal{V}} a_v = 1\}$ be the probability simplex of distributions over $\mathcal{V}$. The target concept is modeled as a function $f : \mathcal{V}^m \to \mathcal{V}$, with finite context length m, that produces the next token probabilistically. Each output is distributed over $\mathcal{V}$ according to its (normalized) logits $F : \mathcal{V}^m \to \Delta^{\mathcal{V}}$. For simplicity we define the probability of outputting a word u given input $v_{1:m}$ to be $(F(v_{1:m}))_u = F_u(v_{1:m})$. In the case where $\mathcal{V} = \{0,1\}$, we let $F(v_{1:m}) = \mathbb{P}[f(v_{1:m}) = 1]$.

For $M \in \mathbb{N}$, $v_{1:M}$ is a **sample from f**, if it collected from f by running it repeatedly on its own output:

$$v_{1:M} = (f(v_{-m+1:0}), f \circ f(v_{-m+1:0}), \ldots, f^{\circ M}(v_{-m+1:0}))$$

where $v_{-m+1:0}$ is chosen from some arbitrary initial distribution π_0.

The goal will be use such a sample to learn a model in a class of feasible models $\mathcal{G}$ where $g \in \mathcal{G}$ has context length $n \leq m$ (i.e. $g : \mathcal{V}^n \to \mathcal{V}$). For $g \in \mathcal{G}$, We mirror the definitions of the logits, $F : \mathcal{V}^m \to \Delta^{\mathcal{V}}, F_u : \mathcal{V}^m \to [0,1]$, by defining $G : \mathcal{V}^m \to \Delta^{\mathcal{V}}, G_u : \mathcal{V}^m \to [0,1]$ identically except corresponding to g instead of f.

In classical learning, the samples come from some distribution over which the error is defined. Since in this case the labels—outputs of f—and inputs—sequences of outputs of f—are generated simultaneously, the distribution over samples must be derived from f. The most natural such distribution is the limiting distribution. In other words, the distribution that captures the m-gram probabilities of an infinitely long sequence of outputs of f. To construct this distribution, we recast f as a Markov chain $\mathbb{M}$ on state space $\mathcal{V}^m$. Define the probability transition matrix $P \in [0,1]^{\mathcal{V}^m \times \mathcal{V}^m}$ of $\mathbb{M}$ to be such that $\forall v_{1:m}, u_{1:m} \in \mathcal{V}^m$, $P_{v_{1:m},u_{1:m}} = \prod_{i\in[m]} F_{u_i}(v_{i:m} \cup u_{1:i-1})$—the probability that $u_{1:m}$ is produced given context $v_{1:m}$. Set $\mathbb{M} = (\mathcal{V}^m, P)$. It is clear to see that it is indeed a Markov chain. We will refer to it as the Markov chain induced by f. Now we can define the limiting distribution of examples in $\mathcal{V}^m$ under f to be the stationary distribution of $\mathbb{M}(\mathcal{V}^m, P)$, π, which exists under mild conditions (see section 5).

It will be useful to extend the definition of π to distributions over strings of arbitrary length. Thus for $m' \in \mathbb{N}$, define $\pi^{(m')}$ such that for all $v_{1:m'} \in \mathcal{V}^*$, $\mathbb{P}_{\pi^{(m')}}[v_{1:m'}] = \mathbb{E}_{u_{1:m}\sim\pi}[\prod_{i\in[m']} F_{v_i}(u_{i:m} \cup v_{1:i-1})]$. This gives us that $\pi^{(m+1)}$, the joint distribution over samples and labels (i.e. $(v_{1:m}, f(v_{1:m})) \sim \pi^{(m+1)}$), is such that $\mathbb{P}_{\pi^{(m+1)}}[v_{1:m+1}] = \mathbb{P}_{\pi}[v_{1:m}]F_{v_{m+1}}(v_{1:m})$.

