Unveiling Induction Heads: Provable Training Dynamics and Feature Learning in Transformers

Siyu Chen[*]
Department of Statistics and Data Science,
Yale University
siyu.chen.sc3226@yale.edu

Heejune Sheen[*]
Department of Statistics and Data Science,
Yale University
heejune.sheen@yale.edu

Tianhao Wang
Toyota Technological Institute at Chicago
tianhao.wang@ttic.edu

Zhuoran Yang
Department of Statistics and Data Science,
Yale University
zhuoran.yang@yale.edu
*

Abstract

In-context learning (ICL) is a cornerstone of large language model (LLM) functionality, yet its theoretical foundations remain elusive due to the complexity of transformer architectures. In particular, most existing work only theoretically explains how the attention mechanism facilitates ICL under certain data models. It remains unclear how the other building blocks of the transformer contribute to ICL. To address this question, we study how a two-attention-layer transformer is trained to perform ICL on n-gram Markov chain data, where each token in the Markov chain statistically depends on the previous n tokens. We analyze a sophisticated transformer model featuring relative positional embedding, multi-head softmax attention, and a feed-forward layer with normalization. We prove that the gradient flow with respect to a cross-entropy ICL loss converges to a limiting model that performs a generalized version of the "induction head" mechanism with a learned feature, resulting from the congruous contribution of all the building blocks. In the limiting model, the first attention layer acts as a *copier*, copying past tokens within a given window to each position, and the feed-forward network with normalization acts as a *selector* that generates a feature vector by only looking at informationally relevant parents from the window. Finally, the second attention layer is a *classifier* that compares these features with the feature at the output position, and uses the resulting similarity scores to generate the desired output. Our theory is further validated by simulation experiments.

1 Introduction

In-context learning (ICL) (Brown et al., 2020) has emerged as a crucial aspect of large language model (LLM) (Radford et al., 2019; Brown et al., 2020; Achiam et al., 2023; Anthropic, 2023; Team et al., 2023) functionality, enabling pre-trained LLMs to solve user-specified tasks during inference without updating model parameters. In ICL, a pre-trained LLM, typically a transformer, receives prompts containing a few demonstration examples sampled from a task-specific distribution and produces the desired output for that task. This capability is noteworthy because the tasks addressed during the ICL might not be part of the original training data set. The success of ICL requires the LLM to perform certain learning processes during inference.

[*] equal contribution

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

https://doi.org/10.52202/079017-2127

Although many previous works aim to demystify ICL from either empirical or theoretical perspectives, the theoretical foundations of ICL remain elusive. This is primarily due to the complexity of transformer architectures, which integrate token and position embeddings, multiple layers of multi-head softmax attention, layer normalization, and feedforward neural networks. When it comes to understanding how the ICL ability emerges in transformers after training, existing works often focus on simplified models, such as linear attention mechanisms or single-layer transformers (Von Oswald et al., 2023), and ICL tasks are typically confined to linear regression (Akyürek et al., 2023). This leaves a gap in understanding how full-fledged transformer architectures facilitate ICL of more complex tasks, especially when latent causal structures exist among the tokens in a sequence.

In this paper, our aim is to narrow this gap by studying **how a two-attention-layer transformer is trained to perform ICL of an n-gram Markov chain model**, where each token in the Markov chain statistically depends on the n tokens before it, known as the parent set. Specifically, we consider a transformer model with relative positional embedding (RPE) (He et al., 2020), multi-head softmax attention, and a feed-forward network (FFN) layer with normalization. We employ such a transformer model to predict the $(L+1)$-th token of an n-gram Markov chain, with the first L tokens given as the prompt, where $L+1$ is the sequence length. Here the L-token sequence is sampled from a random Markov chain model, where a random transition kernel obeying the n-gram Markov property is used to generate sequences. The token sequence is fed into the transformer model, which outputs a probability distribution over the vocabulary set to predict the $(L+1)$-th token. To train the transformer model, we sample token sequences from these random Markov chain models and minimize the cross-entropy loss between the predicted token distribution and the true token distribution.

Under this setting, we aim to answer the following three questions: (i) *Does the gradient flow with respect to the cross-entropy loss converge during training?* (ii) *If yes, how does the limiting model perform ICL?* (iii) *How do the building blocks of the transformer model contribute to ICL?*

Main Results. We provide an affirmative answer to the Question (i) by proving that the gradient flow converges during training. In particular, we identify three phases of training dynamics: in the first stage, FFN learns the potential parent set; in the second stage, each attention head of the first multi-head softmax attention layer learns to focus on a single parent token selected by FFN; and in the final stage, the parameter of the second attention layer increases, and the transformer approaches the limiting model. Moreover, for Questions (ii) and (iii), we show that the limiting model performs a specialized form of exponential kernel regression, dubbed "**generalized induction head**", which requires the congruous contribution of all the building blocks. Specifically, the first attention layer acts as a *copier*, copying past tokens within a given window to each position. The FFN layer acts as a *selector* that generates a feature vector by only looking at informationally relevant parents from the window according to a modified χ^2-mutual information. Finally, the second attention layer is an *exponential kernel classifier* that compares the features at each position with those created for the output position $L+1$, and uses the resulting similarity scores to generate the desired output. When specialized to the case where $n=1$, the limiting model selects the true parent token and implements the *induction head* mechanism (Elhage et al., 2021). In this case, we recover the theory in Nichani et al. (2024). Our theory is complemented by numerical experiments, which validate the three-phase training dynamics and mechanism of generalized induction head.

To our best knowledge, our work is the first to provide a comprehensive understanding of how ICL is empowered by a collaboration of different building blocks in a transformer model. In particular, we identify the pivotal roles played by RPE in the copier component, the FFN layer with normalization in the selector component, and attention in the classifier component. We believe our work will shed light on the theoretical understanding of ICL for more complicated tasks.

Related Works. Our work adds to the rapidly growing literature on understanding in-context learning by transformers. We defer an in-depth discussion on related works in Appendix §A due to the page limit.

Roadmap. The rest of the paper is organized as follows: We introduce the problem setup of ICL of Markov chains in §2. Then in §3, we present the main theoretical results and related discussions. A proof sketch is provided in §D. Finally, we present corresponding experiment results in §B, and the detailed proofs are deferred to the Appendix.

Notation. We denote by $e_1, \ldots, e_d$ the standard basis vectors in $\mathbb{R}^d$ and by $\mathbf{1}$ the all-one vector in $\mathbb{R}^d$. We denote by $\sigma(\cdot)$ the softmax function such that the i-th coordinate of $\sigma(x)$ is $\sigma_i(x) =$

https://doi.org/10.52202/079017-2127

$\exp(x_i)/\sum_{l=1}^{L}\exp(x_l)$ for $x \in \mathbb{R}^L$. By default, the softmax operation will always be applied row-wise. For any integer $n > 0$, we denote $[n] := \{1, \ldots, n\}$. For a vector $w \in \mathbb{R}^M$, we denote by w_i the i-th entry of w and w_{-i} the $(M+1-i)$-th entry of w for positive integer $i \in [M]$. For a matrix W, we denote by $W(i,j)$ the entry at the i-th row and j-th column of W. For two vectors u and v, we write u/v as the vector obtained by taking element-wise division between u and v. We denote by $a \vee b$ and $a \wedge b$ the maximum and minimum of a and b, respectively. We denote by $x_{s:t}$ the sequence $\{x_s, x_{s+1}, \ldots, x_t\}$. For a class $\mathcal{X}$, we denote by $\Delta(\mathcal{X})$ the space of probability measures over $\mathcal{X}$. We use the standard big O notation throughout the paper.

2 Problem Setup: In-Context Learning of Markov Chains

In this section, we present the details of the problem setting. In particular, we first introduce the statistical problem of ICL of n-gram Markov chains in §2.1 and then lay out the details of the transformer model in §2.2.

2.1 In-Context Learning and n-Gram Markov Chains

We study how autoregressive transformers are trained to perform in-context learning (ICL). A pre-trained transformer can be viewed as a conditional distribution $f_{\texttt{tf}}(\cdot \mid \texttt{prompt})$ over a finite vocabulary set $\mathcal{X}$, where $\texttt{prompt}$ is a sequence of tokens in $\mathcal{X}$. We consider an in-context unsupervised learning problem where the pre-trained transformer $f_{\texttt{tf}}$ is used to predict the $(L+1)$-th token x_{L+1} with the first L tokens being the prompt. Here L is a fixed number and the joint distribution of the sequence $x_{1:(L+1)}$ is sampled from a random n-gram Markov chain. In other words, with $x_{1:(L+1)}$ sampled from some distribution, we evaluate how well $f_{\texttt{tf}}(\cdot \mid x_{1:L})$ predicts the distribution of x_{L+1}.

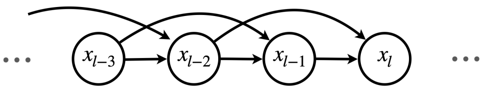


Figure 1: A two-gram Markov chain with parent set $\texttt{pa} = \{-1, -2\}$.

n-Gram Markov Chains. We assume the data comes from a mixture of n-gram Markov chain model, denoted by a tuple $(\mathcal{X}, \texttt{pa}, \mathcal{P}, \mu_0)$, where $\mathcal{X}$ is the state space and $\texttt{pa} = (-r_1, \ldots, -r_n)$ is the parent set with positive integers $r_1 < r_2 < \cdots < r_n$. That is, for each $l > r_n$, x_l only statistically depends on $(x_{l-r_n}, \ldots, x_{l-r_1})$, which is denoted by $X_{\texttt{pa}(l)}$ and referred to as the parent tokens of x_l. We let $d = |\mathcal{X}|$ denote the vocabulary size. Moreover, $\mathcal{P}$ is a probability distribution over the set of Markov transition kernels respecting the parent structure specified by $\texttt{pa}$, and μ_0 is the joint distribution of the first r_n tokens $x_{1:r_n}$. Note that the size of the parent set n can be smaller than or equal to r_n. Thus, the sequence $x_{1:(L+1)}$ is generated as follows: (i) sample initial r_n tokens $(x_1, \ldots, x_{r_n}) \sim \mu_0$, (ii) sample a random transition kernel $\pi \sim \mathcal{P}$, where $\pi \colon \mathcal{X}^n \to \Delta(\mathcal{X})$, and (iii) sample token $x_l \sim \pi(\cdot \mid X_{\texttt{pa}(l)})$ for $l = r_n + 1, \ldots, L+1$. See Figure 1 for an illustration of the generating model of $x_{1:(L+1)}$.

Cross-Entropy Loss. When $x_{1:(L+1)}$ is generated, $x_{1:L}$ is fed into the transformer $f_{\texttt{tf}}$ to predict x_{L+1}. To assess the performance of ICL, we adopt the population cross-entropy (CE) loss

$$\mathcal{L}(f_{\texttt{tf}}) = -\mathbb{E}_{\pi\sim\mathcal{P}, x_{1:(L+1)}}\big[\log\big(f_{\texttt{tf}}(x_{L+1} \mid x_{1:L}) + \epsilon\big)\big], \tag{2.1}$$

where $\epsilon > 0$ is a small constant introduced for numerical stability and in the sequel we will take $\varepsilon = O(L^{-1/2})$. Here, the expectation is taken with respect to the joint distribution of $x_{1:(L+1)}$ (including the randomness of $\pi \sim \mathcal{P}$). When setting $\epsilon = 0$, we note that minimizing this cross-entropy loss is equivalent to minimizing the KL divergence

$$\mathbb{E}_{\pi\sim\mathcal{P}, x_{1:L}}\big[\texttt{KL}(\pi(\cdot \mid X_{\texttt{pa}(L+1)}) \,\|\, f_{\texttt{tf}}(\cdot \mid x_{1:L}))\big].$$

As a remark, we also relax a condition in Nichani et al. (2024) where the last token x_L has to be resampled from a uniform distribution. In addition, our analysis can also be extended to sequential CE loss, which corresponds to predicting every token in the sequence given the past rather than just the last token x_{L+1}. This is closer to the training paradigm used in practice (Brown et al., 2020). See §C.4 for a further discussion on the sequential CE loss.

2.2 A Two-Layer Transformer Model

We consider a class of two-attention-layer transformer model, denoted by $\texttt{TF}(M, H, d, D)$, which incorporates Relative Positional Embedding (RPE) (He et al., 2020), Multi-Head Attention (MHA) (Vaswani et al., 2017), and a Feed-Forward network (FFN) with normalization. Here M is an integer that specifies the window size of RPE, H is the number of heads in the first attention layer, d is the vocabulary size, and D is an integer that controls the complexity of FFN. The details of $\texttt{TF}(M, H, d, D)$ are as follows.

Token Embedding, Input and Output. Note that each token takes values in $\mathcal{X}$ with $d = |\mathcal{X}|$. We embed the tokens into one-hot vectors in $\mathbb{R}^d$, and thus we can identify $\mathcal{X}$ as the canonical basis in $\mathbb{R}^d$, i.e., $\mathcal{X} = \{e_1, \ldots, e_d\}$. A transformer model can be viewed as a mapping from $\mathbb{R}^{(L+1)\times d}$ to $\Delta(\mathcal{X})$. In particular, given the input sequence $x_{1:L}$, we denote $X = (x_1, \ldots, x_L)^\top \in \mathbb{R}^{L\times d}$, and we append a zero vector $\mathbf{0} \in \mathbb{R}^d$ to the sequence, and define $\widetilde{X} = (x_1, \ldots, x_L, \mathbf{0})^\top \in \mathbb{R}^{(L+1)\times d}$. The transformer takes $\widetilde{X}$ as input and outputs a probability distribution over $\mathcal{X}$.

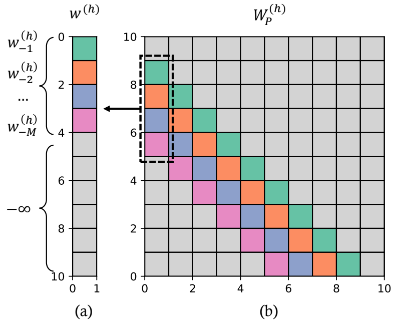


Figure 2: Illustration of the relationship between RPE vector $w^{(h)}$ and corresponding matrix $W_P^{(h)}$.

Relative Positional Embedding. In each head of the first attention layer, we adopt RPE to incorporate positional information. Specifically, RPE is parameterized by a vector $w = (w_{-M}, \ldots, w_{-1})^\top \in \mathbb{R}^M$, and it assigns a scalar value $W_P(i,j)$ to a pair of positions (i,j) satisfying

$$W_P(i,j) = w_{j-i} \ \text{ if } \ i-j \in \{1, \ldots, M\},$$
$$W_P(i,j) = -\infty \ \text{ if } \ j \geq i \ \text{ or } \ |j-i| > M.$$

In other words, as illustrated in Figure 2, the i-th token only attends to tokens with indices in $\{i-1, \ldots, i-M\}$, referred to as the *length-M window of the i-th token*, and the trainable vector w determines the value of positional embedding. Here, we use $-k$ to index the last k-th position.

The First Attention Layer. The input sequence is processed by the first attention layer with H parallel heads. In all heads, we discard the token information and only use RPE to compute the attention score. Specifically, each attention head h maps $\widetilde{X}$ into a sequence in $\mathbb{R}^d$ with length $L+1$, denoted by $V^{(h)} = (v_1^{(h)}, \ldots, v_{L+1}^{(h)})^\top \in \mathbb{R}^{(L+1)\times d}$. For any $l \in [L+1]$, $v_l^{(h)}$ is computed via

$$v_l^{(h)} = \sum_{j=1}^{L} \sigma_j\big(W_P^{(h)}(l,\cdot)\big) \cdot x_j = \sum_{j=1}^{L} \frac{\exp\big(W_P^{(h)}(l,j)\big) \cdot x_j}{\sum_{k=1}^{L} \exp\big(W_P^{(h)}(l,k)\big)}. \tag{2.2}$$

That is, we use the RPE parameter $W_P^{(h)}$ to construct a weighted sum over the input sequence at each position $l \in [L+1]$. Here $W_P^{(h)}$ is the RPE matrix of the h-th head.

Feed-Forward Network with Normalization. Following the first attention layer, we concatenate the outputs of the H attention heads and define $V = (V^{(1)}, \ldots, V^{(H)}) \in \mathbb{R}^{(L+1)\times Hd}$. Here we abuse the notation and write $V = (v_1, \ldots, v_{L+1})^\top$, i.e., each v_l is the l-th row of V. For any vector $v \in \mathbb{R}^{Hd}$, we can split it into $(v^{(1)\top}, \ldots, v^{(H)\top})^\top$ where each block $v^{(h)} \in \mathbb{R}^d$. For embedding dimension d_e, each vector of V is passed through an FFN $\phi(\cdot) : \mathbb{R}^{Hd} \to \mathbb{R}^{d_e}$, which specifies a polynomial kernel such that for any $v, v' \in \mathbb{R}^{Hd}$, we have

$$\langle \phi(v), \phi(v') \rangle = \sum_{\mathcal{S} \in [H]_{\leq D}} c_{\mathcal{S}}^2 \cdot \prod_{h \in \mathcal{S}} \langle v^{(h)}, v'^{(h)} \rangle. \tag{2.3}$$

Here, the low-degree parent set $[H]_{\leq D} := \{\mathcal{S} \subseteq [H] : |\mathcal{S}| \leq D\}$ contains all subsets of $[H]$ with cardinality at most D, and $\{c_{\mathcal{S}} : \mathcal{S} \in [H]_{\leq D}\}$ are the corresponding trainable parameters of $\phi(\cdot)$. Therefore, the FFN $\phi(\cdot)$ specifies a kernel on the output of the multihead attention which induces a special inner product structure. While (2.3) characterizes $\phi(\cdot)$ implicitly, we provide an explicit construction of $\phi(\cdot)$ in Lemma C.1 as a vector-valued mapping whose entries are monomials of the

https://doi.org/10.52202/079017-2127

input's entries. Moreover, the complexity of $\phi(\cdot)$ is controlled by the maximum degree D, which also influences the embedding dimension d_e as we show in the construction.

Furthermore, to control the magnitude of the FFN outputs, we normalize $\phi(\cdot)$ by letting $u_l = \phi(v_l)/\sqrt{C_D}$ for all $l \in [L+1]$, where we define $C_D = \sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2$. Such a normalization scheme is motivated by the standard layer normalization (Ba et al., 2016) in transformer architectures. To motivate the use of $\sqrt{C_D}$ as the normalization, consider a special case where the positional embeddings, after the softmax function, produce attention weights that are close to one-hot for each head. Then $v_l^{(h)}$ in (2.2) is equal to some token in $x_{1:L}$. As a result, each v_l consists of H tokens and

$$\|\phi(v_l)\|_2 = \sqrt{\sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2 \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_l^{(h)}\rangle} = \sqrt{C_D}.$$

Thus, u_l is roughly equivalent to the output of the layer normalization $\phi(v_l)/\|\phi(v_l)\|_2$ (without trainable parameters). Although our theoretical analysis and simulations focus on this simplified version of layer normalization, our additional experiments in §B.2 demonstrate that it aligns well with the performance of the actual layer normalization.

The Second Attention Layer. The normalized vector sequence $U = (u_1, \ldots, u_{L+1})^\top$ and the original sequence $\widetilde{X}$ are then fed into the second attention layer to generate the final output. In particular, u_{L+1} is used as the query to compare with the keys $\{u_{M+1}, \ldots, u_L\}$, and the resulting attention scores are used to aggregate the values $x_{(M+1):L}$. This attention layer has a single head and a scalar trainable parameter a. We let $U_{1:L} = (u_1, \ldots, u_L)^\top \in \mathbb{R}^{L\times d_e}$ and denote by $\texttt{Mask}(\cdot)$ the mask that sets every entry of the first M rows of a matrix to be $-\infty$. The final output is given by

$$y = \sum_{j=M+1}^{L} \sigma_j\big(a \cdot u_{L+1}^\top \texttt{Mask}(U_{1:L}^\top)\big) \cdot x_j = \sum_{j=M+1}^{L} \frac{\exp\big(a \cdot u_{L+1}^\top u_j\big) \cdot x_j}{\sum_{k=M+1}^{L} \exp\big(a \cdot u_{L+1}^\top u_k\big)}. \tag{2.4}$$

Note that the softmax function in (2.4) yields a probability distribution over $[L]$ and that $x_{1:L}$ is a sequence of one-hot vectors. Thus y in (2.4) is a probability distribution over $\mathcal{X}$. The mask operator is included here just to simplify our analysis while in the experiments we are not using the mask.

In summary, given the input $\widetilde{X} \in \mathbb{R}^{(L+1)\times d}$, in the matrix form, our transformer model $\texttt{TF}(M, H, d, D)$ consecutively applies the following operations:

$$\begin{aligned}
&\textbf{First Attention:} && V^{(h)} = \sigma(W_P^{(h)})\widetilde{X} && \in \mathbb{R}^{(L+1)\times d},\ \forall h \in [H];\\
&\textbf{Concatenate:} && V = [V^{(1)}, \ldots, V^{(H)}] && \in \mathbb{R}^{(L+1)\times Hd};\\
&\textbf{FFN \& Normalize:} && U = \phi(V)/\sqrt{C_D} && \in \mathbb{R}^{(L+1)\times d_e};\\
&\textbf{Second Attention:} && y^\top = \sigma\big(a \cdot u_{L+1}^\top \texttt{Mask}(U_{1:L}^\top)\big)X && \in \mathbb{R}^{1\times d}.
\end{aligned} \tag{2.5}$$

The trainable parameters of the above transformer model are denoted by

$$\Theta = \big\{a, \{w_{-1}^{(h)}, \ldots, w_{-M}^{(h)}\}_{h\in[H]}, \{c_{\mathcal{S}} : \mathcal{S} \in [H]_{\le D}\}\big\}.$$

We remark that the transformer model in (2.5) is known as a disentangled transformer (Friedman et al., 2024), which is a version of the transformer model that is more amenable for theoretical analysis. One thing to be noted is that there is a residual connection that directly copies $\widetilde{X}$ to the output of the FFN & Normalize block, which gives us $[U, \widetilde{X}]$, and the second attention layer will treat the copied $\widetilde{X}$ as the value in the attention mechanism. We omit the residual connection in the above paradigm for notation simplicity. As shown in Nichani et al. (2024), any standard transformer model can be expressed as a disentangled transformer by specializing the attention weights to allow feature concatenation.

Our goal is to investigate whether the transformer model $\texttt{TF}(M, H, d, D)$ can perform ICL over n-gram Markov chains and further, whether such capability can be learned from data with common training algorithms like gradient descent.

3 Theoretical Results

In this section, we present the theoretical results. We first show in §3.1 and §C.1 that there exists a transformer in $\texttt{TF}(M, H, d, D)$ that implements a generalized "induction head" mechanism (Olsson et al., 2022) with a learned feature, which serves as a natural algorithm for learning n-gram Markov chains. Then in §3.2 we prove that the gradient flow in (3.4) finds such a desired model asymptotically.

3.1 Generalized Induction Head Mechanism for Learning n-Gram Markov Chains

Recall that we define the mixture of n-gram Markov chain model $(\mathcal{X}, \texttt{pa}, \mathcal{P}, \mu_0)$ in §2.1, where $\mathcal{P}$ is a distribution over the Markov transition kernels. For regularity, we assume existence of a unique stationary distribution for any $\pi \in \operatorname{supp}(\mathcal{P})$, where a rigorous statement is deferred to Assumption 3.5. We also assume the window size $M > r_n$. For any n-gram Markov chain with transition kernel $\pi \sim \mathcal{P}$, we let $\mu^\pi \in \Delta(\mathcal{X}^{M+1})$ denote the stationary distribution of the Markov chain over a window of size $M + 1$. Here we use $\{z_\ell\}_{l\geq 1}$ to denote a random sequence of tokens generated by the Markov chain. Then μ^π denotes the joint distribution of a block of $M + 1$ tokens $(z_{l-M}, \dots, z_{l-1}, z_l)$ under the stationary distribution of π, where $l > M$ is an integer.

In the following, we introduce a generalized induction head (GIH) estimator for the task of predicting x_{L+1} given $x_{1:L}$, which is based on the following simple idea: x_{L+1} *should be similar to a previous token* x_l *if their parents are similar*. As the parent set $\texttt{pa}$ is unknown, GIH adopts an information-theoretic criterion to select a subset of previous tokens as a proxy of the parents. Specifically, GIH uses a modified version of χ^2-mutual information, which is defined as follows.

Definition 3.1 (Modified χ^2-Mutual Information)**.** *We take a length-$(M + 1)$ windows $(z_{l-M}, \dots, z_{l-1}, z_l)$ for some $l > M$ and suppose the sequence is sampled from stationary distribution μ^π with $\pi \sim \mathcal{P}$. Let $Z = (z_{l-M}, \dots, z_{l-1})$. For any subset $\mathcal{S} \subseteq [M]$, we use $Z_{-\mathcal{S}}$ to denote the subvector of Z containing entries of the form z_{l-s}, $\forall s \in \mathcal{S}$. For instance, suppose $\mathcal{S} = \{2, 5\}$, then $Z_{-\mathcal{S}} = (z_{l-5}, z_{l-2})$. The modified χ^2-mutual information for $\mathcal{S}$ is defined as*

$$\widetilde{I}_{\chi^2}(\mathcal{S}) = \mathbb{E}_{\pi\sim\mathcal{P},(z,Z)\sim\mu^\pi}\Bigg[\bigg(\sum_{e\in\mathcal{X}} \frac{[\mu^\pi(z=e\,|\,Z_{-\mathcal{S}})]^2}{\mu^\pi(z=e)} - 1\bigg)\cdot \mu^\pi(Z_{-\mathcal{S}})\Bigg], \tag{3.1}$$

where $\mu^\pi(z = \cdot \mid Z_{-\mathcal{S}})$ is the conditional distribution of z induced by μ^π given the partial history $Z_{-\mathcal{S}}$, and $\mu^\pi(Z_{-\mathcal{S}}), \mu^\pi(z)$ are the marginal distributions of $Z_{-\mathcal{S}}$ and z under $(z, Z) \sim \mu^\pi$.

Intuitively, $\widetilde{I}_{\chi^2}(\mathcal{S})$ is modified from the vanilla χ^2-mutual information (χ^2-MI) between two random variables (Polyanskiy and Wu, 2024) and quantifies how much information the partial history $Z_{-\mathcal{S}}$ contains about z. In particular, we incorporate an additional $\mu^\pi(Z_{-\mathcal{S}})$ term that decreases with the growing size of $\mathcal{S}$. To see the rationality, we first introduce a GIH estimator based on the modified χ^2-mutual information.

Definition 3.2 (Generalized Induction Head)**.** *A GIH estimator with window size $M \in \mathbb{N}$, feature size $D \in \mathbb{N}$ is denoted by $\texttt{GIH}(\cdot; M, D)$, which maps $x_{1:L}$ to a distribution over $\mathcal{X}$. We let $\mathcal{S}^\star$ be the information-optimal subset (referred to as the "information set" in the sequel[2]) of $[M]$ with size no more than D that maximizes the modified χ^2-mutual information $\widetilde{I}_{\chi^2}(\cdot)$ defined in (3.1). That is, we define the information set S^* as*

$$\mathcal{S}^\star = \operatorname{argmax}_{\mathcal{S}\in[M]_{\leq D}} \widetilde{I}_{\chi^2}(\mathcal{S}). \tag{3.2}$$

Then $\texttt{GIH}(x_{1:L}; M, D)$ outputs

$$y^\star := \begin{cases} N^{-1} \cdot \sum_{l=M+1}^{L} x_l \cdot \mathbb{1}(X_{l-\mathcal{S}^\star} = X_{L+1-\mathcal{S}^\star}), & \text{if } N \geq 1, \\ (L-M)^{-1} \cdot \sum_{l=M+1}^{L} x_l, & \text{otherwise.} \end{cases} \tag{3.3}$$

Here, we define $X_{l-\mathcal{S}^\star}$ as the set $\{x_{l-s} : s \in \mathcal{S}^\star\}$ and $N = \sum_{l=M+1}^{L} \mathbb{1}(X_{l-\mathcal{S}^\star} = X_{L+1-\mathcal{S}^\star})$.

Note that $\mathcal{S}^\star$ defined in (3.2) depends on the choices of M and D and serves as a proxy of the unknown parent set $\texttt{pa}$ based on $\widetilde{I}_{\chi^2}(\cdot)$ defined in (3.1). In a nutshell, the GIH estimator checks

[2] With a slight abuse of notation, we also call $X_{l-\mathcal{S}^\star} := (x_{l-s} : s \in \mathcal{S}^\star)$ the information set of the l-th token x_l.

whether the partial histories of $X_{l-\mathcal{S}^\star}$ and $X_{L+1-\mathcal{S}^\star}$ match and aggregate all the tokens x_l that have a matching partial history to predict x_{L+1}. As a remark, using the modified χ^2-MI as the information criterion rules out redundancy in the information set $\mathcal{S}^\star$ in the following sense:

• **$\mathcal{S}^\star$ cannot be a superset of the true parents.** Note that if $\mathcal{S}$ is a superset of the true parent set, by the Markov property, z and Z_{-S} are conditionally independent given the true parents $Z_{\mathtt{pa}}$. Thus, maximizing the vanilla χ^2-mutual information yields multiple maximizers, i.e., all the supersets of the true parent set. However, with the modification in (3.1), any superset yields a strictly smaller $\widetilde{I}_{\chi^2}$ compared to the exact parent set, making them suboptimal.

• **The modified χ^2-MI selects informative partial history.** Even a true parent may bear relatively little information about the target compared to other parents sometimes. Meanwhile, exact match of a larger set of partial history becomes much harder as it tends to appear less frequently in the context sequence, leading to poor estimation accuracy for the estimator in (3.3). The modified χ^2-MI reaches a balance by selecting the informative partial history while penalizing the size of the information set.

The term involving $\mu^\pi(z = \cdot \,|\, Z_{-\mathcal{S}})$ can be viewed as the *signal* part which helps us to find an informative subset $\mathcal{S}$. The term $\mu^\pi(Z_{-\mathcal{S}})$ can be viewed as *penalty on the model complexity* which favors smaller subsets. Thus, the modified χ^2-MI strikes a balance between these two objectives and enables us to find a good proxy $\mathcal{S}^\star$ of $\mathtt{pa}$ when L is finite. Moreover, when L is sufficiently large, we identify two scenarios in which maximizing $\widetilde{I}_{\chi^2}(\cdot)$ yields the true parent set (see §C.7 for details). Moreover, the GIH estimator is a generalization of the induction head mechanism (Elhage et al., 2021) to the stochastic setting with multiple parents, where we give the model more flexibility to learn based on a partial history that does not necessarily correspond to the true parent set. As we will show in §C.1, the GIH mechanism can be implemented by the transformer model.

3.2 Convergence Guarantee of Gradient Flow

In the following, we present the convergence guarantee for gradient flow. To simplify the discussion, we consider the case where $H = M$, meaning there are enough heads to implement the GIH mechanism by having each head copy a unique parent token from a window of size M. Let us first introduce the paradigm of training by gradient flow.

Training Paradigm. Consider training a transformer $\mathtt{TF}(M, H, d, D)$ in (2.5) with $M = H$ to perform ICL on the n-gram Markov chain model introduced in §2.1. Specifically, we define $\mathcal{L}(\Theta)$ as the population cross-entropy loss in (2.1), where the transformer model $f_{\mathtt{tf}}$ is given by (2.5) with a parameter Θ. Ideally, when training the parameter Θ with gradient flow, the dynamics with respect to the loss $\mathcal{L}(\Theta)$ is given by:

$$\partial_t \Theta(t) = -\nabla \mathcal{L}\big(\Theta(t)\big). \tag{3.4}$$

We consider a three-stage training paradigm where, in each stage, only a specific subset of the weights is trained by gradient flow. The three stages are outlined in Table 1. Specifically, in the first stage, we only train the FFN layer via gradient flow while keeping other weights fixed. We then only train the RPE weights in the first attention layer in the second stage. Finally, we only train the weight a in the second attention layer in the last stage, while fixing the rest of the parameters. This training approach is primarily used for analytical convenience; in practice, the entire model can be trained simultaneously, and similar convergence results are reported in §B.2. From a theoretical standpoint, we will also justify the three-stage paradigm in the discussion following Theorem 3.6.

Initialization Conditions. Before presenting our main results about how training by gradient flow induces the GIH structure, let us introduce the following assumption on the initialization of the weights. We define the *information gap* within the D-degree parent set $[H]_{\le D}$ as

$$\Delta \widetilde{I}_{\chi^2} = \widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \max_{S \in [H]_{\le D} \setminus \{\mathcal{S}^\star\}} \widetilde{I}_{\chi^2}(\mathcal{S}), \tag{3.5}$$

where we recall that $\mathcal{S}^\star$ defined in (3.2) maximizes the modified χ^2 mutual information.

Assumption 3.3 (Initialization)**.** *We assume that the following holds at initialization:*

https://doi.org/10.52202/079017-2127

Stage	Weights to Train	Description
I	$\{c_{\mathcal{S}}\}_{\mathcal{S}\in[H]_{\le D}}$ in the FFN layer	Ratio $c_{\mathcal{S}^\star}(t)/c_{\mathcal{S}}(t)$ grows exponentially, learning the low-degree features with $\mathcal{S}^\star$,
II	$\{w^{(h)}\}_{h\in[H]}$ in the RPE of the first attention layer,	$1-\prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)}(t))^2$ decays polynomially training each head in $\mathcal{S}^\star$ to be a copier,
III	a in the weight of the second attention layer	$a(t)$ experiences a two-stage growth, learning the softmax aggregator for GIH,

Table 1: Three-stage training paradigm for gradient flow. Here, the "Weights to Train" column indicates the weights updated in each stage, and the "Description" column summarizes the corresponding results from Theorem 3.6.

1. *For the first attention layer's RPE weights, $w_{-h}^{(h)} \ge w_{-j}^{(h)} + \Delta w$ for all $h, j \in [H]$ with $j \ne h$, where $\Delta w > 0$ is a positive scalar satisfying*

$$\Delta w \ge \log(M-1) - \log\bigg[\Big(1+\Delta\widetilde{I}_{\chi^2}/(14\widetilde{I}_{\chi^2}(\mathcal{S}^\star))\bigg]^{\frac{1}{2H}} - 1\Big). \tag{3.6}$$

2. *The scalar parameter a in the second attention layer satisfies $0 < a \le O(L^{-3/2})$.*

The first assumption on the RPE is used to induce the correspondence between parents and heads during the training by slightly breaking the symmetry between different attention heads. The second assumption on the scale of a ensures that the attention probability given by the second attention layer is close to the uniform distribution over $[L]$. These initialization conditions enable us to derive clean descriptions for the dynamics of the first attention layer and the FFN, shedding light on their respective roles in executing ICL.

We now outline our assumptions on the Markov chain used in the data generation process. Recall that r_n is the largest absolute integer in the parent set $\mathtt{pa}$. For any position l, we define the history $Z = (z_{l-r_n}, \ldots, z_{l-1})$ as the last state and $Z' = (z_{l-r_n+1}, \ldots, z_l)$ as the current state. Since the parent of the new token z_l is already included in Z, Z' is independent of all prior history given Z, forming a Markov chain.

We define P_π as the $d^{r_n} \times d^{r_n}$ transition matrix for this Markov chain, where states are successive r_n-tokens. Each row of P_π is indexed by Z' and each column by Z. The matrix element $P_\pi(Z', Z)$ is thus given by

$$P_\pi(Z', Z) = \pi(z'_l \mid Z_{\mathtt{pa}(l)}) \cdot \mathbb{1}(Z'_{l-r_n+1:-1} = Z_{l-r_n+1:-1}).$$

This means that to transition from Z to Z', all elements of Z' except for z'_{-1} must match the last $r_n - 1$ tokens of Z. The token z'_l is then sampled according to the transition kernel π and depends only on the parent $Z_{\mathtt{pa}(l)}$. The above definition is in fact independent of the position l as the transition kernel π is the same across all positions. Note that P_π is also a stochastic matrix but with zero entries due to the indicator. To proceed, we need the following notion of primitive matrix to state our assumption on P_π.

Definition 3.4 (Primitive Matrix). *A nonnegative and irreducible square matrix P is called primitive if there exists a positive integer k such that all entries of P^k are positive.*

We defer more details about the above definition to §C.3. By the celebrated Perron-Frobenius theorem, if a stochastic matrix P_π is also primitive, then (i) there exists a unique stationary distribution for the Markov chain; (ii) P_π has a unique leading eigenvalue equal to 1, and the corresponding eigenvector is the stationary distribution. Next, we state the assumptions on the mixture of Markov chains for data generation.

Assumption 3.5 (Markov Chain). *For any $\pi \in \mathrm{supp}(\mathcal{P})$, we assume that:*

1. *The transition matrix P_π is primitive. In particular, we assume that there exists $\lambda < 1$ such that the eigenvalue of P_π with the second largest magnitude satisfies $|\lambda_2(P_\pi)| \le \lambda$. Note that $\lambda_2(P_\pi)$ can be complex-valued.*

2. *There exists $\gamma > 0$ such that the transition kernel satisfies $\pi(x \mid X_{\mathtt{pa}}) \geq \gamma$ for any $(x, X_{\mathtt{pa}})$.*

In fact, the second condition $\pi(\cdot \mid X_{\mathtt{pa}}) > \gamma$ already ensures that P_π must be primitive, as is required by the first condition. See Corollary F.14 for details. On the high level, the first assumption guarantees a unique stationary distribution as well as a fast mixing rate of the Markov chain by ensuring a spectral gap for P_π. The second assumption implies a lower bound on the probability for any set $\mathcal{S} \subseteq [M]$ under the stationary distribution, i.e., $\mu^\pi(X_{l-\mathcal{S}}) \geq \gamma^{|\mathcal{S}|}$ for any $l > M$. See Corollary F.15 for details.

Now we are ready to present our main theoretical result on training transformers by gradient flow.