Now we can finally define the error. A natural objective for sequence learning is minimizing the L_2 distance between f and $g \in \mathcal{G}$'s outputs, as seen in chapter 16 of Anthony and Bartlett [2009], so we set the error to be:

$$er(f,g) = \mathop{\mathbb{E}}_{v_{1:m}\sim\pi}[\|F(v_{1:m}) - G(v_{m-n+1:m})\|_2^2].$$

Thus, the goal of a learning algorithm is to take in a sample from f, and output some $g \in \mathcal{G}$ which approximately minimizes the error with respect to f. This type of algorithm is dubbed a *sequential learning algorithm*:

Definition 1. *An algorithm $\mathcal{L}$ outputting in $\mathcal{G}$ is a **sequence learning algorithm** for $f : \mathcal{V}^m \to \mathcal{V}$ if for all $\epsilon, \delta \in (0,1)$, there exists an integer $M_0(\epsilon,\delta)$ such that if $\mathcal{L}$ receives a sample from f of length $M \geq M_0(\epsilon,\delta)$ then:*

$$\mathbb{P}[er_\pi(f,\mathcal{L}) \leq \inf_{g\in\mathcal{G}} er_\pi(f,g) + \epsilon] \geq 1 - \delta$$

where the probability is over the sample and $\mathcal{L}$'s internal randomness.

4 Main Result

Given this new model of learning, one of the most natural questions to ask is: what is the sample complexity of an empirical risk minimizing algorithm for sequential learning? Before we can answer this, it first needs to be decided what the empirical risk even is. The following observation gives the finite sample loss that an ERM algorithm should minimize over $\mathcal{G}$. Recall that $\pi^{(m+1)}$ is a joint distribution over $v_{1:m} \in \mathcal{V}^m$ and its label $f(v_{1:m})$.

Observation 2. *For $Z^{(i\in\mathbb{N})} \sim \pi^{(m+1)}$ i.i.d:*

$$\begin{aligned}\arg\inf_{g\in\mathcal{G}} er(f,g) &= \arg\inf_{g\in\mathcal{G}} \mathop{\mathbb{E}}_{v_{1:m}\sim\pi}[G^2(v_{m-n+1:m}) - 2F(v_{1:m})G(v_{m-n+1:m})] \\ &= \arg\inf_{g\in\mathcal{G}} \lim_{M\to\infty} \frac{1}{M}\sum_{i\in[M]} G^2(Z^{(i)}_{m-n+1:m}) - \frac{2}{M}\sum_{i\in[M]} Z^{(i)}_{m+1}\cdot G(Z^{(i)}_{m-n+1:m})\end{aligned}$$

Therefore we define:

$$\hat{er}(Z^{(1)},\dots,Z^{(M)},g) \equiv \frac{1}{M}\sum_{i\in[M]} G^2(Z^{(i)}_{m-n+1:m}) - \frac{2}{M}\sum_{i\in[M]} Z^{(i)}_{m+1}\cdot G(Z^{(i)}_{m-n+1:m})$$

Before presenting the uniform convergence result, we define the key parameters that captures how predictable the output of f is:

Definition 3. *Let d_{TV} be the total variation distance. For any target concept $f : \mathcal{V}^m \to \mathcal{V}$, if the sequence of random variables $v_1, v_2, v_3, \dots$ is a sample from f, then the ξ-mixing time of f is:*

$$T(\xi) = \min\{t \geq 0 :\ \forall i \geq 0 \max_{u_{0:m}\in\mathcal{V}^{m+1}} d_{TV}(\mathbb{P}[v_{i+t:i+m+t} \in \cdot\ | v_{i:i+m} = u_{0:m}], \pi^{(m+1)}) \leq \xi\}$$

Now we present the main result of this section:

Theorem 4. *Let $\mathcal{G}$ be a hypothesis space of functions from $\mathcal{V}^n$ to $\mathcal{V} = \{0,1\}$ and $f : \mathcal{V}^m \to \mathcal{V}$ be any target concept with $\frac{\epsilon\delta}{6}$-mixing time at most $T < \infty$. If $d < \infty$ is the pseudo dimension of $er_{\mathcal{G}} \equiv \{er_g(v_{1:m}, v_{m+1}) = G^2(v_{1:m}) - 2v_{m+1}G(v_{1:m})\ : g \in \mathcal{G}\}$ and*