Theorem 3.6 (Convergence of Gradient Flow)**.** *Suppose Assumption 3.3 and Assumption 3.5 hold. Consider $H \geq M$. We set $\varepsilon = L^{-1/2}$ for the cross-entropy loss and assume L is sufficiently large. Then the following holds for the three-stage training of gradient flow:*

Stage I: Parent Selection by FFN. *Let $C_D(t) = \sum_{\mathcal{S} \in [H]_{\leq D}} c_{\mathcal{S}}(t)^2$ and $p_{\mathcal{S}^\star}(t) = c^2_{\mathcal{S}^\star}(t)/C_D(t)$. Then in the first stage with duration $t_1 \asymp C_D(0) \log L/(a(0)\Delta\widetilde{I}_{\chi^2})$, the ratio $c_{\mathcal{S}^\star}/c_{\mathcal{S}}$ grows exponentially fast for any $\mathcal{S} \neq \mathcal{S}^\star$, and $\mathcal{S}^\star$ dominates exponentially fast in the sense that,*

$$1 - p_{\mathcal{S}^\star}(t) \leq (1 - p_{\mathcal{S}^\star}(0)) \cdot \exp\bigl(-(2C_D)^{-1} \cdot a(0) \cdot p_{\mathcal{S}^\star}(0) \cdot \Delta\widetilde{I}_{\chi^2} \cdot t\bigr), \quad \forall t \in [0, t_1).$$

Stage II: Concentration of The First Attention. *Define $\sigma^{(h)}(t) = \sigma(w^{(h)}(t)) \in \mathbb{R}^M$, and let $\sigma_{\min}(t) := \min_{h \in \mathcal{S}^\star} \sigma^{(h)}_{-h}(t)$. Then in the second stage with duration $t_2 - t_1 \asymp L/(a(0)\Delta\widetilde{I}_{\chi^2})$, the first layer's attention heads have attention probabilities concentrated on the optimal information set $\mathcal{S}^\star$ in the sense that for any $t \in [t_1, t_1 + t_2)$,*

$$1 - \prod_{h \in \mathcal{S}^\star} (\sigma^{(h)}_{-h}(t))^2 \leq \frac{2|\mathcal{S}^\star| \cdot (M-1)}{a(0) \cdot \Delta\widetilde{I}_{\chi^2} \cdot \sigma_{\min}(0) \cdot (t - t_1)/2 + \exp(\Delta w) + (M-1)} \wedge 1.$$

Stage III: Growth of The Second Attention. *For some constants c_1, c_2 depending on $(\mathcal{P}, \mathcal{S}^\star)$ with $0 < c_1 < c_2$, there exists a small constant $\delta > 0$ such that the growth of $a(t)$ exhibits the following two sub-stages: (i) When $a(t) \leq \log(c_1/\delta)$, it holds that $\partial a(t) \asymp e^{a(t)}$; (ii) After $a(t)$ has grown such that $a(t) \geq \log(c_2/\delta)$, then $\partial_t a(t) \asymp 1/a(t)$ until it reaches the value $\log L/8$.*

See §D for a proof sketch and §E for the detailed proof. We require that L is sufficiently large, and the specific conditions for L are deferred to §E.1.

Interpretation of Training Dynamics. We empirically verify Theorem 3.6 by conducting a simulation experiment. In particular, we train a transformer with $H = M = 3$ and $D = 2$ based on Markov chain data with $d = 2$, $L = 100$ and $\mathtt{pa} = \{-1, -2\}$. We sample the transition kernel from a Dirichlet prior such that $\mathcal{S}^\star = \{1, 2\}$ also matches the parent set. For more details on this simulation, see §B. The results are shown in Figure 3 and align perfectly with Theorem 3.6. From Theorem 3.6, we can interpret the three stages of training dynamics as follows.

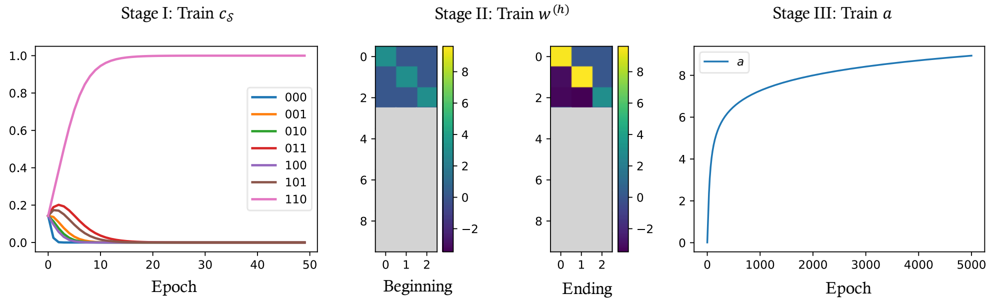
- In the first stage, the training of FFN parameters learns a ***selector*** that selects an informative set $\mathcal{S}^\star$ by realizing the corresponding feature embedding through the polynomial kernel. That is, when t is sufficiently large, we have $p_{\mathcal{S}^\star}(t) \approx 1$ and $p_{\mathcal{S}}(t) \approx 0$ for all $\mathcal{S} \neq \mathcal{S}^\star$. In this case, for any input vectors $v, v' \in \mathbb{R}^{Hd}$, the inner product in (2.3) reduces to

$$\langle \phi(v), \phi(v') \rangle \approx c^2_{\mathcal{S}^\star} \cdot \prod_{h \in \mathcal{S}^\star} \langle v^{(h)}, v'^{(h)} \rangle.$$

  That is, FFN only selects the blocks in $\mathcal{S}^\star$ as the feature. We observe this phenomenon in the experiment, where we set $\mathcal{S}^\star = \{1, 2\}$. As shown in Figure 3-(a), it is clear that $c_{\mathcal{S}^\star}$ immediately dominates the rest of $c_{\mathcal{S}}$'s within only a few gradient epochs.

- In the second stage, we update the parameters of the RPE. This stage turns the first attention layer into a ***copier*** by establishing the correspondence between the attention heads and the parents in the selected $\mathcal{S}^\star$. That is, each attention head copies a particular parent in $\mathcal{S}^\star$. Specifically, when t is sufficiently large, for any $h \in \mathcal{S}^\star$, $\sigma^{(h)}(t) = \sigma(w^{(h)}(t)) \approx 1$. Recalling the construction of RPE, this implies that $v^{(h)}_l$ in (2.2) becomes x_{l-h} for all

https://doi.org/10.52202/079017-2127



(a)

(b)

(c)

Figure 3: An illustration of the transformer parameters during the three-stage training. We train a transformer in $\texttt{TF}(M = 3, H = 3, d = 3, D = 2)$ with $L = 100$, $\texttt{pa} = \{-1, -2\}$. See §B and Figure 4 for more details of the simulation.

$h \in \mathcal{S}^\star$. As shown in Figure 3-(b), in the experiment, the first two heads initialized towards the first two parents will deterministically copy parents -1 and -2 eventually. The third head stays close to its initial value. This head has a negligible effect on the output because $3 \notin \mathcal{S}^\star$ and $p_{\mathcal{S}^\star} \approx 1$.

- After the first two stages are completed, we know that the features constructed approximately satisfy (C.1) up to a proportionality factor. Then, in the final training stage, the scalar weight a in the second attention layer keeps increasing. Thus, this stage learns an exponential kernel ***classifier*** as specified in (C.2). When $a(t)$ is sufficiently large, the learned transformer is close to a classifier that uses covariate-label pairs of the form $(X_{l-\mathcal{S}^\star}, x_l)$ to predict x_{L+1}. In particular, when $a(t)$ goes to infinity, the transformer exactly becomes the GIH mechanism given in Definition 3.2. Moreover, we theoretically prove that the increasing trajectory of $a(t)$ has two stages, where $\mathrm{d}a(t)/\mathrm{d}t$ is initially large and gradually decays, this is also clearly observed in the experiment. See Figure 3-s(c) for details.

In summary, we theoretically show that the limiting model obtained by three-stage training approximately implements the GIH mechanism. We will prove that the difference between these two estimators is at most $O(L^{-1/8})$. We defer the formal statement and proof to §E.5. Moreover, as an answer to the Question (iii) raised in §1, the different components of the transformer architecture are all critical for achieving this: FFN with normalization realizes the ***selector***, the multi-head design of attention supports the ***copier***, and finally, the softmax operation facilitates the exponential kernel ***classifier***. These components work organically as a whole system, yielding the trained transformer's capability of ICL of n-gram Markov chains.

Another takeaway from Theorem 3.6 is a strict separation in the growth rate of these three stages. In particular, the convergence rates of the corresponding components of the transformer model in these three stages range from exponentially fast (Stage I), polynomially fast (Stage II), to logarithmically slow (Stage III). With such two exponential separations of convergence rates, we expect that these three stages naturally arise when we simultaneously train the whole model via gradient descent/flow. We empirically verify this argument and the details are deferred to §B.2.

In §C.7, we provide more intuitive interpretation of the modified χ^2-mutual information, which demonstrates a balance of model complexity and information richness.

4 Conclusion and Future Work

In this paper, we have studied the training dynamics of a two-attention-layer transformer model for learning n-gram Markov chains in an in-context way. Our work opens new directions for developing a rigorous understanding of the transformer models, which includes understanding the induction head mechanism with standard FFN layer and investigating the training dynamics beyond a single loop of this induction head mechanism. We defer readers to §C.8 for more discussions.

5 Acknowledgement

We acknowledge Shaobo Wang for his help with the experiments. We also thank Jason D. Lee, Alex Damian, and Eshaan Nichani for their helpful discussions. Zhuoran Yang acknowledges the support of NSF under the award DMS-2413243.

https://doi.org/10.52202/079017-2127

References

Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S. et al. (2023). Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Ahn, K., Cheng, X., Daneshmand, H. and Sra, S. (2023). Transformers learn to implement preconditioned gradient descent for in-context learning. *arXiv preprint arXiv:2306.00297*.

Ahuja, K., Panwar, M. and Goyal, N. (2023). In-context learning through the bayesian prism. *arXiv preprint arXiv:2306.04891*.

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T. and Zhou, D. (2023). What learning algorithm is in-context learning? investigations with linear models. In *The Eleventh International Conference on Learning Representations*.

Alayrac, J.-B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M. et al. (2022). Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, **35** 23716–23736.

Anthropic (2023). Model card and evaluations for claude models.

Ba, J. L., Kiros, J. R. and Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bai, Y., Chen, F., Wang, H., Xiong, C. and Mei, S. (2023). Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*.

Bietti, A., Cabannes, V., Bouchacourt, D., Jegou, H. and Bottou, L. (2024). Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, **36**.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A. et al. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, **33** 1877–1901.

Cabannes, V., Arnal, C., Bouaziz, W., Yang, A., Charton, F. and Kempe, J. (2024). Iteration head: A mechanistic study of chain-of-thought. *arXiv preprint arXiv:2406.02128*.

Chen, S. and Li, Y. (2024). Provably learning a multi-head attention layer. *arXiv preprint arXiv:2402.04084*.

Chen, S., Sheen, H., Wang, T. and Yang, Z. (2024). Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*.

Chen, S., Yang, D., Li, J., Wang, S., Yang, Z. and Wang, Z. (2022). Adaptive model design for markov decision process. In *International Conference on Machine Learning*. PMLR.

Chen, X. and Zou, D. (2024). What can transformer learn with varying depth? case studies on sequence learning tasks. *arXiv preprint arXiv:2404.01601*.

Cheng, X., Chen, Y. and Sra, S. (2023). Transformers implement functional gradient descent to learn non-linear functions in context. *arXiv preprint arXiv:2312.06528*.

Collins, L., Parulekar, A., Mokhtari, A., Sanghavi, S. and Shakkottai, S. (2024). In-context learning with transformers: Softmax attention adapts to function lipschitzness. *arXiv preprint arXiv:2402.11639*.

Deora, P., Ghaderi, R., Taheri, H. and Thrampoulidis, C. (2023). On the optimization and generalization of multi-head attention. *arXiv preprint arXiv:2310.12680*.

Edelman, B. L., Edelman, E., Goel, S., Malach, E. and Tsilivis, N. (2024). The evolution of statistical induction heads: In-context learning markov chains. *arXiv preprint arXiv:2402.11004*.

Edelman, B. L., Goel, S., Kakade, S. and Zhang, C. (2022). Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*. PMLR.

https://doi.org/10.52202/079017-2127

Elhage, N., Nanda, N., Olsson, C., Henighan, T., Joseph, N., Mann, B., Askell, A., Bai, Y., Chen, A., Conerly, T. et al. (2021). A mathematical framework for transformer circuits. *Transformer Circuits Thread*, **1** 1.

Friedman, D., Wettig, A. and Chen, D. (2024). Learning transformer programs. *Advances in Neural Information Processing Systems*, **36**.

Fu, D., Chen, T.-Q., Jia, R. and Sharan, V. (2023). Transformers learn higher-order optimization methods for in-context learning: A study with linear models. *arXiv preprint arXiv:2310.17086*.

Giannou, A., Rajput, S., Sohn, J.-Y., Lee, K., Lee, J. D. and Papailiopoulos, D. (2023). Looped transformers as programmable computers. In *Proceedings of the 40th International Conference on Machine Learning* (A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato and J. Scarlett, eds.), vol. 202 of *Proceedings of Machine Learning Research*. PMLR.

Giannou, A., Yang, L., Wang, T., Papailiopoulos, D. and Lee, J. D. (2024). How well can transformers emulate in-context newton's method? *arXiv preprint arXiv:2403.03183*.

Guo, T., Hu, W., Mei, S., Wang, H., Xiong, C., Savarese, S. and Bai, Y. (2023). How do transformers learn in-context beyond simple functions? a case study on learning with representations. *arXiv preprint arXiv:2310.10616*.

He, J., Chen, S., Zhang, F. and Yang, Z. (2024). From words to actions: Unveiling the theoretical underpinnings of llm-driven autonomous systems. *arXiv preprint arXiv:2405.19883*.

He, P., Liu, X., Gao, J. and Chen, W. (2020). Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Honovich, O., Shaham, U., Bowman, S. R. and Levy, O. (2022). Instruction induction: From few examples to natural language task descriptions. *arXiv preprint arXiv:2205.10782*.

Huang, Y., Cheng, Y. and Liang, Y. (2023). In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*.

Jelassi, S., Sander, M. and Li, Y. (2022). Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, **35** 37822–37836.

Jeon, H. J., Lee, J. D., Lei, Q. and Van Roy, B. (2024). An information-theoretic analysis of in-context learning. *arXiv preprint arXiv:2401.15530*.

Kim, J. and Suzuki, T. (2024). Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. *arXiv preprint arXiv:2402.01258*.

Li, Y., Huang, Y., Ildiz, M. E., Rawat, A. S. and Oymak, S. (2024). Mechanics of next token prediction with self-attention. In *International Conference on Artificial Intelligence and Statistics*. PMLR.

Li, Y., Li, Y.-F. and Risteski, A. (2023). How do transformers learn topic structure: Towards a mechanistic understanding. *arXiv preprint arXiv:2303.04245*.

Lin, L., Bai, Y. and Mei, S. (2023). Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. *arXiv preprint arXiv:2310.08566*.

Liu, B., Ash, J., Goel, S., Krishnamurthy, A. and Zhang, C. (2022). Transformers learn shortcuts to automata. *ArXiv*, **abs/2210.10749**.

Mahankali, A., Hashimoto, T. B. and Ma, T. (2023). One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*.

Makkuva, A. V., Bondaschi, M., Ekbote, C., Girish, A., Nagle, A., Kim, H. and Gastpar, M. (2024a). Local to global: Learning dynamics and effect of initialization for transformers. *arXiv preprint arXiv:2406.03072*.

Makkuva, A. V., Bondaschi, M., Girish, A., Nagle, A., Jaggi, M., Kim, H. and Gastpar, M. (2024b). Attention with markov: A framework for principled analysis of transformers via markov chains. *arXiv preprint arXiv:2402.04161*.

https://doi.org/10.52202/079017-2127

Meyer, C. D. (2023). *Matrix analysis and applied linear algebra*. SIAM.

Muller, S., Hollmann, N., Arango, S. P., Grabocka, J. and Hutter, F. (2021). Transformers can do bayesian inference. *ArXiv*, **abs/2112.10510**.

Nichani, E., Damian, A. and Lee, J. D. (2024). How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*.

Olsson, C., Elhage, N., Nanda, N., Joseph, N., DasSarma, N., Henighan, T., Mann, B., Askell, A., Bai, Y., Chen, A. et al. (2022). In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*.

Polyanskiy, Y. and Wu, Y. (2024). *Information Theory: From Coding to Learning*. Cambridge University Press.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I. et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, **1** 9.

Rajaraman, N., Bondaschi, M., Ramchandran, K., Gastpar, M. and Makkuva, A. V. (2024a). Transformers on markov data: Constant depth suffices. *arXiv preprint arXiv:2407.17686*.

Rajaraman, N., Jiao, J. and Ramchandran, K. (2024b). Toward a theory of tokenization in llms. *arXiv preprint arXiv:2404.08335*.

Sanford, C., Hsu, D. and Telgarsky, M. (2023). Representational strengths and limitations of transformers. *arXiv preprint arXiv:2306.02896*.

Sheen, H., Chen, S., Wang, T. and Zhou, H. H. (2024). Implicit regularization of gradient flow on one-layer softmax attention. *arXiv preprint arXiv:2403.08699*.

Sinii, V., Nikulin, A., Kurenkov, V., Zisman, I. and Kolesnikov, S. (2023). In-context reinforcement learning for variable action spaces. *arXiv preprint arXiv:2312.13327*.

Song, J. and Zhong, Y. (2023). Uncovering hidden geometry in transformers via disentangling position and context. *arXiv preprint arXiv:2310.04861*.

Tarzanagh, D. A., Li, Y., Thrampoulidis, C. and Oymak, S. (2023a). Transformers as support vector machines. *ArXiv*, **abs/2308.16898**.

Tarzanagh, D. A., Li, Y., Zhang, X. and Oymak, S. (2023b). Max-margin token selection in attention mechanism. *arXiv preprint arXiv:2306.13596*.

Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.-B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A. M., Hauth, A. et al. (2023). Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Thrampoulidis, C. (2024). Implicit bias of next-token prediction. *arXiv preprint arXiv:2402.18551*.

Tian, Y., Wang, Y., Chen, B. and Du, S. (2023a). Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *arXiv preprint arXiv:2305.16380*.

Tian, Y., Wang, Y., Zhang, Z., Chen, B. and Du, S. (2023b). Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. *arXiv preprint arXiv:2310.00535*.

Vasudeva, B., Deora, P. and Thrampoulidis, C. (2024). Implicit bias and fast convergence rates for self-attention. *arXiv preprint arXiv:2402.05738*.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł. and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, **30**.

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A. and Vladymyrov, M. (2023). Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*. PMLR.

Wang, K., Variengien, A., Conmy, A., Shlegeris, B. and Steinhardt, J. (2022). Interpretability in the wild: a circuit for indirect object identification in gpt-2 small. *arXiv preprint arXiv:2211.00593*.

https://doi.org/10.52202/079017-2127

Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M. and Le, Q. V. (2021). Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q. V., Zhou, D. et al. (2022). Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, **35** 24824–24837.

Wu, J., Zou, D., Chen, Z., Braverman, V., Gu, Q. and Bartlett, P. L. (2023). How many pretraining tasks are needed for in-context learning of linear regression? *arXiv preprint arXiv:2310.08391*.

Xie, S. M., Raghunathan, A., Liang, P. and Ma, T. (2021). An explanation of in-context learning as implicit bayesian inference. *arXiv preprint arXiv:2111.02080*.

Zhang, R., Frei, S. and Bartlett, P. L. (2023a). Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*.

Zhang, Y., Liu, B., Cai, Q., Wang, L. and Wang, Z. (2022). An analysis of attention via the lens of exchangeability and latent variable models. *arXiv preprint arXiv:2212.14852*.

Zhang, Y., Zhang, F., Yang, Z. and Wang, Z. (2023b). What and how does in-context learning learn? bayesian model averaging, parameterization, and generalization. *arXiv preprint arXiv:2305.19420*.

Zhou, D., Schärli, N., Hou, L., Wei, J., Scales, N., Wang, X., Schuurmans, D., Cui, C., Bousquet, O., Le, Q. et al. (2022). Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*.

https://doi.org/10.52202/079017-2127

Contents

https://doi.org/10.52202/079017-2127

Organization of The Appendix

The appendices are organized as follows:

- In §A, we present an in-depth discussion on the related works.
- In §B, we discuss the experimental details.
- In §C, we discuss the implementation of GIH mechanism, provide explicit expressions for the FFN realizing a low-degree polynomial kernel, and review basics related to concepts mentioned in the main text.
- In §D, we provide a high-level overview of the proof of our main results.
- In §E, we present the proof for Theorem 3.6.
- In §F, we collect auxiliary results used in the proof of Theorem 3.6.

A Related Works

In Context Learning (ICL). Commercial Large Language Models (LLMs) such as ChatGPT (Brown et al., 2020), GPT-4 (Achiam et al., 2023), and Gemini (Team et al., 2023) typically operate in an autoregressive manner. These models exhibit remarkable ICL capabilities, without requiring further training. Previous research explores various aspects of the in-context learning (ICL) ability of these models. This includes their performance in zero-shot and few-shot learning scenarios (Honovich et al., 2022; Wei et al., 2021), the use of the chain of thought method to enhance reasoning (Wei et al., 2022; Zhou et al., 2022), and learning with multi-modalities (Alayrac et al., 2022). Moreover, recent research highlights the properties and advantages of using transformers beyond the traditional ICL setting, thereby broadening our understanding of their capabilities and applications (Edelman et al., 2022; Li et al., 2023; Jelassi et al., 2022; Sanford et al., 2023; Giannou et al., 2023; Liu et al., 2022; Tarzanagh et al., 2023a,b; Tian et al., 2023b,a; Song and Zhong, 2023; Deora et al., 2023; Chen and Li, 2024; Rajaraman et al., 2024b).

There is a large and growing body of literature on understanding how transformer architecture enables ICL. One strand of research proposes to understand ICL by casting it as a version of Bayesian inference expressed by the transformer architecture. See, e.g., Xie et al. (2021); Muller et al. (2021); Zhang et al. (2022, 2023b); Ahuja et al. (2023); Jeon et al. (2024); He et al. (2024) and the references therein. Another line of work investigates how transformers internally emulate specific algorithms to solve ICL tasks, where Akyürek et al. (2023); Von Oswald et al. (2023); Fu et al. (2023); Ahn et al. (2023); Mahankali et al. (2023); Giannou et al. (2024); Wu et al. (2023) focus on learning with linear regression tasks and Bai et al. (2023); Cheng et al. (2023); Collins et al. (2024); Guo et al. (2023) investigate transformers' capabilities in learning with nonlinear functions. However, all of these works above focus on regression tasks where token (or token pairs) in the prompt sequences are i.i.d. or uncorrelated, which may not capture the more sophisticated data structures in real-world applications.

In addition, to study ICL with correlated data, there is also substantial interest in understanding how ICL operates over data drawn from Markov chains, providing insight into how transformer architectures contribute to ICL in these settings (Edelman et al., 2024; Makkuva et al., 2024b; Chen

https://doi.org/10.52202/079017-2127

and Zou, 2024). Furthermore, Lin et al. (2023); Sinii et al. (2023) show how transformers can solve reinforcement learning problems in an in-context fashion.

While many of the aforementioned works focus on the expressivity of the transformer model on different ICL tasks and the statistical properties of the learned models, understanding training dynamics from an optimization perspective is also crucial for comprehending ICL by transformers. The training dynamics for one-layer attention models have been investigated under different data models for both regression and classification tasks (Zhang et al., 2023a; Huang et al., 2023; Tarzanagh et al., 2023a,b; Kim and Suzuki, 2024; Chen et al., 2024; Vasudeva et al., 2024; Li et al., 2024; Thrampoulidis, 2024; Sheen et al., 2024). These studies offer a thorough characterization of the training process, yet they have limitations — they are not directly applicable to data drawn from Markov processes and are confined to single-layer attention. Our work belongs to this line of research and we adopt a two-attention-layer transformer architecture, which is more complicated than the transformer studied in these works.

Induction Head. Elhage et al. (2021) introduce the concept of "induction heads" as the mechanism underlying the ICL capabilities of transformers. Since then, there has been a surge of interest in understanding the induction head mechanism and its role in ICL. At a high level, the induction head mechanism works by matching the history of the current token with those seen previously in the sequence and then predicting the next token based on the matched historical sub-sequences. Olsson et al. (2022) provide empirical evidence highlighting that induction heads are crucial in facilitating the ICL capabilities of transformers. Bietti et al. (2024); Edelman et al. (2024) conduct a further empirical investigation into the development of induction heads specifically tailored for the ICL of bi-gram data models. Rajaraman et al. (2024a) provide explicit constructions of single-head transformers with constant depths that can learn n-gram data. Also, a wider range of functionalities exhibited by induction heads that interact with various other mechanisms have been observed by Wang et al. (2022).

From a theoretical perspective, Nichani et al. (2024) study the ICL of *first-order* Markov chains using a two-layer transformer and demonstrate the formation of the induction head mechanism. Makkuva et al. (2024a) also prove that training a single layer attention with a feed-forward layer on *first-order* Markov data (with $\{0, 1\}$ vocabulary) can converge to either to global or local minima depending on the initialization. However, the *first-order* assumption seems to be quite restrictive, especially when modeling the natural language, where the tokens can depend on multiple previous tokens. Most related to our work is Nichani et al. (2024), where they analyzed how training by gradient descent enables a two-layer transformer to learn the latent causal graph underlying the ICL data. However, the analysis in Nichani et al. (2024) applies to Markov chains where each token has at most one parent, and it remains unclear how to extend the analysis to more general n-gram Markov chains.

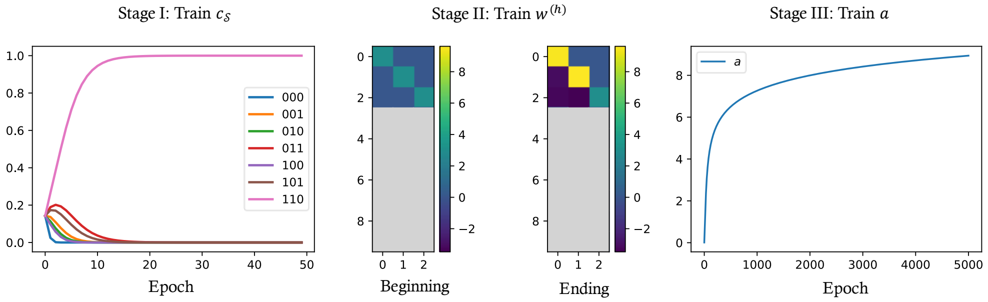
In this work, we show that a generalized version of the induction head mechanism can emerge when training a two-layer transformer on n-gram Markov chains. Moreover, our transformer models are more sophisticated, incorporating features like relative positional embedding, multi-head attention, an FNN layer, and normalization. Notably, we provide an in-depth dynamics analysis of the corresponding FFN layer and two-layer multi-head attention.

B Experiments

In this section, we first detail the setup for the experiment in Figure 3, and then provide additional results for training a model that also incorporates the word embedding matrices W_Q, W_K, W_V and the output embedding matrix W_O in the first attention layer. Let us first detail the data setup that is used for all the experiments in this work.

Data generation. The dataset for the ICL task is generated as n-gram Markov chains as described in §2.1. We take $\mathtt{pa} = \{-1, -2\}$ as the parent set. Thus, the number of parents is $n = 2$ and the token embedding dimension is $d = 3$. Note that for each sequence, the transition matrix $\pi(x \,|\, x_{\mathtt{pa}})$ is of shape $d \times d^n$. We assign a prior distribution $\mathcal{P}$ for the transition matrix, which is defined such that each column of the transition matrix of kernel π is independently drawn from a symmetric Dirichlet distribution with parameter $\alpha = 0.01$, i.e., $\pi(\cdot|x_{\mathtt{pa}}) \sim \mathrm{Dir}(\alpha \cdot \mathbf{1}_d)$. Note that each chain has different transition kernel π but follows the same prior distribution $\mathcal{P}$. We randomly sample 10,000 Markov chains with $L = 100$ from the prior distribution $\mathcal{P}$; 9,000 are used for training and 1,000 for validation.

https://doi.org/10.52202/079017-2127



(a) (b) (c)

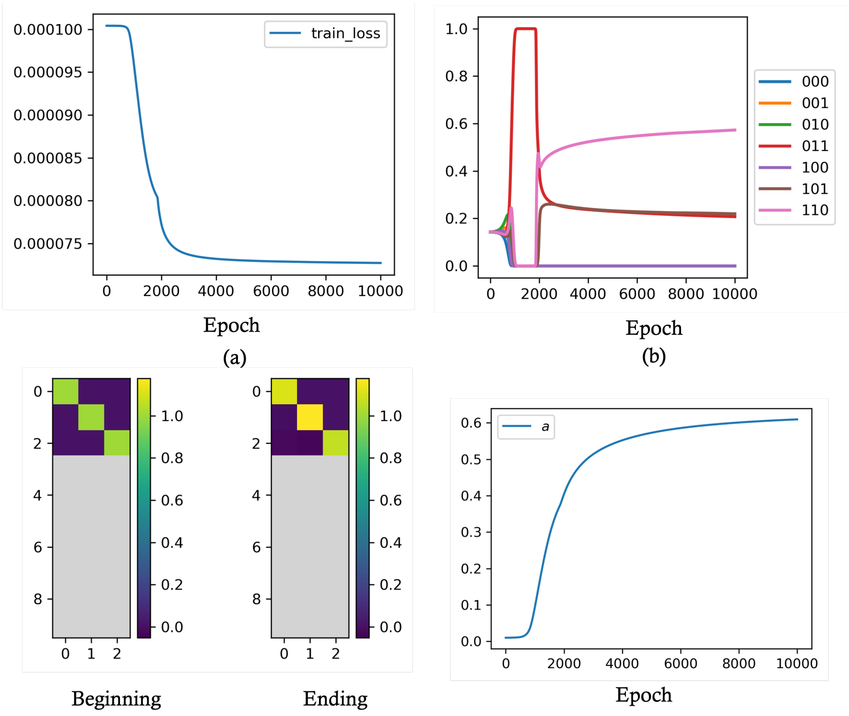
Figure 4: An illustration of the transformer parameters during the three-stage training. This is the same figure as Figure 3. We train a transformer in $\mathtt{TF}(M = 3, H = 3, d = 3, D = 2)$ with $L = 100$, $\mathtt{pa} = \{-1, -2\}$. In (a) we show the evolution of $\{p_{\mathcal{S}}\}_{\mathcal{S}\in[H]_{\le D}}$ in the first stage of training where $p_{\mathcal{S}} = c_{\mathcal{S}}^2 / \sum_{\mathcal{S}'\in[H]_{\le D}} c_{\mathcal{S}'}^2$. We use a binary coding in $\{0, 1\}^3$ to indicate each subset $\mathcal{S}$. Recall that "110" represents $= \{1, 2\}$, which is exactly $\mathcal{S}^\star$. This figure shows that $p_{\mathcal{S}^\star}$ gradually increases to one while the any other $p_{\mathcal{S}}$ decays to zero. In (b) we plot the RPE weights of the first attention layer before and after the second stage of training. Here the h-th column corresponds to the RPE weight vector of head h. This figure shows that $w_{-1}^{(1)}$ and $w_{-2}^{(2)}$ increase to a large number after training, while $w_{-3}^{(3)}$ stays close to its initial value. Thus, we have $\sigma(w^{(1)}) \approx \sigma(w^{(2)}) \approx 1$. That is, the first two heads are trained to attend to parents -1 and -2, respectively. In (c) we plot the evolution of a in the last stage of training. This figure clearly exhibits a two-step growth pattern and a keeps increasing throughout this stage. In summary, the results of the simulation experiments coincide with the theoretical results.

B.1 Training with Stage Splitting

we present the simulation results with model $\mathtt{TF}(M, H, d, D)$ in (2.5) and training in the three-stage manner. We configure the model with window size $M = 3$, number of heads $H = 3$, vocabulary size $d = 3$ and maximal FFN degree $D = 2$.

Model initialization. The RPE weight matrix $W_P^{(h)}$ is initialized such that the $(-i)$-th diagonal of $W_P^{(h)}$ has value $w_{-i}^{(h)}$ for $i = 1, 2, \ldots, M$, while all other entries are initialized to $-\infty$. See Figure 2 for an interpretation. We initialize $w_{-h}^{(h)} = 3$ and set the remaining entries within the size-M window to 0.01 to ensure symmetrization-breaking and some initial correspondence between heads and parents. For the FFN layer that learns the polynomial features, all $c_{\mathcal{S}}$ for $\mathcal{S} \in [H]_{\le D}$ are initialized to 0.01. The initial value of a in the second attention layer is set to 0.01.

Training settings. The models are trained using gradient descent with respect to the cross-entropy loss and a constant learning rate that is set to one for all stages. We train the model in Stage I (update parameters $\{c_{\mathcal{S}}\}$ only) for 2000 epochs, in Stage II (update parameters $\{w^{(h)}\}$ only) for 50,000 epochs, and in Stage III (update parameter a only) for 5000 epochs, respectively. All experiments are conducted using a single Nvidia A100 GPU. The results are shown in Figure 4, which matches our theoretical results.



(a)

(b)

(c)

(d)

Figure 5: An illustration of the evolution of gradient descent dynamics when training a transformer model specified in §B.2 with word embedding matrices $\{W_Q, W_K, W_V\}$. Here the dynamics are not split into three stages and each gradient descent step updates all parameters. We set $M = H = 3$, $d = 3$, and $D = 2$, the number of input token is $L = 100$, and Markov chain has parent set $\mathtt{pa} = \{-1, -2\}$. In (a) we show the training loss of the model, which shows that the loss decreases and converges to some value. In (b) we show the evolution of $p_{\mathcal{S}}$ where we use binary coding $\{0, 1\}^3$ to indicate each subset $\mathcal{S}$. Here, $p_{\mathcal{S}^\star}$ has code "110", which corresponds to the true parent set. This figure shows that initially a wrong $p_{\mathcal{S}}$ dominates at the early stage of training, which corresponds to $\mathcal{S} = \{2, 3\}$ (code "011"). Then eventually $p_{\mathcal{S}^\star}$ increases and becomes dominant. However, $p_{\mathcal{S}^\star}$ does not increase to one and is about 0.6, and there are two $p_{\mathcal{S}}$'s that are about 0.2. In (c) we show the RPE weights of the first attention layer before and after training. The entries corresponding to the true parents, $w_{-1}^{(1)}$ and $w_{-2}^{(2)}$, significantly increase after training, while $w_{-3}^{(3)}$ slightly increases from initialization. This figure shows that each attention head focuses on copying a single previous token. In (d) we show the evolution of the weight a in the second attention layer. We observe a similar "elbow" curve as in Figure 3-(c).

B.2 Training without Stage Splitting

Previously in §B.1, we show the simulation results on the simplified model (2.5). Now we present the results of additional experiments based on the full model defined as follows.

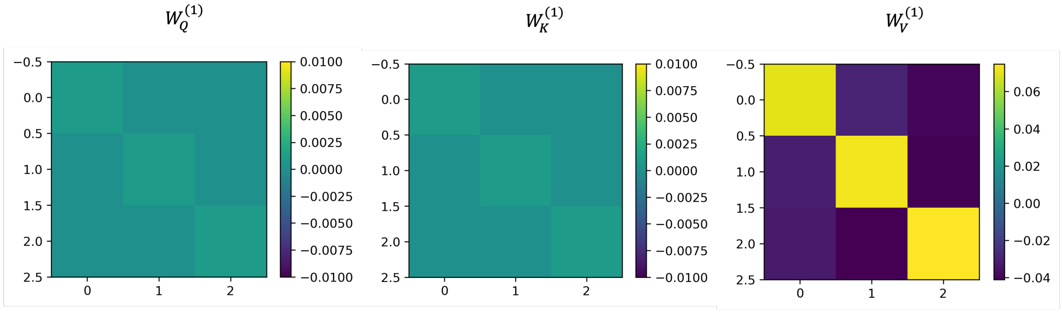
$$\begin{aligned}
&\textbf{First Attention:} && \widetilde{V}^{(h)} = \sigma\big(\widetilde{X} W_Q^{(h)} W_K^{(h)\top} \widetilde{X}^\top + W_P^{(h)}\big) \widetilde{X} W_V^{(h)\top} && \in \mathbb{R}^{(L+1)\times d};\\
&\textbf{Concatenate \& Normalize:} && V = \mathrm{LN}\big([\widetilde{V}^{(1)}, \ldots, \widetilde{V}^{(H)}, \widetilde{X}]\big) && \in \mathbb{R}^{(L+1)\times (H+1)d};\\
&\textbf{FFN \& Normalize:} && \widetilde{U} = \phi(V)/\sqrt{C_D} && \in \mathbb{R}^{(L+1)\times d_e};\\
&\textbf{Concatenate} && \widetilde{X}' = [\widetilde{U}, V] && \in \mathbb{R}^{(L+1)\times ((H+1)d + d_e)};\\
&\textbf{Second Attention:} && Y = \sigma\big(a \cdot (\tilde{x}'_{L+1})^\top (\widetilde{X}'_{1:L})^\top\big) X && \in \mathbb{R}^{(L+1)\times d}.
\end{aligned}$$

In head h of the first attention layer, $W_P^{(h)}$ is the relative positional embedding matrix, and we include $W_Q^{(h)} \in \mathbb{R}^{d\times d}$, $W_K^{(h)} \in \mathbb{R}^{d\times d}$ and $W_V^{(h)} \in \mathbb{R}^{d\times d}$ as the weight matrices for the query, key, and value projections, respectively. That is, in the full model, we the attention heads has more weight matrices than the simplified model. Another difference is that we also explicitly include the residual link that copies $\widetilde{X}$ to the output of the first attention layer. For the FFN layer, $\phi : \mathbb{R}^{(H+1)d} \to \mathbb{R}^{d_e}$ is the same feed-forward network specified in (2.3). Here, we use a standard ℓ_2-layer-normalization $\mathrm{LN}(\cdot)$, defined as

$$\mathrm{LN}([x, y]) = \left[\frac{x}{\|x\|_2}, \frac{y}{\|y\|_2}\right].$$

The second attention layer takes X as the value, which comes from the residual link (i.e., concatenation of $\widetilde{U}$ and V while $\widetilde{X}$ in V remains the same after ℓ_2-normalization). In comparison to the simplified model in (2.5), here we incorporate the query, key and value projections for the first layer as in a standard transformer architecture.

Our training setup is similar to that in §B.1. We use the same dataset and a similar training settings. All these weight matrices $W_Q^{(h)}$, $W_K^{(h)}$ and $W_V^{(h)}$ are initialized as identity matrices scaled by 0.001. We initialized the RPE vector $w^{(h)}$ as $w_{-h}^{(h)} = 1$ for $h = 1, 2, 3$, and leave the remaining entries within the length-M window to 0.01. We trained the model with all parameters together for 10,000 epochs with the same loss function and learning rate. As illustrated in Figure 5, the full model converged to a state comparable to our simplified model. We further plot the $W_Q^{(1)}, W_K^{(1)}, W_V^{(1)}$ for the first head after training in Figure 6. The results demonstrate that the model converges to a point where the query and key projections are close to zero, which leaves the RPE weights to dominate the attention mechanism. This fact justifies our simplification in (2.5) where we remove the query and key projection weights and set $W_V^{(h)}$ to be identity matrix.



(a)

(b)

(c)

Figure 6: A visualization of the word embedding matrices $W_Q^{(1)}, W_K^{(1)}, W_V^{(1)}$ of a pre-trained transformer with $M = H = 3$, $d = 3$, and $D = 2$. These are the parameters in of the first attention head in the first attention layer. Since $d = 3$, all word embedding matrices are of shape 3×3. As shown in (a) and (b), $W_Q^{(1)}$ and $W_K^{(1)}$ do not change much compared to their initialization value 0.001. Thus, they are both close to the zero matrix and play a negligible role in the first attention layer. Besides, in (c) we plot $W_V^{(1)}$, which establishes a clear diagonal structure, with the diagonal entries growing to 0.07 compared to the initialization value 0.001. Thus, $W_V^{(1)}$ is proportional to the identity matrix.