$$M' \geq \frac{256\,(m+1)\,T}{\epsilon^2}\left(2d\ln\left(\frac{32e}{\epsilon}\right) + \ln\left(\frac{18}{\delta}\right)\right) + (m+1)T$$

then, for any $v_1,\dots,v_{M'}$ sampled from f with arbitrary initial distribution π_0, there exists a subsample of $v_{1:M'}$ $X^{(1)},\dots,X^{(M)} \in \mathcal{V}^{m+1}$, such that:

$$\mathbb{P}[\forall g \in \mathcal{G},\ \textit{it holds that } |\hat{er}(X^{(1)},\dots,X^{(M)},g) - er(f,g)| \leq \epsilon] \geq 1-\delta$$

where the probability is over the generation of $v_1,\dots,v_{M'}$.

The key observation behind this proof is that any sequence of weakly dependent-data can be related to a carefully constructed set of i.i.d samples:

Lemma 5. *If $X^{(0)}, \dots, X^{(M)} \in \mathcal{V}^{m+1}$ form a Markov chain such that for all $i \in [1, M]$:*

$$\max_{v_{0:m} \in \mathcal{V}^{m+1}} d_{TV}(\mathbb{P}[X^{(i)} \in \cdot \mid X^{(i-1)} = v_{0:m}], \pi^{(1)}) \leq \xi$$

then there exist $Z^{(1)}, \dots, Z^{(M)}$ i.i.d from $\pi^{(1)}$ with for all $i \in [M]$:

$$\mathbb{P}[Z^{(i)} \neq X^{(i)}] \leq \frac{\xi}{2}$$

Proof of lemma 5. For all $i \in [1, M]$, consider the random variable $(X^{(i-1)}, Z^{(i)})$ such that $Z^{(i)} \sim \pi^{(1)}$ is independent of $X^{(i-1)}$. By theorem 2.12 of den Hollander [2012], there exists a joint distribution, λ_i over $(X^{(i-1)}, X^{(i)}), (X^{(i-1)}, Z^{(i)})$ such that:

$$d_{TV}((X^{(i-1)}, X^{(i)}), (X^{(i-1)}, Z^{(i)})) = 2\,\mathbb{P}[(X^{(i-1)}X^{(i)}) \neq (X^{(i-1)}, Z^{(i)})].$$

By the assumption on $X^{(1)}, \dots, X^{(M)}$ and construction of λ_i:

$$\begin{aligned}
\mathbb{P}[X^{(i)} \neq Z^{(i)}] &= \frac{1}{2} d_{TV}((X^{(i-1)}, X^{(i)}), (X^{(i-1)}, Z^{(i)})) \\
&= \frac{1}{4} \sum_{v_{0:m}, u_{0:m} \in \mathcal{V}^{m+1}} |\,\mathbb{P}[X^{(i-1)} = v_{0:m}]\,\mathbb{P}[X^{(i)} = u_{0:m} | X^{(i-1)} = v_{0:m}] \\
&\quad - \mathbb{P}[X^{(i-1)} = v_{0:m}] \pi^{(1)}(u_{0:m})| \\
&\leq \frac{1}{4} \max_{v_{0:m} \in \mathcal{V}^{m+1}} \sum_{u_{0:m} \in \mathcal{V}^{m+1}} |\,\mathbb{P}[X^{(i)} = u_{0:m} | X^{(i-1)} = v_{0:m}] - \pi^{(1)}(u_{0:m})| \\
&= \frac{1}{2} \max_{v_{0:m} \in \mathcal{V}^{m+1}} d_{TV}(\mathbb{P}[X^{(i)} \in \cdot | X^{(i-1)} = v_{0:m}], \pi^{(1)}) \\
&\leq \frac{\xi}{2}.
\end{aligned}$$