B.3 Prior and Length Generalization

We further test the model learned by the three-stage training on sequences coming from different priors and of different lengths. Note that our pre-trained transformer learns to perform GIH. As introduced in §3.1, the GIH estimator can be applied to a sequence with an arbitrary length and does not concern the prior distribution of the underlying Markov chain. Thus, it is natural to see if the pre-trained transformer can also generalize to different lengths and prior distributions.

Recall that we train the transformer model with sequence length $L = 100$ and the concentration parameter of the Dirichlet prior is $\alpha = 0.01$. Here, we test the pre-trained transformer on new sequences of different lengths and sampled from different prior distributions. That is, with a different

https://doi.org/10.52202/079017-2127

concentration parameter α, we sample a random Markov chain, and generate a sequence of length L, and evaluate of cross-entropy loss for predicting x_{L+1}. Here we choose $\alpha \in \{0.05, 0.1, 0.2\}$ and range L from 10 to 1000. When generating the data, the Markov chains share the same parent set $\mathtt{pa} = \{-1, -2\}$ with the pre-training data. The results are shown in Figure 7. The results show a decreasing trend in testing loss as the sequence length increases. For $\alpha = 0.2$, we observe first a small increase in the test loss when L just exceeds 100, but then the loss decreases as L increases further. This experiment shows that the pre-trained transformer indeed generalizes in length and is robust to the change of prior distribution.

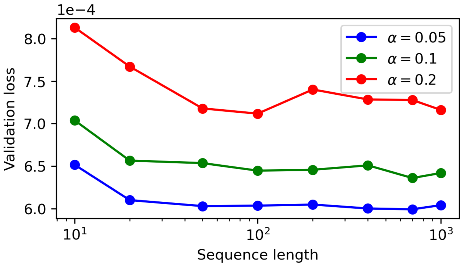


Figure 7: Generalization capability of our model to different sequence lengths and prior distributions. We plot the cross-entropy loss of the pre-trained transformer model on sequences with different lengths sampled from Markov chains with different prior distributions. The prior is Dirichlet distribution with $\alpha \in \{0.05, 0.1, 0.2\}$ and we vary the length L in $\{10, 20, 50, 100, 200, 400, 700, 1000\}$. The pre-training data contains sequences of length $L = 100$ and $\alpha = 0.01$. For different α, we see that the error has a decreasing trend as L increases. This shows that the pre-trained transformer can generalize in length and is robust to the distributional shift due to a change of prior.

C Additional Background and Discussions

C.1 How Does Transformer Implement the GIH Mechanism?

In the following, we briefly illustrate how a two-attention-layer transformer model as introduced in (2.5) implements the GIH mechanism. As we will show in §3.2, gradient flow with respect to the cross-entropy loss converges to this transformer in the limit.

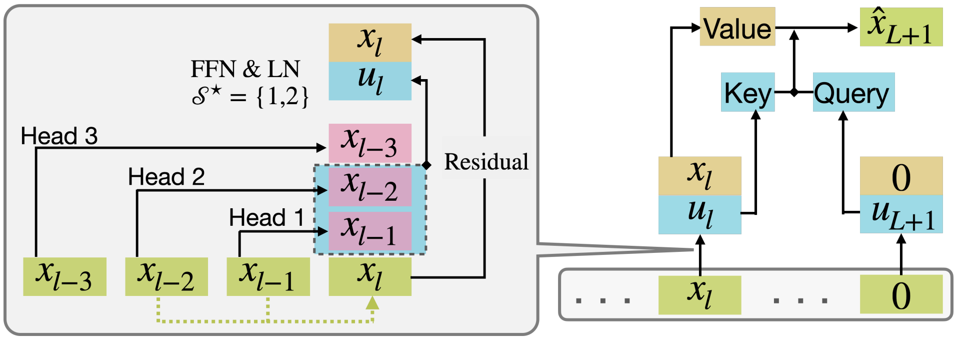


Figure 8: Illustration of the GIH mechanism in a two-attention-layer transformer model. Here, $\mathtt{pa} = \{-1, -2\}$, $M = 3$ and $\mathcal{S}^\star = \{1, 2\}$. The first attention layer copies the parents (including the information set $\mathcal{S}^\star$) to the current position. Then the FFN layer together with layer normalization generates the features u_l using the parent tokens within the information set $\mathcal{S}^\star$. The second attention layer treats each x_l as the value, and aggregates x_l as the prediction by matching the keys and query that come from the learned features using the attention mechanism. The $L + 1$-th token is padded with zeros in the input.

Step I: The First Attention Layer Copies the Information Set $\mathcal{S}^\star$ to the Current Position. Suppose the number of heads is equal to the window size for simplicity, i.e., $H = M$. Then, attention head $h \in \mathcal{S}^\star$ can attend to the h-th parent token by setting the RPE weights in the softmax function to be $w^{(h)} = \rho \cdot e_{-h}$ for a sufficiently large ρ, where $e_{-h} \in \mathbb{R}^M$ is the canonical basis vector with the $(M+1-h)$-th entry being one and all other entries being zero. As a result, each $v_l^{(h)}$ for $h \in \mathcal{S}^\star$ satisfies $v_l^{(h)} \approx x_{l-h}$.

Step II: FFN Generates the Polynomial Features of the Information Set $\mathcal{S}^\star$. As we have introduced in (2.3), each learnable $c_\mathcal{S}$ in the FFN layer determines the contribution of the corresponding subset $\mathcal{S}$ to the output feature. To let the optimal information set $\mathcal{S}^\star$ dominate the output, we set $c_{\mathcal{S}^\star} = 1$ whereas $c_\mathcal{S} = 0$ for all $\mathcal{S} \neq \mathcal{S}^\star$. The exact form of the output of the FFN layer, $\phi(v_l)$, is deferred to §C.2. Here the only property we require is that

$$s_l := \langle \phi(v_l), \phi(v_{L+1}) \rangle = \prod_{h \in \mathcal{S}^\star} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle \approx \mathbb{1}(X_{l-\mathcal{S}^\star} = X_{L+1-\mathcal{S}^\star}), \tag{C.1}$$

Here $X_{l-\mathcal{S}^\star} := (x_{l-s} : s \in \mathcal{S}^\star)$ and in the last equation we use the orthogonality and normalization of the vocabulary embeddings.

Step III: The Second Attention Layer Aggregates Tokens with Matching History on $\mathcal{S}^\star$. We can interpret s_l in (C.1) as an indicator for whether the information set of a token x_l matches the information set of the token x_{L+1}. Then for the second attention layer, by setting a to be sufficiently large, the output will become

$$y = \sum_{l=M+1}^{L} \frac{\exp\big(a \cdot s_l\big) \cdot x_l}{\sum_{k=M+1}^{L} \exp\big(a \cdot s_k\big)} \approx \begin{cases} N^{-1} \cdot \sum_{l=M+1}^{L} x_l \cdot \mathbb{1}(X_{l-\mathcal{S}^\star} = X_{L+1-\mathcal{S}^\star}), & \text{if } N \geq 1, \\ (L-M)^{-1} \cdot \sum_{l=M+1}^{L} x_l, & \text{otherwise}, \end{cases} \tag{C.2}$$

where $N = \sum_{l=M+1}^{L} \mathbb{1}(X_{l-\mathcal{S}^\star} = X_{L+1-\mathcal{S}^\star})$. That is, if at least one token x_l has a matching information set as x_{L+1}, i.e., their histories restricted to $\mathcal{S}^\star$ are the same, the second attention layer outputs the average of such tokens. Otherwise, it outputs the average of previous tokens from x_{M+1} to x_L. In Lemma E.6 in the appendix, we will show that the model learned by gradient flow implements the GIH mechanism up to a diminishing approximation error.

The weights of the transformer constructed above are illustrated in Figure 9. We consider the transformer model with $M = H = 3$, $d = 3$, and $D = 2$. In this case, in the first attention layer, for each $h \in [3]$, $W_P^{(h)}$ has three finite parameters $w_{-1}^{(h)}, w_{-2}^{(h)}$, and $w_{-3}^{(h)}$. By our construction, we have $w_{-h}^{(h)} = \rho$ for all $h \in [3]$ and the rest of the entries of $\{w^{(h)}\}_{h\in[3]}$ are all equal to zero. In Figure 9-(a) we plot the top ten by ten block of $W_P^{(1)}$, where $w_{-1}^{(1)} = \rho$ is shown in yellow and $w_{-2}^{(1)} = w_{-3}^{(1)}$ are shown in purple. The gray color stands for $-\infty$ entries. In Figure 9-(b) we plot $\{w^{(h)}\}_{h\in[3]}$. In Figure 9-(c) we plot the parameters of the FFN. Since $H = 3$ and $D = 2$, $[H]_{\leq D}$ contains seven elements: $\varnothing, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}$, and $\{2,3\}$. We use binary strings of length 3 to index these seven subsets, where the i-th bit indicates whether element i is included in the subset. For instance, "110" represents $\{1,2\}$. We set $\mathcal{S}^\star = \{1,2\}$, $c_{\mathcal{S}^\star} = 1$, and $c_\mathcal{S} = 0$ for any other $\mathcal{S}$.

C.2 Feed-Forward Network for Polynomial Kernel

Lemma C.1. *Recall the FFN satisfying (2.3), which maps a vector $z \in \mathbb{R}^{dH}$ to a vector in $\mathbb{R}^{d_e}$. We write z as $(z^{(1)}, \ldots, z^{(H)})$ where $z^{(h)} \in \mathbb{R}^d$ for all $h \in [H]$. Let $z_i^{(h)}$ be the i-th entry of $z^{(h)}$. Then we can explicitly construct $\phi(\cdot)$ by letting*

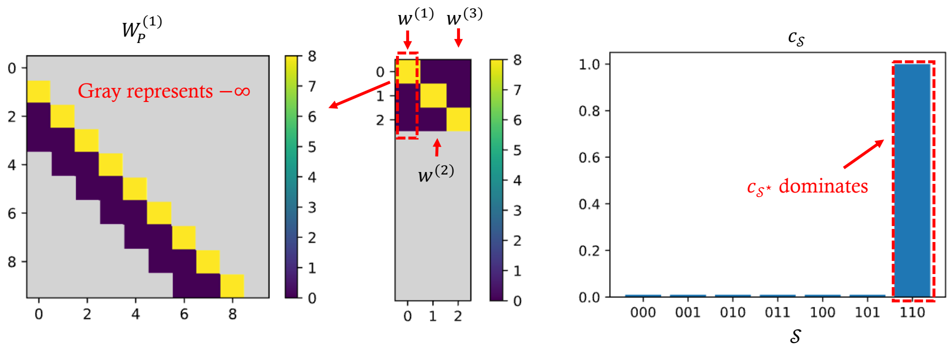
$$\phi\big((z^{(1)}, \ldots, z^{(H)})\big) = \bigg(c_\mathcal{S} \cdot \prod_{h \in \mathcal{S}} z_{i_h}^{(h)} : \{i_h\}_{h\in\mathcal{S}} \subseteq [d], \mathcal{S} \in [H]_{\leq D}\bigg), \tag{C.3}$$

which is equivalent to

$$\phi\big((z^{(1)}, \ldots, z^{(H)})\big) = \Big(c_\mathcal{S} \cdot \mathrm{vec}\big(\otimes_{h\in\mathcal{S}}(z^{(h)})\big)\Big)_{\mathcal{S}\in[H]_{\leq D}},$$

where $\mathrm{vec}(\cdot)$ is the vectorization operator that transforms a tensor into a vector by stacking all the entries in the tensor. That is, for any $\mathcal{S}$, we consider the $|\mathcal{S}|$ vectors in $\mathbb{R}^d$, $\{z^{(h)}\}_{h\in\mathcal{S}}$. In (C.3) we

https://doi.org/10.52202/079017-2127



(a) (b) (c)

Figure 9: Limiting model of $\mathtt{TF}(M = 3, H = 3, d = 3, D = 2)$ that implements the GIH mechanism with $L = 100$, $\mathtt{pa} = \{-1, -2\}$. (a): The top left 10 by 10 block of $W_P^{(1)}$ that attends to the -1 parent. (b): The RPE weight heatmap for all 3 heads, where the h-th column corresponds to the RPE weight vector of head h. (c): In the GIH mechanism, only one $c_{\mathcal{S}}^\star$ for the optimal information set $\mathcal{S}^\star$ dominates. For the label of the x-axis, we use a binary coding $\{0, 1\}^3$ to indicate each subset $\mathcal{S}$. Here, $\mathcal{S}^\star = \{1, 2\}$ is the parent set, which is represented by "110".

compute all possible products of the entries of these vectors and multiply them by $c_{\mathcal{S}}$. In particular, for each $\mathcal{S} \in [H]_{\le D}$, we enumerate $i_h \in [d]$ for all $h \in \mathcal{S}$. Therefore, the output dimension of ϕ is given by

$$d_e = \sum_{\mathcal{S} \in [H]_{\le D}} d^{|\mathcal{S}|}. \tag{C.4}$$

Proof. First, we note that the indices of $\phi(\cdot)$ have a grouped structure — we first enumerate all subsets in $[H]_{\le D}$ and then enumerate all monomials with superscripts in $\mathcal{S}$. Since there are $d^{|\mathcal{S}|}$ monomials, the output dimension is given by (C.4).

It remains to verify (2.3) with $\phi(\cdot)$ defined in (C.3). To this end, we note that for any $u, v \in \mathbb{R}^{dH}$ and any $\mathcal{S} \in [H]_{\le D}$, we have

$$\sum_{(i_h)_{h \in \mathcal{S}} \in [d]^{|\mathcal{S}|}} \left(\prod_{h \in \mathcal{S}} u_{i_h}^{(h)} \cdot v_{i_h}^{(h)} \right) = \prod_{h \in \mathcal{S}} \left(\sum_{i_h \in [d]} u_{i_h}^{(h)} \cdot v_{i_h}^{(h)} \right) = \prod_{h \in \mathcal{S}} \langle u^{(h)}, v^{(h)} \rangle,$$

which directly implies (2.3). Therefore, we conclude the proof of this lemma. □

C.3 Perron-Frobenius Theorem

Next, we review the basics for the celebrated Perron-Frobenius theorem on non-negative matrices (Meyer, 2023, Chapter 7). We consider the following class of irreducible matrices.

Definition C.2 (Irreducible Matrix)**.** *A non-negative square matrix $P \in \mathbb{R}_+^{d \times d}$ is called irreducible if the induced directed graph $\mathcal{G}$ is strongly connected, i.e., for any pair of nodes in the graph, there always exists a directed path that connects these two nodes. Here, the induced graph $\mathcal{G}$ is defined based on d nodes with adjacent matrix A given by $A_{ij} = \mathbb{1}(P_{ij} \neq 0)$.*

In particular, if P is a stochastic matrix that corresponds to a d-state Markov chain, then starting from any state, we can reach any other state with positive probability in a finite number of steps. The irreducibility property also has an equivalent definition in the matrix form. That is, for any permutation matrix T, TPT^{-1} cannot be written as an upper triangular block matrix with the following form

$$\begin{bmatrix} M_1 & M_2 \\ 0 & M_3 \end{bmatrix}.$$

In other words, an irreducible matrix does not have a nontrivial absorbing subspace that aligns with the standard basis.

In this work, we require more than the irreducibility property from the transition matrix P_π defined in §3.2. In fact, we need the existence of a unique stationary distribution (which is not guaranteed by the irreducibility) so that the chain has a sufficiently fast mixing rate. This enables us to learn with a finite sequence length L. To achieve this, one typically needs the second largest magnitude of the eigenvalues of P_π, denoted by λ, to be bounded away from 1, which is the leading eigenvalue of the transition matrix. The difference $1 - \lambda$ is also referred to as the spectral gap. It is well-known that if all the entries of P_π are positive, then P_π is irreducible and there is only one leading eigenvalue on the spectral circle with the corresponding eigenvector given by the chain's stationary distribution μ^π, and all the other eigenvalues have magnitude strictly less than 1. However, for our case, the transition matrix P_π has zero entries by definition. Fortunately, the nice property on the existence of spectral gap can be generalized to a class called *primitive* matrix.

Definition C.3 (Primitive Matrix)**.** *A nonnegative and irreducible square matrix P is called primitive if there exists an integer k such that all the entries of P^k are positive.*

By definition of the primitive matrix, one can immediately see that for any $k' > k$, $P_\pi^{k'}$ is a positive matrix. The following is the celebrated Perron-Frobenius theorem that characterizes the spectral structure of the primitive matrices.

Theorem C.4 (Perron-Frobenius Theorem for Primitive Matrices)**.** *Let P be a primitive matrix. Then the following statements hold:*

1. *The leading eigenvalue of P is real and positive, and it is the unique eigenvalue with the largest magnitude. In particular, if P is a stochastic matrix, then the leading eigenvalue is* 1.
2. *The leading eigenvector of P is positive and unique up to a scaling factor. In particular, if P is a stochastic matrix, then the leading eigenvector is the stationary distribution of the Markov chain with transition kernel P.*

The Perron-Frobenius theorem guarantees the existence of a unique stationary distribution μ^π when the transition matrix P_π is primitive. In particular, when we further assume that the transition matrix P_π has a spectral gap, the chain is sufficiently mixed, meaning that we can thus approximate sum over the entire sequence with an average with respect to the stationary distribution. In particular, the approximation error will decays with the sequence length L.

C.4 Sequential CE Loss

In this work, we only consider the prediction error on the last token in the sequence as in (2.1):

$$\mathcal{L}(f_{\texttt{tf}}) = -\mathbb{E}_{\pi\sim\mathcal{P},x_{1:(L+1)}}\big[\log\big(f_{\texttt{tf}}(x_{L+1}\,|\,x_{1:L}) + \epsilon\big)\big].$$

In practice however, people often train the transformer model by minimizing the cross-entropy (CE) loss over the entire sequence. We demonstrate that our analysis can be extended to training on the entire sequence. In this vein, we define the sequential CE loss as

$$\mathcal{L}_{\texttt{seq}}(f_{\texttt{tf}}) = \sum_{l=1}^{L} -\mathbb{E}_{\pi\sim\mathcal{P},X}\big[\log\big(f_{\texttt{tf}}(x_{l+1}\,|\,x_{1:l}) + \epsilon\big)\big]. \tag{C.5}$$

One can equivalently view this sequential CE loss as an aggregation of the CE loss for sequence length ranging from 1 to L. We argue from the following two perspectives that our analysis can be extended to the sequential cross-entropy (CE) loss:

1. Due to the use of relative positional embedding (RPE), the transformer's predictions are invariant to the *absolute* positions of tokens within a sequence. Intuitively, this implies that even if we choose a different sequence length L', the model can still handle the task in the same manner.
2. By Assumption 3.5, the chain is sufficiently mixed for large L. In the analysis, we actually use $X_{l-M:l} = (x_l, x_{l-1}, \ldots, x_{l-M}) \sim \mu^\pi$, where μ^π is the stationary distribution over a length-$(M+1)$ window, to approximate the aggregation over $X_{l-M:l}$ for $l = M+1, \ldots, L$ in the sequence. For example, this approximation is reflected in the transition from (D.3) to (D.4) in the proof sketch in §D. Since changing the sequence length does not affect the underlying stationary distribution, the only issue is the approximation error. In particular, for sufficiently large L, the CE loss at large l constitutes the majority of the sequential CE loss in (C.5), making the CE loss at small l negligible.

https://doi.org/10.52202/079017-2127

C.5 Standard χ^2-Divergence and Mutual Information

The χ^2-divergence (or χ^2-distance) between two probability distributions P and Q in the same probability space is defined as:

$$D_{\chi^2}(P\|Q) = \sum_{x\in\mathrm{supp}(Q)} \frac{(P(x)-Q(x))^2}{Q(x)},$$

where the summation is taken over all elements x in the sample space where $Q(x) > 0$. The χ^2-mutual information between two random variables X and Y with joint distribution P_{XY} and marginal distributions P_X and P_Y is defined as:

$$I_{\chi^2}(X;Y) = D_{\chi^2}(P_{XY}\|P_X \otimes P_Y) = \sum_y D_{\chi^2}(P_{X\,|\,Y}(\cdot\,|\,y)\|P_X(\cdot))P_Y(y).$$

where $P_X \otimes P_Y$ is the product of the marginals, meaning $(P_X \otimes P_Y)(x,y) = P_X(x)P_Y(y)$. For a Markov chain $X \to Y \to Z$, the χ^2-mutual information satisfies the data processing inequality

$$I_{\chi^2}(X;Z) \le I_{\chi^2}(Y;Z),$$

which follows from the observation that χ^2-divergence is also an f-divergence.

C.6 More Details on the Generalized Induction Head Mechanism

Recall that we define the Generalized Induction Head (GIH) estimator in (3.3). Specifically, $\mathtt{GIH}(x_{1:L}; M, D)$ is constructed in two steps. First, we find the information-optimal subset $\mathcal{S}^\star$ of $[M]$ by solving (3.2). Second, we build a d-class kernel classifier to predict x_{L+1}, where the "data" used by such a classifier are $\{\psi_{\mathcal{S}^\star}(l), x_l\}_{l\in[M+1,L]}$. Here $\{\psi_{\mathcal{S}^\star}(l), l \in [M+1, L+1]\}$ are features constructed at each position based on the partial history given $\mathcal{S}^\star$. In particular, similar to (C.3), for any subset $\mathcal{S}$ of $[M]$, any input token sequence $x_{1:L}$, and any position $l \in [M+1, L+1]$, we define $\psi_{\mathcal{S}}(l) = \psi_{\mathcal{S}}(l; x_{1:L})$ as

$$\psi_{\mathcal{S}}(l) = \mathrm{vec}\Big(\bigotimes_{s\in\mathcal{S}} x_{l-s}\Big) = \bigg(\prod_{s\in\mathcal{S}} (x_{l-s})_{i_s} : \{i_s\}_{s\in\mathcal{S}} \subseteq [d]\bigg) \in \mathbb{R}^{d^{|\mathcal{S}|}}.$$

In other word, $\psi_{\mathcal{S}}(l)$ is given by expanding the rank-1 tensor spanned by $\{x_{l-s}\}_{s\in\mathcal{S}}$ into a vector. Here $x_{l-s} \in \mathcal{X}$ is a vector in $\mathbb{R}^d$ and we let $(x_{l-s})_{i_s}$ denote its i_s-th entry. The rationale behind $\psi_{\mathcal{S}}(l)$ is similar to ϕ introduced in (C.3). We form a long vector containing all the products of the entries of vectors $\{x_{l-s}\}_{s\in\mathcal{S}}$. Here we omit the dependency of $\psi_{\mathcal{S}}$ on the input sequence $x_{1:L}$ to simplify the notation. Furthermore, $\psi_{\mathcal{S}}$ induces a polynomial kernel such that for any $l, m \in [M+1, L+1]$, we have

$$\langle \psi_{\mathcal{S}}(l), \psi_{\mathcal{S}}(m)\rangle = \prod_{s\in\mathcal{S}} \langle x_{l-s}, x_{m-s}\rangle = \mathbb{1}\{x_{l-s} = x_{m-s}, \forall s \in \mathcal{S}\}.$$

That is, feature $\psi_{\mathcal{S}}$ selects the token position pairs (l, m) such that the partial histories induced by $\mathcal{S}$ at position l and m are exactly the same.

Based on $\{\psi_{\mathcal{S}^\star}(l), x_l\}_{l\in[M+1,L]}$, GIH forms a kernel classifier using the indicator kernel. Specifically, for any $j \in [d]$, by (3.3), $\mathtt{GIH}(x_{1:L}; M, D)$ outputs each $e_j \in \mathcal{X}$ with probability

$$\mathbb{P}\big(\mathtt{GIH}(x_{1:L}; M, D) = e_j\big) = \frac{\sum_{l=M+1}^L \mathbb{1}\{x_{l-s} = x_{L+1-s}, \forall s \in \mathcal{S}^\star\} \cdot \mathbb{1}\{x_l = e_j\}}{\sum_{m=M+1}^L \mathbb{1}\{x_{m-s} = x_{L+1-s}, \forall s \in \mathcal{S}^\star\}}.$$

C.7 Further Discussions on the GIH Mechanism

We conclude this section with further discussions on the modified χ^2-mutual information and low-degree polynomial kernel for the FFN within the GIH mechanism.

On the Modified χ^2-Mutual Information. Now that we have shown how gradient flow approaches the desired GIH model, it is natural to ask the following questions: What is the optimal subset $\mathcal{S}^\star$ that the model selects? How well does the model perform? For the purpose of illustration, let us consider a symmetric case where the stationary distribution μ^π over a length-r_n window is uniform over $\mathcal{X}^{r_n}$. One can verify that in this case, the stationary distribution over a window of any other length is uniform as well, and the modified mutual information can be simplified into

$$\log \widetilde{I}_{\chi^2}(\mathcal{S}) = \log I_{\chi^2}(\mathcal{S}) - |\mathcal{S}| \log d, \tag{C.6}$$

where $I_{\chi^2}(\mathcal{S})$ is the standard χ^2 mutual information between $\mu^\pi(z \mid Z_{-\mathcal{S}})$ and $\mu^\pi(z)$, and the second term $|\mathcal{S}| \log d$ serves as a penalty on the *model complexity*. Thus, the GIH mechanism is *reaching a balance between the model complexity and the information richness*. Below we characterize two scenarios where the model will select the exact parent set, i.e., $\mathcal{S}^\star = \texttt{pa}$.

1. If $n = 1$, i.e., each token only has one parent, then $\mathcal{S}^\star = \texttt{pa}$. This is because $\mathcal{S}^\star$ simultaneously maximizes both terms in (C.6), thus reproducing the results in Nichani et al. (2024).
2. If n is known a priori and restricting the polynomial kernel to $\mathcal{S} \in [H]_{=n} = \{\mathcal{S} \in [H] : |\mathcal{S}| = n\}$ for the FFN layer, then $\mathcal{S}^\star = \texttt{pa}$. Here, the penalty term does not influence the selection and the exact parent set maximizes the mutual information by the data-processing inequality.

In the general case, however, the model could be much more flexible, and it is possible that the model selects only a subset of the true parent set or even some non-parent tokens that are also informative. The rationale is that with a more complex model, e.g., selecting a large $\mathcal{S}$, the model are able to make more accurate predictions for large L but may endure a large estimation error for small L, as the exact matching $X_{l-\mathcal{S}} = X_{L+1-\mathcal{S}}$ may appear rarely in the sequence.

On the Low-Degree Polynomial Kernel. The goal of using a low-degree polynomial kernel in (2.3) is to strike a balance between model complexity (which is also related to computational cost) and the model's accuracy. In this regard, we have the following corollary.

Corollary C.5. *We always have* $|\mathcal{S}^\star| \leq n$ *regardless of the choice of* D*, where* $\mathcal{S}^\star = \operatorname{argmax}_{[H]_{\leq D}} \log \widetilde{I}_{\chi^2}(\mathcal{S})$ *for* $\widetilde{I}_{\chi^2}(\mathcal{S})$ *in* (C.6)

The reasoning behind this corollary is as follows. Consider any set $\mathcal{S}$ with $|\mathcal{S}| > n$, we have $I_{\chi^2}(\mathcal{S}) \leq I_{\chi^2}(\texttt{pa})$ as the true parent set is the most informative. Moreover, since $|\texttt{pa}| = n < |\mathcal{S}|$, $\mathcal{S}$ suffers from a larger penalty. As a result, we have $\log \widetilde{I}_{\chi^2}(\mathcal{S}) < \log \widetilde{I}_{\chi^2}(\texttt{pa})$ when $\mathcal{S}$ has more than n elements. In other words, it is without loss of generality to set $D \leq n$.

C.8 Conclusion and Future Directions

In this paper, we have studied the training dynamics of a two-attention-layer transformer model for learning n-gram Markov chains in an in-context way. Our theoretical analysis underscores a congruous interplay between the multihead attention mechanism, the feed-forward network, and layer normalization that yields a generalized version of the induction head mechanism during the training. In particular, we prove that the generalized induction head mechanism adopts a modified χ^2-mutual information criterion for parent selection that strikes a balance between information richness and model complexity. To our best knowledge, our work gives the first theoretical evidence for learning an induction head mechanism with n-gram Markov data, which potentially sheds light on the inner workings of large-scale transformer models.

Our work opens new directions for developing a rigorous understanding of the transformer models. A natural direction would be that if one can find such a mechanism with standard FFN layer using multi-layer perceptron and standard layer normalization in the more practical transformer model. The intuition is that our FFN layer in (2.3), which is further instantiated in (C.3), lies in the space of low-degree polynomials and can be well represented by a MLP with sufficient dimensions and proper activation functions. Initial attempts to learn nonlinear features have also been made by Kim and Suzuki (2024). Another direction is to investigate the training dynamics beyond a single loop of this induction head mechanism, e.g., iteration head with recursively refined predictions (Cabannes et al., 2024), and how the induction head mechanism occurs in multi-layer transformer models.

D Proof Sketch

In this section, we discuss the main ingredients of analysis of gradient flow. First, we show in §D.1 how to simplify the model based on our choice of the initialization and the structure of the disentangled transformer. We then proceed to present the main proof ideas for the three stages of the gradient flow dynamics, where the training yields the following behaviors:

https://doi.org/10.52202/079017-2127

- Stage I: A unique $\mathcal{S}^\star \in [H]_{\le D}$ stands out such that the associated parameter $c_{\mathcal{S}^\star}$ dominates those of the other sets. As a result, $p^*_{\mathcal{S}}(t) = c^2_{\mathcal{S}^*}(t)/C_D(t)$ approaches to one.
- Stage II: For each $h \in \mathcal{S}^\star$, $\sigma(w^{(h)})$ approaches a one-hot vector $e_{M+1-h} \in \mathbb{R}^M$, where $w^{(h)}$ contains the parameters of RPE of the h-th head. During this stage, each head concentrates on copying a particular parent.
- Stage III: Finally, a grows and reaches $\mathcal{O}(\log L)$. As a result, the trained model approximately implements the GIH mechanism $\texttt{GIH}(x_{1:L}; M, D)$.

D.1 Simplification of the Transformer Model at Initialization

We first simplify the expression of the transformer model at initialization under Assumption 3.3, by showing that the attention scores of the second attention layer admit a simpler form.

For the second attention layer, we write the output as $y^\top = \sigma(as)X$ where $s := u_{L+1}^\top \texttt{Mask}(U_{1:L}^\top) \in \mathbb{R}^{1\times L}$ is the row vector of the similarity scores. Recall from (2.5) that the FFN layer with normalization outputs $U = \phi(V)/\sqrt{C_D} \in \mathbb{R}^{(L+1)\times d_e}$, and we denote the l-th row of U by $u_l = \phi(v_l)/\sqrt{C_D}$. For $l = M+1, \ldots, L$, the l-th entry of s is given by

$$s_l = \langle u_l, u_{L+1}\rangle = \langle \phi(v_l), \phi(v_{L+1})\rangle / C_D,$$

and the other entries are all $-\infty$. By the property of the FFN layer in (2.3) and the definition $C_D = \sum_{\mathcal{S}\in[H]_{\le D}} c^2_{\mathcal{S}}$, we can rewrite the above attention score as

$$s_l = \frac{\sum_{\mathcal{S}\in[H]_{\le D}} c^2_{\mathcal{S}} \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle}{\sum_{\mathcal{S}\in[H]_{\le D}} c^2_{\mathcal{S}}}, \quad \text{for } l = M+1, \ldots, L. \tag{D.1}$$

Note that under Assumption 3.3, by the definition of Δw in (3.6), we have a sufficiently large gap $w_{-h}^{(h)} - w_{-j}^{(h)}$ for all $j \neq h$ at initialization. Thus, $\exp(w_{-h}^{(h)}) \gg \exp(w_{-j}^{(h)})$ for all $j \neq h$, which implies the following approximation:

$$v_l^{(h)} = \sum_{k=1}^{M} \frac{\exp(w_{-k}^{(h)})}{\sum_{j=1}^{M} \exp(w_{-j}^{(h)})} \cdot x_{l-k} \approx x_{l-h}, \quad \text{for } l = M+1, \ldots, L.$$

This further implies that for $l = M+1, \ldots, L$, we have

$$\prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle \approx \prod_{h\in\mathcal{S}} \langle x_{l-h}, x_{L+1-h}\rangle = \mathbb{1}\{x_{l-i} = x_{L+1-i} \text{ for } i \in \mathcal{S}\}, \tag{D.2}$$

which is a binary value indicating whether the query and the key token's history match on the subset $\mathcal{S}$. Combining (D.1) and (D.2), we obtain the following simplified expression for s_l:

$$s_l \approx \frac{\sum_{\mathcal{S}\in[H]_{\le D}} c^2_{\mathcal{S}} \cdot \mathbb{1}\{x_{l-i} = x_{L+1-i} \text{ for } i \in \mathcal{S}\}}{\sum_{\mathcal{S}\in[H]_{\le D}} c^2_{\mathcal{S}}} = \sum_{\mathcal{S}\in[H]_{\le D}} p_{\mathcal{S}} \cdot \mathbb{1}\{x_{l-i} = x_{L+1-i} \text{ for } i \in \mathcal{S}\}$$

where we denote $p_{\mathcal{S}} = c^2_{\mathcal{S}} / \sum_{\mathcal{S}\in[H]_{\le D}} c^2_{\mathcal{S}}$ for $\mathcal{S} \in [H]_{\le D}$.

In summary, when Δw is sufficiently large, $v_l^{(h)}$ approximately copies the token x_{l-h}. As a result, the attention score s_l satisfies

$$s_l \approx \sum_{\mathcal{S}\in[H]_{\le D}} p_{\mathcal{S}} \cdot \mathbb{1}\{x_{l-i} = x_{L+1-i} \text{ for } i \in \mathcal{S}\}.$$

D.2 Analysis for Training the FFN and the First Attention Layer

The first two training stages involve the dynamics of the weights of the FFN, $\{c_{\mathcal{S}}\}_{\mathcal{S}\in[H]_{\le D}}$, and the weights of the first attention layer, $\{w^{(h)}\}_{h=1}^{H}$. The analyses of these two stages have similar structures and contain the following essential steps:

1. Derive the explicit expression of the dynamics of the weights, via direct calculations.

2. Unveil the key quantities (related to the modified χ^2-MI) that dominantly drive the dynamics, by replacing the empirical average over the context sequence with the expectation over the stationary distribution, along with other approximations.
3. Then based on the above characterization of the dynamics, we can show the convergence of the weights to the desired values.

D.2.1 Training the FFN: Identification of the Information Set $\mathcal{S}^\star$

In the first stage, we track the dynamics of $c_{\mathcal{S}}^2(t)$ for each $\mathcal{S} \in [H]_{\le D}$. For convenience, we drop the dependence on t in the sequel.

Recall the output of the model is $y = (\sigma(a \cdot s)X)^\top$ and the cross-entropy loss function is $\mathcal{L}(\Theta) = \mathbb{E}_{\pi \sim \mathcal{P}, x_{1:L}}[\ell(\Theta)]$, where $\ell(\Theta)$ can be written as $\ell(\Theta) = -\langle x_{L+1}, \log(y + \varepsilon \mathbf{1})\rangle$. We ignore the small constant ε in the following proof sketch for simplicity. We also abbreviate the vector of attention probabilities in the second attention layer as $\sigma \in \mathbb{R}^L$.

Calculation of the Dynamics of $c_{\mathcal{S}}^2$. By a direct calculation for the loss ℓ and s_l in (D.1),

$$\frac{\partial \ell}{\partial s_l} = -a \cdot \sigma_l(a \cdot s) \cdot \left(\frac{x_{L+1}}{y}\right)^\top (x_l - y)\,, \quad \frac{\partial s_l}{\partial c_{\mathcal{S}}} = \frac{2c_{\mathcal{S}} \prod_{h \in \mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle}{C_D} - \frac{2c_{\mathcal{S}} s_l}{C_D}.$$

Here the vector x_{L+1}/y is obtained by element-wise division and $\sigma_l(a \cdot s)$ is the l-th entry of $\sigma(a \cdot s)$. Then applying the chain rule, we obtain the following dynamics for $c_{\mathcal{S}}^2$ along the gradient flow:

$$\partial_t \log c_{\mathcal{S}}^2 = -\frac{2}{c_S} \sum_{l=M+1}^{L} \mathbb{E}\left[\frac{\partial \ell}{\partial s_l} \frac{\partial s_l}{\partial c_S}\right] = \frac{4a}{C_D} \sum_{l=M+1}^{L} \mathbb{E}\left[\sigma_l(a \cdot s) \cdot \prod_{h \in \mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle \cdot \left(\frac{x_{L+1}}{y}\right)^\top (x_l - y)\right]$$
$$\underbrace{- \frac{4a}{C_D} \sum_{l=M+1}^{L} \mathbb{E}\left[\sigma_l(a \cdot s) \cdot s_l \cdot \left(\frac{x_{L+1}}{y}\right)^\top (x_l - y)\right]}_{f(t)}.$$

Note that here the second term $f(t)$ is independent of $\mathcal{S}$, and it will be canceled out when we consider the difference of the derivatives, $\partial_t \log c_{\mathcal{S}}^2 - \partial_t \log c_{\mathcal{S}'}^2$, for two sets $\mathcal{S}, \mathcal{S}' \in [H]_{\le D}$. This is why we focus on the time derivative of $\log c_{\mathcal{S}}^2$.

Relate the Dynamics to the Modified χ^2-MI by Approximations. Now using the approximation in (D.2) for $\prod_{h \in \mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle$, expanding $(x_{L+1}/y)^\top (x_l - y)$ coordinate-wise, and noting that $\sigma_l(a \cdot s) \approx 1/(L-M)$ as we have small a in the second attention layer, we arrive at

$$\partial_t \log c_{\mathcal{S}}^2 \approx \frac{4a}{(L-M)C_D} \sum_{l=M+1}^{L} \mathbb{E}\left[\mathbb{1}(X_{l-\mathcal{S}} = X_{L+1-\mathcal{S}}) \cdot \left(\sum_{k=1}^{d} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)} - 1\right)\right] - f(t). \tag{D.3}$$

where $y(k)$ denotes the k-th entry of y and $X_{l-\mathcal{S}} := (x_{l-i} : i \in \mathcal{S})$ denotes the history of x_l on the set $\mathcal{S}$, similar for $X_{L+1-\mathcal{S}}$. Note that $y(k) \approx (L-M)^{-1} \sum_{l=M+1}^{L} \mathbb{1}(x_l = e_k) \approx \mu^\pi(e_k)$, which follows from the mixing assumption of the Markov chain that allows us to replace the average over $l = M+1, \ldots, L$ by the expectation over the stationary distribution. Also for the same reason, we can replace $(x_l, X_{l-\mathcal{S}}), (x_{L+1}, X_{L+1-\mathcal{S}})$ with *two independent copies* from the stationary distribution μ^π, i.e.,

$$\partial_t \log c_{\mathcal{S}}^2 \approx \frac{4a}{C_D} \mathbb{E}_{(x,X),(z,Z) \sim \mu^\pi \times \mu^\pi}\left[\mathbb{1}(Z_{-\mathcal{S}} = X_{-\mathcal{S}}) \cdot \left(\sum_{k=1}^{d} \frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(e_k)} - 1\right)\right] - f(t). \tag{D.4}$$

See the approximation from $g_{2,\mathcal{S}}$ to $g_{3,\mathcal{S}}$ in §E.2. Indeed, the first term in (D.4) becomes the modified χ^2-MI, $\widetilde{I}_{\chi^2}(\mathcal{S})$, which is defined in Definition 3.1. This gives rise to the following approximation:

$$\partial_t \log c_{\mathcal{S}}^2 \approx \frac{4a}{C_D} \widetilde{I}_{\chi^2}(\mathcal{S}) - f(t). \tag{D.5}$$

Since the value of $f(t)$ is independent of the specific choice of set $\mathcal{S}$, it is clear that the set $\mathcal{S}$ achieving the fastest growth rate is the information-optimal set $\mathcal{S}^* = \mathrm{argmax}_{\mathcal{S}\in[H]_{\le D}} \widetilde{I}_{\chi^2}(\mathcal{S})$ that maximizes the modified χ^2-MI within $[H]_{\le D}$.