Now we verify that the induced distribution over $Z^{(1)}, \dots, Z^{(M)}$ is indeed the i.i.d distribution $\pi^{(1)} \times \cdots \times \pi^{(1)}$. By the Markov property, each $X^{(i)}$ is independent of the past given $X^{(i-1)}$ and, by construction, $Z^{(i)}$ is only dependent on $X^{(i)}$ (since it is independent of $X^{(i-1)}$). Let $\lambda_i | X^{(i)}$ be the distribution over $Z^{(i)}$ given $X^{(i)}$. Given $X^{(1)}, \dots, X^{(M)}$, sample $Z^{(1)}, \dots, Z^{(M)}$ from $\lambda_1 | X^{(1)} \times \cdots \times \lambda_M | X^{(M)}$. Then for any sets $A_1, \dots, A_M \subseteq \mathcal{V}^{m+1}$:

$$\begin{aligned}
&\mathbb{P}[Z^{(1)} \in A_1, \dots, Z^{(M)} \in A_M] \\
&\quad = \mathop{\mathbb{E}}_{X^{(1)}, \dots, X^{(M)}} [\mathbb{P}[Z^{(1)} \in A_1, \dots, Z^{(M)} \in A_M | X^{(1)}, \dots, X^{(M)}]] \\
&\quad = \mathop{\mathbb{E}}_{X^{(1)}, \dots, X^{(M)}} [\mathop{\mathbb{P}}_{Z^{(1)} \sim \lambda_1 | X^{(1)}} [Z^{(1)} \in A_1] \cdots \mathop{\mathbb{P}}_{Z^{(M)} \sim \lambda_M | X^{(M)}} [Z^{(M)} \in A_M]] \\
&\quad = \mathop{\mathbb{E}}_{X^{(1)}, \dots, X^{(M-1)}} [\mathop{\mathbb{P}}_{Z^{(1)} \sim \lambda_1 | X^{(1)}} [Z^{(1)} \in A_1] \cdots \mathop{\mathbb{P}}_{Z^{(M-1)} \sim \lambda_1 | X^{(M-1)}} [Z^{(M-1)} \in A_{M-1}] \\
&\qquad \mathop{\mathbb{E}}_{X^{(M)}} [\mathop{\mathbb{P}}_{Z^{(M)} \sim \lambda_M | X^{(M)}} [Z^{(M)} \in A_M] | X^{(M-1)}]].
\end{aligned}$$

Since $Z^{(M)}$ is independent of $X^{(M-1)}$ we get that:

$$
\begin{aligned}
&= \mathop{\mathbb{E}}_{X^{(1)},\ldots,X^{(M-1)}}\Big[\mathop{\mathbb{P}}_{Z^{(1)}\sim\lambda_1|X^{(1)}}[Z^{(1)}\in A_1]\cdots \mathop{\mathbb{P}}_{Z^{(M-1)}\sim\lambda_1|X^{(M-1)}}[Z^{(M-1)}\in A_{M-1}] \\
&\quad \mathop{\mathbb{E}}_{X^{(M)}}\Big[\mathop{\mathbb{P}}_{Z^{(M)}\sim\lambda_M|X^{(M)}}[Z^{(M)}\in A_M]\Big]\Big] \\
&= \mathop{\mathbb{E}}_{X^{(1)},\ldots,X^{(M-1)}}\Big[\mathop{\mathbb{P}}_{Z^{(1)}\sim\lambda_1|X^{(1)}}[Z^{(1)}\in A_1]\cdots \mathop{\mathbb{P}}_{Z^{(M-1)}\sim\lambda_1|X^{(M-1)}}[Z^{(M-1)}\in A_{M-1}]\Big] \\
&\quad \mathop{\mathbb{P}}_{Z^{(M)}\sim\pi^{(1)}}[Z^{(M)}\in A_M]
\end{aligned}
$$

and by induction:

$$
= \mathop{\mathbb{P}}_{Z^{(1)}\sim\pi^{(1)}}[Z^{(1)}\in A_1]\cdots \mathop{\mathbb{P}}_{Z^{(M)}\sim\pi^{(1)}}[Z^{(M)}\in A_M].
$$

□

Now we prove Theorem 4:

Proof of theorem 4. The proof proceeds by reducing a sample $v_{1:M'}$ produced by f to a smaller i.i.d sample from the stationary distribution such that the empirical error is close to the original. To finish the proof, established results in learning theory are applied.