Convergence of $p_{\mathcal{S}^\star}$. Note that $p_{\mathcal{S}} = c_{\mathcal{S}}^2 / \sum_{\mathcal{S}'\in[H]_{\le D}} c_{\mathcal{S}'}^2$ quantifies the contribution of the set $\mathcal{S}$ to the feature produced by the FFN layer. Thus, it is the *relative growth rate* of $c_{\mathcal{S}}^2$ that matters. Towards this end, it follows from (D.5) that, for all $\mathcal{S} \in [H]_{\le D}\backslash\{\mathcal{S}^\star\}$,

$$\partial_t \log \frac{c_{\mathcal{S}^\star}^2}{c_{\mathcal{S}}^2} \approx \frac{4a}{C_D} \cdot \Big(\widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \widetilde{I}_{\chi^2}(\mathcal{S})\Big) \ge \frac{4a}{C_D} \cdot \Delta \widetilde{I}_{\chi^2}. \tag{D.6}$$

Here we recall from (3.5) that $\Delta\widetilde{I}_{\chi^2}$ quantifies the minimal gap between the modified χ^2-MI of $\mathcal{S}^\star$ and any other set in $[H]_{\le D}$. The lower bound given by (D.6) ensures that for all $\mathcal{S} \ne \mathcal{S}^\star$, the ratio $c_{\mathcal{S}^\star}^2 / c_{\mathcal{S}}^2$ grows exponentially fast, which further implies that $p_{\mathcal{S}^\star}$ approaches one exponentially fast. This concludes the first stage of the training dynamics.

D.2.2 Training the First Attention Layer: Convergence of $\sigma(w^{(h)})$ to One-Hot Vector

As we proceed to the second stage after $p_{\mathcal{S}^\star} \approx 1$, it suffices to show how $\sigma(w^{(h)})$ converges to a one-hot vector e_{M+1-h} for $h \in \mathcal{S}^\star$ in order to show that the model converges to the GIH mechanism. Recall that we denote $X = (x_1, \ldots, x_L) \in \mathbb{R}^{L\times d}$. For notational convenience, we denote $\sigma^{(h)} := \sigma(w^{(h)})$ and let $X_{(l-M):(l-1)} \in \mathbb{R}^{M\times d}$ denote the submatrix of X with rows $l-M, \ldots, l-1$ for any l. Following our convention, we let $\sigma_{-i}^{(h)}$ denote the $(M+1-i)$-th entry of $\sigma^{(h)}$ and similarly for $w_{-i}^{(h)}$.

Calculation of the Dynamics of $w^{(h)}$. The main idea for analyzing $\{w^{(h)}\}_{h=1}^H$ is the same as that in the previous stage: It suffices to analyze the *difference between the growth rates* of different coordinates of $w^{(h)}$ for $h \in \mathcal{S}^\star$. In particular, we care about how quickly $w_{-h}^{(h)}$ grows compared to other coordinates if $w_{-h}^{(h)}$ is initialized to be larger than the remaining coordinates:

$$\begin{aligned}\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} &= \sum_{l=M+1}^{L} \mathbb{E}\bigg[\frac{\partial \ell}{\partial s_l}\bigg(\frac{\partial s_l}{\partial w_{-h}^{(h)}} - \frac{\partial s_l}{\partial w_{-i}^{(h)}}\bigg)\bigg] \\ &= a \sum_{l=M+1}^{L} \mathbb{E}\bigg[\sigma_l(as)\bigg(\sum_{k=1}^{d} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)} - 1\bigg)\bigg(\frac{\partial s_l}{\partial w_{-h}^{(h)}} - \frac{\partial s_l}{\partial w_{-i}^{(h)}}\bigg)\bigg].\end{aligned} \tag{D.7}$$

Now, we invoke the result obtained in the previous stage that $p_{\mathcal{S}^\star} \approx 1$, which gives us $s_l \approx \prod_{h\in\mathcal{S}^\star} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle$. Consequently, for any $h \in \mathcal{S}^\star$, we have

$$\frac{\partial s_l}{\partial w_{-i}^{(h)}} \approx \frac{\partial}{\partial w_{-i}^{(h)}} \prod_{h'\in\mathcal{S}^\star} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle = \bigg(\prod_{h'\in\mathcal{S}^\star\backslash\{h\}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle\bigg) \cdot b_l^\top (e_{M+1-i} - (\sigma^{(h)})^\top)\sigma_{-i}^{(h)} \tag{D.8}$$

where the equality follows from the fact that $w_{-i}^{(h)}$ only affects $(v_l^{(h)}, v_{L+1}^{(h)})$ and differentiating through the softmax function. Here we define $b_l := X_{(l-M):(l-1)} v_{L+1}^{(h)} + X_{(L+1-M):L} v_l^{(h)}$ to simplify the notation. Combining (D.7) and (D.8), we obtain

$$\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} \approx a g_h^\top \bigg(\sigma_{-i}^{(h)}\,(e_{M+1-h} - e_{M+1-i}) + (\sigma_{-h}^{(h)} - \sigma_{-i}^{(h)}) \sum_{j\ne h} \sigma_{-j}^{(h)} (e_{M+1-h} - e_{M+1-j})\bigg), \tag{D.9}$$

where we introduce the following notation

$$g_h := \sum_{l=M+1}^{L} \mathbb{E}\bigg[\sigma_l(a\cdot s)\cdot\bigg(\sum_{k=1}^{d} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)} - 1\bigg)\cdot \prod_{h'\in\mathcal{S}\backslash\{h\}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle b_l\bigg].$$

https://doi.org/10.52202/079017-2127

A detailed deviation of (D.9) can be found in (E.16). Notice that $\sigma_{-h}^{(h)} - \sigma_{-i}^{(h)}$ is positive at initialization. Now suppose $\sigma_{-h}^{(h)} - \sigma_{-i}^{(h)} > 0$ holds at current time t. Then, lower bounding $\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)}$ boils down to lower bounding $g_h^\top (e_{M+1-h} - e_{M+1-i})$ for $i \neq h$. Furthermore, if we can show that $\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)}$ is lower bounded by some positive value, the gap $\sigma_{-h}^{(h)} - \sigma_{-i}^{(h)}$ will further increase. Since $\sum_{i=1}^M \sigma_{-i}^{(h)} \equiv 1$, this will create a reinforcing loop that makes $\sigma_{-h}^{(h)}$ monotonically increase.

Relate the Dynamics to the Modified χ^2-MI by Approximations. We demonstrate next that $g_h^\top (e_{M+1-h} - e_{M+1-i})$ for $i \neq h$ admits a lower bound depending on the information gap $\Delta\widetilde{I}_{\chi^2}$. Specifically, using the same strategy for (D.3), we have by definition that

$$g_h^\top e_{M+1-i} \approx \frac{1}{L-M} \sum_{l=M+1}^{L} \mathbb{E}\left[\left(\sum_{k=1}^{d} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)} - 1\right) \cdot \mathbb{1}(x_{l-j} = x_{L+1-j}, j \in \mathcal{S}^\star \setminus \{h\}) \cdot b_l^\top e_{M+1-i}\right] \tag{D.10}$$

where for b_l we have by the same approximation $v_l^{(h)} \approx x_{l-h}$ and $v_{L+1}^{(h)} \approx x_{L+1-h}$ as in (D.2) that

$$b_l^\top e_{M+1-i} = {v_{L+1}^{(h)}}^\top x_{l-i} + {v_l^{(h)}}^\top x_{L+1-i} \approx \mathbb{1}(x_{L+1-h} = x_{l-i}) + \mathbb{1}(x_{l-h} = x_{l-i}). \tag{D.11}$$

Now we consider the case $i = h$ and $i \neq h$ separately:

(i) ($i = h$) For $g_h^\top e_{M+1-h}$, we simply set $i = h$ in (D.11), and the indicator $\mathbb{1}(x_{L+1-h} = x_{l-h})$ will exactly compensate for the exclusion of h in the indicator function of (D.10). Drawing an analogy to how we go from (D.3) to (D.5), we obtain

$$g_h^\top e_{M+1-h} \approx 2\widetilde{I}_{\chi^2}(\mathcal{S}^\star).$$

(ii) ($i \neq h$) For $g_h^\top e_{M+1-i}$ with $i \neq h$ in (D.11), we apply the same reasoning as in the previous case. Additionally, by using the Cauchy-Schwarz inequality, the following inequality holds up to a small error (see Lemma F.7 for a detailed derivation):

$$g_h^\top e_{M+1-i} \leq \widetilde{I}_{\chi^2}(\mathcal{S}^\star) + \widetilde{I}_{\chi^2}(\mathcal{S}^\star \backslash \{h\} \cup \{i\}) \leq 2\widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \Delta\widetilde{I}_{\chi^2}.$$

Plugging this back into the dynamics in (D.9), we conclude that for all $i \neq h$,

$$\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} \geq a \cdot \sigma_{-i}^{(h)} \cdot \Delta\widetilde{I}_{\chi^2}.$$

Convergence of $\sigma(w^{(h)})$. Combining the arguments in the previous two steps, we can now say that $\sigma_{-h}^{(h)}$ will monotonically increase. It remains to show that $\sigma_{-h}^{(h)}$ converges to one. Note that $\log(\sigma_{-h}^{(h)}/\sigma_{-i}^{(h)}) = w_{-h}^{(h)} - w_{-i}^{(h)}$ by the definition of the softmax function. Therefore,

$$\partial_t \log\big(\sigma_{-h}^{(h)}/\sigma_{-i}^{(h)}\big) = \partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} \geq a \cdot \sigma_{-i}^{(h)} \cdot \Delta\widetilde{I}_{\chi^2} = a \cdot \Delta\widetilde{I}_{\chi^2} \cdot \sigma_{-h}^{(h)}(0) \cdot \big(\sigma_{-i}^{(h)}/\sigma_{-h}^{(h)}\big)$$

where $\sigma_{-h}^{(h)}(0)$ is the initial value of $\sigma_{-h}^{(h)}$ at time $t = 0$. One can now rearrange the term and pick the ratio $\sigma_{-i}^{(h)}/\sigma_{-h}^{(h)}$ as the variable to track in the dynamics. A refined analysis in the convergence analysis in §E.3 shows that $\sigma^{(h)}$ converges to a one-hot vector with $\sigma_{-h}^{(h)}$ going to one. In particular, the convergence rate is determined by the information gap $\Delta\widetilde{I}_{\chi^2}$ according to the above formula.

D.3 Analysis for the Training of the Second Attention Layer

In the last stage, we turn to the training of a given that all $\sigma^{(h)}$'s for $h \in \mathcal{S}^\star$ are approximately one-hot vectors. The following approximation of the dynamics of $a(t)$ is performed in the region $a \leq O(\log L)$, where the signal term in the dynamics dominates the approximation error.

https://doi.org/10.52202/079017-2127

Calculation of the Dynamics of a. After Stages I and II, the output is approximated as $y(k) \approx y^\star(k) := \sum_{l=1}^{L} \sigma_l^\star \, \mathbb{1}(x_l = e_k)$ for each $k \in [d]$. Here the weighting coefficients $\sigma_1^\star, \ldots, \sigma_L^\star$ satisfy

$$\sigma_l^\star \propto \exp\left(a \cdot \mathbb{1}(X_{l-\mathcal{S}^\star} = X_{L+1-\mathcal{S}^\star})\right).$$

Note that for each $l \in [L]$, $\sigma_l^\star$ indicates the importance assigned to the l-th token based on the corresponding history of x_l over the information set $\mathcal{S}^\star$. In the population counterpart, when the chain has sufficiently mixed, for given $X_{L+1-\mathcal{S}^\star}$, we can roughly view each $(x_l, X_{l-\mathcal{S}^\star})$ as being sampled from a *reweighed version of the stationary distribution*:

$$\widetilde{\mu}^\pi(x_l, X_{l-\mathcal{S}^\star} \,|\, X_{L+1-\mathcal{S}^\star}) \propto \mu^\pi(x_l, X_{l-\mathcal{S}^\star}) \cdot \exp\left(a \cdot \mathbb{1}(X_{l-\mathcal{S}^\star} = X_{L+1-\mathcal{S}^\star})\right).$$

Following the same argument as those in the previous stages, replacing the sum over l with the expectation over the stationary distribution, we arrive at

$$\partial_t a \approx \mathbb{E}_{\pi\sim\mathcal{P},(x,X_{-\mathcal{S}^\star},z,Z_{-\mathcal{S}^\star})\sim q^\pi}\left[\mathbb{1}(X_{-\mathcal{S}^\star} = Z_{-\mathcal{S}^\star}) \cdot \left(\sum_{k=1}^{d} \frac{\mathbb{1}(x = z = e_k)}{\widetilde{\mu}^\pi(z = e_k \,|\, X_{-\mathcal{S}^\star})} - 1\right)\right]. \tag{D.12}$$

See detailed derivations of the above approximation in §E.4. Comparing the above expression with (D.4) in Stage I, one can see that here $(x, X_{-\mathcal{S}^\star})$ and $(z, Z_{-\mathcal{S}^\star})$ are no longer independent because now the model has learned to perform a *information-theoretic feature selection*, i.e., focusing on tokens sharing the same set of features based on the information set $\mathcal{S}^\star$, which is defined according to the modified χ^2-mutual information. In fact, the underlying joint distribution q^π is given by $q^\pi = \mu^\pi(x, X_{-\mathcal{S}^\star}) \cdot \widetilde{\mu}^\pi(z, Z_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star})$.

Divergence of a. As the dynamics of a has no closed-form expression due to the nonlinearity in the reweighed distribution $\widetilde{\mu}^\pi$, we resort to providing characterization for cases where a is either sufficiently small or large. In both cases, the lower and upper bounds of (D.12) can be derived, respectively. Using these bounds, we can argue rigorously that for small a, it undergoes super-exponential growth until it reaches a critical "elbow" value. After that, when a becomes even larger, it grows logarithmically until it reaches $\Omega(\log L)$.

E Analysis of the Training Dyanamics

Masking the Simplified Model. Recall that we apply a mask to the first M position in the simplified model. Therefore, we only allow index l to run from $M+1$ to L in the following analysis. In the following, we first specify the conditions on L that are required for the analysis of the training dynamics and then present the proof of Theorem 3.6.

E.1 Conditions on the Sequence Length

We first introduce the following condition on L:

$$L \ge \Omega\left(\frac{1}{\Delta \widetilde{I}_{\chi^2}^2 (1-\lambda)\gamma^{r_n+2}}\right), \quad L \ge (1-\lambda)^{-1}\gamma^{-D}, \quad \sqrt{L} \ge M \vee d, \tag{E.1}$$

where Ω only hides a universal constant that does not depend on the model parameters. The conditions in (E.1) will facilitate our analysis for Stage I and Stage II. For the last stage, we require

$$L \ge 2M + r_n \frac{\log \gamma^{-1}}{\lambda^{-1}}, \qquad \frac{L}{(\log L)^4} \ge \Omega\left(\frac{1}{\kappa^4 \gamma^{8+2|\mathcal{S}^\star|}} \cdot \left(\frac{\sqrt{M}+d}{(1-\lambda)^{1/2}\gamma^{|\mathcal{S}^\star|+2+r_n/4}}\right)^4\right), \tag{E.2}$$

where

$$\kappa := \mathbb{E}\left[D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}))\right] \wedge \mathbb{E}\left[D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot))\right] \wedge 1,$$

and Ω only hides universal constants that do not depend on the model parameters. Here, $\mu^\pi(x, X_{-\mathcal{S}^\star})$ denotes the stationary distribution of the Markov chain over token x and its parents $X_{-\mathcal{S}^\star}$, with $\mathcal{S}^\star$ being the information set defined in (3.2).

E.2 Analysis for Stage I

In this section, we analyze the dynamics of the parameters $\{c_{\mathcal{S}}^2\}_{\mathcal{S}\in[H]_{\le D}}$ in the first stage of training. We will show that there is a unique $\mathcal{S}_* \in [H]_{\le D}$ such that $c_{\mathcal{S}^\star}^2$ dominates all the other $c_{\mathcal{S}}^2$'s at the end of the first stage. In addition, we will characterize how fast this happens and provide a corresponding convergence rate.

Proof Strategy. At a high level, the strategy is to analyze $\partial_t \log c_{\mathcal{S}^\star}^2 - \partial_t \log c_{\mathcal{S}}^2$ for all $\mathcal{S} \neq \mathcal{S}^\star$ via the following steps:

1. **Dynamics Calculation.** First, we calculate the dynamics of $\log c_{\mathcal{S}}^2$ for each fixed $\mathcal{S}$. By selecting sufficiently small values for a and ε, and leveraging the mixing properties of the Markov chain with large L, the dynamics of $\log c_{\mathcal{S}}^2$ is approximately governed by the modified mutual information $\widetilde{I}_{\chi^2}(\mathcal{S})$.
2. **Lower Bound for The Growth Rate.** Consequently, we are able to lower bound the difference between the growth rates, $\partial_t \log c_{\mathcal{S}^\star}^2 - \partial_t \log c_{\mathcal{S}}^2$, in terms of $\Delta\widetilde{I}_{\chi^2}$, the gap between the modified mutual information of $\mathcal{S}^\star$ and the second-best set.
3. **Convergence.** Finally, we derive the convergence using the above lower bound.

Before presenting the proof, we first remind the readers of a few definitions and notations. Recall that our simplified model is given by

$$y = (\sigma(as)X)^\top = \sum_{l=M+1}^{L} \sigma_l(as)\cdot x_l, \quad \text{where} \quad s_l = \frac{\sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2 \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle}{\sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2}$$

Also recall that $C_D(t) = \sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2(t)$ and $p_{\mathcal{S}}(t) = c_{\mathcal{S}}^2(t)/C_D(t)$ for each $\mathcal{S}\in[H]_{\le D}$. The loss function can be rewritten as

$$\mathcal{L} = \mathbb{E}[\ell], \quad \text{where} \quad \ell = -\langle x_{L+1}, \log(y+\varepsilon\mathbf{1})\rangle.$$

Here the expectation $\mathbb{E}$ is taken over both the sequence $(x_1,\ldots,x_{L+1})$ and the Markov kernel $\pi \sim \mathcal{P}$. We abbreviate $\sigma \equiv \sigma(as)$ for convenience and denote by σ_l the l-th element of σ. By direct calculation, we have

$$\frac{\partial \ell}{\partial y} = -\frac{x_{L+1}}{y+\varepsilon\mathbf{1}}, \quad \frac{\partial y}{\partial \sigma} = X^\top, \quad \frac{\partial \sigma}{\partial s_l} = a\cdot\sigma_l(as)\cdot(e_l^\top - \sigma),$$

Then applying the chain rule, we have

$$\frac{\partial \ell}{\partial s_l} = \frac{\partial \ell}{\partial y}\frac{\partial y}{\partial \sigma}\frac{\partial \sigma}{\partial s_l} = -a\left(\frac{x_{L+1}}{y+\varepsilon\mathbf{1}}\right)^\top (x_l - y)\cdot\sigma_l(as). \tag{E.3}$$

In addition,

$$\frac{\partial s_l}{\partial c_{\mathcal{S}}} = \frac{2c_{\mathcal{S}}\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle}{\sum_{\mathcal{S}'\in[H]_{\le D}} c_{\mathcal{S}'}^2} - \frac{2c_{\mathcal{S}} s_l}{\sum_{\mathcal{S}'\in[H]_{\le D}} c_{\mathcal{S}'}^2} = \frac{2c_{\mathcal{S}}}{C_D}\left(\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle - s_l\right).$$

Now, we are ready to present the proof of Theorem 3.6 for the first stage of training. We remind readers that here only $\{c_{\mathcal{S}}\}_{\mathcal{S}[H]_{\le D}}$ are trained, and we omit the dependence on t for convenience.