Let $M' \geq \frac{256(m+1)T}{\epsilon^2}\left(2d\ln\left(\frac{32e}{\epsilon}\right)+\ln\left(\frac{18}{\delta}\right)\right)+(m+1)T$ and define $M = \left\lfloor \frac{\lfloor M'/m+1\rfloor}{T}\right\rfloor \geq \frac{256}{\epsilon^2}\left(2d\ln\left(\frac{32e}{\epsilon}\right)+\ln\left(\frac{18}{\delta}\right)\right)$

Let $v_1,\ldots,v_{M'}$ be the sample generated by f with initial distribution π_0. Relabel this sample as $\tilde{X}^{(0)},\ldots,\tilde{X}^{(\lfloor M'/m+1\rfloor)}$ where $\tilde{X}^{(i)} = (v_{(m+1)(i-1)+1},\ldots,v_{(m+1)i})$. We define the sub-sample $X^{(0)},\ldots,X^{(M)}$ of $\tilde{X}^{(0)},\ldots,\tilde{X}^{(M)}$ such that $X^{(i)} = X^{(iT)}$. By construction, $X^{(0)},\ldots,X^{(M)}$ satisfies the conditions of lemma 5 so there there exist $Z^{(1)},\ldots,Z^{(M)}$ i.i.d from $\pi^{(1)}$ with $\mathbb{P}[Z^{(i)}\neq X^{(i)}]\leq\frac{\epsilon\delta}{12}$. Crucially, this last property gives us that all $g\in\mathcal{G}$ have similar error. Denote $Z=\{Z^{(1)},\ldots,Z^{(M)}\}$ and $X = X^{(1)},\ldots,X^{(M)}$. Then:

$$
\mathbb{P}\left[\exists g\in\mathcal{G}\text{ s.t. }|\hat{er}(X,g)-\hat{er}(Z,g)|<\frac{\epsilon}{2}\right]\geq 1-\frac{\delta}{2}.
$$

Indeed for any $g\in\mathcal{G}$:

$$
\begin{aligned}
|\hat{er}(X,g)-\hat{er}(Z,g)| &= \left|\frac{1}{M}\sum_{i\in[M]} G^2(Z^{(i)}_{m-n+1:m}) - G^2(X^{(i)}_{1:m}) - 2Z^{(i)}_{m+1}G(Z^{(i)}_{m-n+1:m}) + 2X^{(i)}_{m+1}G(X^{(i)}_{m-n+1:m})\right| \\
&\leq \frac{1}{M}\sum_{i\in[M]}\left|G^2(Z^{(i)}_{m-n+1:m}) - G^2(X^{(i)}_{m-n+1:m}) - 2Z^{(i)}_{m+1}G(Z^{(i)}_{m-n+1:m}) + 2X^{(i)}_{m+1}G(X^{(i)}_{m-n+1:m})\right| \\
&\leq \frac{1}{M}\sum_{i\in[M]} 3\,\mathbb{1}[Z^{(i)}\neq X^{(i)}]
\end{aligned}
$$

where in the last inequality we use the fact that $0\leq G, Z^{(i)}_{m+1}, X^{(i)}_{m+1}\leq 1$. By a simple Markov inequality:

$$\mathbb{P}\left[\exists g \in \mathcal{G} \text{ s.t. } |\hat{er}(X,g) - \hat{er}(Z,g)| \geq \frac{\epsilon}{2}\right] \leq \mathbb{P}\left[\frac{3}{M}\sum_{i\in[M]} \mathbb{1}[Z^{(i)} \neq X^{(i)}] \geq \frac{\epsilon}{2}\right]$$
$$\leq \frac{6}{\epsilon M}\sum_{i\in[M]} \mathbb{P}[Z^{(i)} \neq X^{(i)}] \leq \frac{\delta}{2}$$

To finish the proof, the above is used to show that if learning on X fails then learning on Z must fail as well. Suppose there exists $g \in \mathcal{G}$ such that $|er(f,g) - \hat{er}(X,g)| \geq \epsilon$ and $\forall g \in \mathcal{G}\ |\hat{er}(X,g) - \hat{er}(Z,g)| \leq \frac{\epsilon}{2}$, then by triangle inequality it holds that $|er(f,g) - \hat{er}(Z,g)| \geq \frac{\epsilon}{2}$. Hence:

$$\mathbb{P}[g \in \mathcal{G} \text{ s.t } |er(f,g) - \hat{er}(X,g)| \geq \epsilon] \leq \mathbb{P}\left[\exists g \in \mathcal{G} \text{ s.t } |er(f,g) - \hat{er}(Z,g)| \geq \frac{\epsilon}{2}\right]$$
$$+ \mathbb{P}\left[\exists g \in \mathcal{G} \text{ s.t. } |\hat{er}(X,g) - \hat{er}(Z,g)| \geq \frac{\epsilon}{2}\right].$$

By the above calculations and theorem 5.1 of Kearns and Schapire [1990], since $M \geq \frac{256}{\epsilon^2}\left(2d\ln\left(\frac{32e}{\epsilon}\right) + \ln\left(\frac{18}{\delta}\right)\right)$ then:

$$\mathbb{P}[g \in \mathcal{G} \text{ s.t } |er(f,g) - \hat{er}(X,g)| \geq \epsilon] \leq \frac{\delta}{2} + \frac{\delta}{2} = \delta.$$

□

References

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, USA, 1st edition, 2009. ISBN 052111862X.

F. den Hollander. Probability theory : The coupling method, 2012. URL `https://api.semanticscholar.org/CorpusID:32225702`.

Y. Hao, A. Orlitsky, and V. Pichapati. On learning markov chains. *CoRR*, abs/1810.11754, 2018. URL `http://arxiv.org/abs/1810.11754`.

M. Kearns and R. Schapire. Efficient distribution-free learning of probabilistic concepts. In *Proceedings [1990] 31st Annual Symposium on Foundations of Computer Science*, pages 382–391 vol.1, 1990. doi: 10.1109/FSCS.1990.89557.

B. Yu. Rates of convergence for empirical processes of stationary mixing sequences. *The Annals of Probability*, 22(1):94–116, 1994. ISSN 00911798, 2168894X. URL `http://www.jstor.org/stable/2244496`.

5 Appendix: A Quick Note on the Existence of a Stationary Distribution of $\mathbb{M}$

$\mathbb{M}$ has a unique stationary distribution if it is aperiodic and irreducible. In other words, $\mathbb{M}$ can go from any state to any other state in finite time (irreducible) and it does not follow a periodic pattern (aperiodic). For these conditions to be met, it is sufficient that $F_u(v_{1:m})$ is every so slightly larger than 0 for all $u \in \mathcal{V}$ and $v_{1:m} \in \mathcal{V}$. This often holds in practice as models usually do not return exactly 0 for any logits. Indeed if this holds, $\mathbb{M}$ is clearly irreducible, and to see that $\mathbb{M}$ is aperiodic consider that for some $v \in \mathcal{V}$, $F_v(v, v, \ldots, v) > 0$ hence all states must have period 1.

It will also be useful to note that if $\mathbb{M}$ has a stationary distribution, the stationary distribution is also its limiting distribution, i.e. for any initial distribution (represented as a vector) $\pi_0 \in [0,1]^{\mathcal{V}^m}$:

$$\pi = \lim_{M \to \infty} P^M \pi_0.$$

Thus, it is still possible to construct a reasonable "limiting" distribution π if $\mathbb{M}$ is not irreducible or aperiodic. For one, since $\mathcal{V}^m$ is finite, $\mathbb{M}$ has at least one recurrent class which implies that we can ignore any $v_{1:m} \in \mathcal{V}^m$ which are outside this class as the probability they will appear in a sequence generated by f goes to 0 as the sequence length increases. Hence, $\mathbb{M}$ can be assumed to be irreducible if we simply restrict ourselves to this recurrent class. Further, if $\mathcal{F}$ is periodic with some period ρ (which must be finite since at least one state in the irreducible class is recurrent) then we can redefine P to be the ρ step transition matrix (i.e. $P \mapsto P^\rho$) which will be aperiodic. Thus π can be defined as the distribution over the recurrent class (again represented as a vector) such that $P^\rho \pi = \pi$.