Proof of Theorem 3.6: Stage I. As discussed in the proof strategy above, we first derive the dynamics of $\log c_{\mathcal{S}}^2$ for each fixed $\mathcal{S}\in[H]_{\le D}$. Then we compare the growth rate of $c_{\mathcal{S}^\star}^2$ with any other $c_{\mathcal{S}}^2$.

Calculation of The Dynamics of $\log c_{\mathcal{S}}^2$. We fix a $\mathcal{S}\in[H]_{\le D}$ and apply the chain rule $\partial\ell/\partial c_{\mathcal{S}} = \sum_{l=M+1}^{L} \partial\ell/\partial s_l \cdot \partial s_l/\partial c_{\mathcal{S}}$ and the gradient flow formula that $\partial_t c_{\mathcal{S}}^2 = -2c_{\mathcal{S}}\cdot\partial\mathcal{L}/\partial c_{\mathcal{S}}$. We have

$$\partial_t c_{\mathcal{S}}^2 = \frac{4ac_{\mathcal{S}}^2}{C_D}\sum_{l=M+1}^{L}\mathbb{E}\left[\sigma_l(as)\cdot\left(\frac{x_{L+1}}{y+\varepsilon\mathbf{1}}\right)^\top (x_l - y)\cdot\left(\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle - s_l\right)\right]. \tag{E.4}$$

In the following, we consider a fixed π for error analysis and take expectation over π again when plugging in everything back into the dynamics. To simplify the expression of $\partial_t c_{\mathcal{S}}^2$, we define quantities $g_{0,\mathcal{S}}$ and f as

$$g_{0,\mathcal{S}} := \sum_{l=M+1}^{L} \mathbb{E}_{X|\pi}\Big[\sigma_l(as)\sum_{k=1}^{d}\Big(\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y(k)+\varepsilon}-\frac{y(k)\,\mathbb{1}(x_{L+1}=e_k)}{y(k)+\varepsilon}\Big)\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Big],$$

$$f := \sum_{l=M+1}^{L} \mathbb{E}_{X|\pi}\Big[\sigma_l(as)\sum_{k=1}^{d}\Big(\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y(k)+\varepsilon}-\frac{y(k)\,\mathbb{1}(x_{L+1}=e_k)}{y(k)+\varepsilon}\Big)\cdot s_l\Big].$$

Note that here f does not depend on $\mathcal{S}$. Based on the above definitions, we can rewrite (E.4) as

$$\partial_t \log c_{\mathcal{S}}^2 = \frac{1}{c_{\mathcal{S}}^2}\cdot\partial_t c_{\mathcal{S}}^2 = \frac{4a}{C_D}\cdot\mathbb{E}_{\pi\sim\mathcal{P}}[g_{0,\mathcal{S}}-f]. \tag{E.5}$$

Using this, it can be shown that $C_D(t)$ does not change during the training, as described in the following lemma.

Lemma E.1. *The quantity $C_D(t)=\sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2(t)$ is preserved along the gradient flow over $\{c_{\mathcal{S}}\}_{\mathcal{S}\in[H]_{\le D}}$, i.e., $\partial_t C_D(t)\equiv 0$.*

This lemma will be useful in the following analysis, and we defer its proof to §E.2.1. Next, we proceed to further simplify the dynamics in (E.5) by approximating $g_{0,\mathcal{S}}$.

Simplification of $\partial_t \log c_{\mathcal{S}}^2$. To approximiate $g_{0,\mathcal{S}}$, we introduce the following quantities:

$$g_{1,\mathcal{S}} := \frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}_{X|\pi}\Big[\Big(\sum_{k=1}^{d}\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\bar{y}(k)+\varepsilon}-\frac{\bar{y}(k)\,\mathbb{1}(x_{L+1}=e_k)}{\bar{y}(k)+\varepsilon}\Big)\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Big],$$

$$g_{2,\mathcal{S}} := \frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}_{X|\pi}\Big[\Big(\sum_{k=1}^{d}\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\mu^\pi(e_k)}-1\Big)\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Big],$$

$$g_{3,\mathcal{S}} := \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\Big[\Big(\sum_{k=1}^{d}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(e_k)}-1\Big)\prod_{h\in\mathcal{S}}\langle v^{(h)}(Z), v^{(h)}(X)\rangle\Big],$$

where $Z=(z_{-M},\dots,z_{-1})$ is independent of $X=(x_{-M},\dots,x_{-1})$ and we define

$$v^{(h)}(X) := \sum_{i=1}^{M}\sigma_{-i_h}^{(h)}x_{-i_h},\quad v^{(h)}(Z) := \sum_{i=1}^{M}\sigma_{-i_h}^{(h)}z_{-i_h},\quad \text{and } \bar{y} := \frac{1}{L-M}\sum_{l=M+1}^{L}x_l.$$

Here $\bar{y}(k)$ is the k-th entry of $\bar{y}$. We remark that each of $g_{1,\mathcal{S}}, g_{2,\mathcal{S}}, g_{3,\mathcal{S}}$ is a function of π and t, but we omit the dependence for brevity.

From $g_{0,\mathcal{S}}$ to $g_{1,\mathcal{S}}$, we replace attention probability $\sigma_l(as)$ by the uniform average with factor $1/L$, which yields $\bar{y}$. From $g_{1,\mathcal{S}}$ to $g_{2,\mathcal{S}}$, we replace the empirical distribution $\bar{y}$ with the stationary distribution μ^π and drop the small constant ε. Finally, from $g_{2,\mathcal{S}}$ to $g_{3,\mathcal{S}}$, we replace the average over the sequence by the expectation over the stationary distribution μ^π of the underlying Markov chain. We will show that the approximation error in each step is small, given that a and ε are sufficiently small and the Markov chain mixes well for a large L.

- For the approximation of $g_{0,\mathcal{S}}$ by $g_{1,\mathcal{S}}$, note that when a is small, the attention probability $\sigma_l(as)\approx 1/(L-M)$ for all $l\in[L]$. More specifically, it follows from Lemma F.3 that

$$|g_{0,\mathcal{S}}-g_{1,\mathcal{S}}| \le \frac{8ad}{\varepsilon^2}.$$

- For the approximation of $g_{1,\mathcal{S}}$ by $g_{2,\mathcal{S}}$, we leverage the approximation $\bar{y}(k)\approx\mu^\pi(e_k)$ due to the mixing of the Markov chain for large L. The result in Lemma F.4 implies that

$$|g_{1,\mathcal{S}}-g_{2,\mathcal{S}}| \le 4\cdot\frac{(1-\lambda)^{-1/2}(D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1)^{1/4}+2\sqrt{M}}{L^{1/2}\gamma}+\gamma^{-1}\varepsilon$$

https://doi.org/10.52202/079017-2127

where $\mu_0(\cdot)$ is the initial distribution over the first r_n tokens. Here we abuse the notation of μ^π in $D_{\chi^2}(\mu_0 \,\|\, \mu^\pi)$ to denote the stationary distribution over the last r_n tokens. Since $\mu^\pi_{\min} \geq \gamma$ by Assumption 3.5, we have

$$D_{\chi^2}(\mu_0 \| \mu^\pi) = \sum_X (\mu(X) - \mu^\pi(X))^2/\mu^\pi(X) \leq \sum_X 1/\mu^\pi(X) \leq (2/\gamma)^{r_n}. \tag{E.6}$$

Therefore, we can further simplify the above bound as

$$|g_{1,\mathcal{S}} - g_{2,\mathcal{S}}| = O\biggl(\frac{1}{\sqrt{L(1-\lambda)\gamma^{r_n+2}}} + \frac{\varepsilon}{\gamma}\biggr).$$

- Finally, the approximation of $g_{2,\mathcal{S}}$ by $g_{3,\mathcal{S}}$ follows from the mixing property of the Markov chain. In particular, it follows from Lemma F.5 that
$$|g_{2,\mathcal{S}} - g_{3,\mathcal{S}}| \leq \frac{8M}{L\gamma} + \frac{16\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda)\gamma^{|\mathcal{S}|/2+1}} \leq O\biggl(\frac{1}{L(1-\lambda)\gamma^{|\mathcal{S}|/2+r_n/2+1}}\biggr).$$

Combining the above results, and by the assumption that $a = a(0) = O(1/L^{3/2})$ and $\varepsilon = 1/\sqrt{L}$, we obtain the following approximation error:

$$\begin{aligned}|g_{0,\mathcal{S}} - g_{3,\mathcal{S}}| &= O\left(\frac{ad}{\varepsilon^2}\right) + O\biggl(\frac{1}{\sqrt{L(1-\lambda)\gamma^{r_n+2}}} + \frac{\varepsilon}{\gamma}\biggr) + O\biggl(\frac{1}{L(1-\lambda)\gamma^{|\mathcal{S}|+2+r_n/2}}\biggr)\\ &\leq O\biggl(\frac{1}{\sqrt{L(1-\lambda)\gamma^{r_n+2}}} + \frac{1}{L(1-\lambda)\gamma^{D/2+r_n/2+1}}\biggr) \leq O\biggl(\frac{1}{\sqrt{L(1-\lambda)\gamma^{r_n+2}}}\biggr),\end{aligned}$$

where we note that $|\mathcal{S}| \leq D$ for any $\mathcal{S} \in [H]_{\leq D}$ and the last inequality holds by also noting our condition on L in (E.1) that $L \geq \Omega((1-\lambda)^{-1}\gamma^{-D})$. As a result, the dynamics of $c^2_{\mathcal{S}}$ in (E.5) can be approximated as follows:

$$\partial_t \log c^2_{\mathcal{S}} = \frac{4a}{C_D} \cdot \mathbb{E}_{\pi\sim\mathcal{P}}[g_{3,\mathcal{S}} - f] + \mathcal{E}, \quad \text{where } |\mathcal{E}| \leq O\biggl(\frac{a}{C_D\sqrt{L(1-\lambda)\gamma^{r_n+2}}}\biggr), \tag{E.7}$$

where $\mathcal{O}(\cdot)$ hides universal constants that do not depend on the model parameters. Here and in the sequel, we let $\mathcal{E}$ denote an error term that is of the order $O(a/\sqrt{C_D^2 L(1-\lambda)\gamma^{r_n+2}})$ where the specific constant hidden in $O(\cdot)$ may change from line to line, but does not depend on the model parameters. In fact, we can show C_D remains constant by Lemma E.1 and a is not updated during this stage. Thus, the error term $|\mathcal{E}|$ is of scale $O(aL^{-1/2})$.

Lower Bound for The Difference $\partial_t \log c^2_{\mathcal{S}^\star} - \partial_t \log c^2_{\mathcal{S}}$. The reason for approximating $g_{0,\mathcal{S}}$ by $g_{3,\mathcal{S}}$ in the previous step is that the latter is more interpretable, in the sense that we can relate it to the modified χ^2 mutual information $\widetilde{I}_{\chi^2}(\mathcal{S})$. Recall that for each $\mathcal{S} \in [H]_{\leq D}$, the modified χ^2-mutual information is

$$\widetilde{I}_{\chi^2}(\mathcal{S}) = \mathbb{E}_{\pi\sim\mathcal{P},(z,Z)\sim\mu^\pi}\biggl[\biggl(\sum_{e\in\mathcal{X}} \frac{\mu^\pi(z=e \,|\, Z_{-\mathcal{S}})^2}{\mu^\pi(z=e)} - 1\biggr) \cdot \mu^\pi(Z_{-\mathcal{S}})\biggr].$$

Note that f in (E.7) is independent of $\mathcal{S}$, and will be canceled when computing $\partial_t \log c^2_{\mathcal{S}^\star} - \partial_t \log c^2_{\mathcal{S}}$:

$$\partial_t \log c^2_{\mathcal{S}^\star} - \partial_t \log c^2_{\mathcal{S}} = \frac{4a}{C_D} \cdot \mathbb{E}_{\pi\sim\mathcal{P}}[g_{3,\mathcal{S}^\star} - g_{3,\mathcal{S}}] \pm 2|\mathcal{E}|.$$

Thus, it suffices to consider $\mathbb{E}_{\pi\sim\mathcal{P}}[g_{3,\mathcal{S}^\star} - g_{3,\mathcal{S}}]$. It follows from Lemma F.6 that for each $\mathcal{S} \in [H]_{\leq D}$, $\mathbb{E}_{\pi\sim\mathcal{P}}[g_{3,\mathcal{S}}]$ satisfies

$$\biggl|\mathbb{E}_{\pi\sim\mathcal{P}}[g_{3,\mathcal{S}}] - \prod_{h\in\mathcal{S}} (\sigma^{(h)}_{-h})^2 \cdot \widetilde{I}_{\chi^2}(\mathcal{S})\biggr| \leq \biggl(1 - \prod_{h\in\mathcal{S}} (\sigma^{(h)}_{-h})^2\biggr) \cdot \widetilde{I}_{\chi^2}(\mathcal{S}^\star).$$

This yields a lower bound for $\mathbb{E}_{\pi\sim\mathcal{P}}[g_{3,\mathcal{S}^\star}]$ and an upper bound for $\mathbb{E}_{\pi\sim\mathcal{P}}[g_{3,\mathcal{S}}]$ for each $\mathcal{S} \neq \mathcal{S}^\star$, i.e.,

$$\begin{aligned}\mathbb{E}_{\pi\sim\mathcal{P}}[g_{3,\mathcal{S}^\star}] &\geq \prod_{h\in\mathcal{S}^\star} (\sigma^{(h)}_{-h})^2 \cdot \widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \biggl(1 - \prod_{h\in\mathcal{S}^\star} (\sigma^{(h)}_{-h})^2\biggr) \cdot \widetilde{I}_{\chi^2}(\mathcal{S}^\star),\\ \mathbb{E}_{\pi\sim\mathcal{P}}[g_{3,\mathcal{S}}] &\leq \prod_{h\in\mathcal{S}} (\sigma^{(h)}_{-h})^2 \cdot \widetilde{I}_{\chi^2}(\mathcal{S}) + \biggl(1 - \prod_{h\in\mathcal{S}} (\sigma^{(h)}_{-h})^2\biggr) \cdot \widetilde{I}_{\chi^2}(\mathcal{S}^\star), \quad \text{for all } \mathcal{S} \neq \mathcal{S}^\star.\end{aligned}$$

Consequently,

$$\begin{aligned}
\partial_t \log c^2_{\mathcal{S}^\star} - \partial_t \log c^2_{\mathcal{S}} &= \frac{4a}{C_D}\cdot \mathbb{E}_{\pi\sim\mathcal{P}}[g_{3,\mathcal{S}^*} - g_{3,\mathcal{S}}] \pm 2|\mathcal{E}| \\
&\ge \frac{4a}{C_D}\bigg(\prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)})^2\cdot \widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \prod_{h\in\mathcal{S}}(\sigma_{-h}^{(h)})^2\cdot\widetilde{I}_{\chi^2}(\mathcal{S})\bigg)\\
&\qquad - \frac{4a}{C_D}\bigg(2 - \prod_{h\in\mathcal{S}^\star}(\sigma^{(h)}_{-h})^2 - \prod_{h\in\mathcal{S}}(\sigma^{(h)}_{-h})^2\bigg)\widetilde{I}_{\chi^2}(\mathcal{S}^\star) - 2|\mathcal{E}|\\
&\ge \frac{4a}{C_D}\bigg(\bigg(2\prod_{h\in\mathcal{S}^\star}(\sigma^{(h)}_{-h})^2 - 2\bigg)\widetilde{I}_{\chi^2}(\mathcal{S}^\star) + \prod_{h\in\mathcal{S}}(\sigma^{(h)}_{-h})^2\cdot\Delta\widetilde{I}_{\chi^2}\bigg) - 2|\mathcal{E}|,
\end{aligned}$$

where the second inequality follows from the definition $\Delta\widetilde{I}_{\chi^2} = \min_{\mathcal{S}\in[H]_{\le D}\setminus\{\mathcal{S}^\star\}} \widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \widetilde{I}_{\chi^2}(\mathcal{S})$. Moreover, since each $(\sigma^{(h)}_{-h})^2 \in (0,1)$, we have $\prod_{h\in\mathcal{S}}(\sigma^{(h)}_{-h})^2 \ge \prod_{h=1}^{H}(\sigma^{(h)}_{-h})^2$ for any $\mathcal{S}\in[H]_{\le D}$. Appling this to the above inequality, we obtain

$$\partial_t \log c^2_{\mathcal{S}^\star} - \partial_t \log c^2_{\mathcal{S}} \ge \frac{4a}{C_D}\bigg(2\prod_{h=1}^{H}(\sigma^{(h)}_{-h})^2\cdot\widetilde{I}_{\chi^2}(\mathcal{S}^\star) + \prod_{h=1}^{H}(\sigma^{(h)}_{-h})^2\cdot\Delta\widetilde{I}_{\chi^2} - 2\widetilde{I}_{\chi^2}(\mathcal{S}^\star)\bigg) - 2|\mathcal{E}|, \tag{E.8}$$

Exponential Growth of $c^2_{\mathcal{S}^\star}$. We proceed to show that the first term in (E.8) dominates the error term $\mathcal{E}$ and thus leads to the exponential growth of $c^2_{\mathcal{S}^\star}$.

Note that by Assumption 3.3, $w^{(h)}_{-h} \ge w^{(h)}_{-j} + \Delta w$ for all $j \ne h$ and $h\in[H]$, where the quantity Δw satisfies

$$\Delta w \ge \log(M-1) - \log\bigg(\bigg(1 + \frac{\Delta\widetilde{I}_{\chi^2}}{14\widetilde{I}_{\chi^2}(\mathcal{S}^\star)}\bigg)^{\frac{1}{2H}} - 1\bigg). \tag{E.9}$$

Recall that we are not updating the RPE parameters during this stage, so $\sigma^{(h)}$ is fixed for all $h\in[H]$. ***So the gap condition* (E.9) *holds throughout Stage I.*** This conditions ensures that $w^{(h)}_{-h} \gg w^{(h)}_{-j}$, so $\prod_{h\in[H]}(\sigma^{(h)}_{-h})^2$ is sufficiently large. More precisely, given that head h is more focused on the $(-h)$-th position by having a gap Δw in the initialization, we can further show by definition of the softmax function that

$$\sigma^{(h)}_{-h} \ge \frac{1}{1+(M-1)\exp(-\Delta w)}, \forall h\in[H] \Rightarrow \prod_{h=1}^{H}(\sigma^{(h)}_{-h})^2 \ge \frac{1}{\big(1+(M-1)\exp(-\Delta w)\big)^{2H}}. \tag{E.10}$$

Plugging (E.9) into (E.10), we have by additionally noting that $\widetilde{I}_{\chi^2}(\mathcal{S}^\star) \ge \Delta\widetilde{I}_{\chi^2} > 0$ that

$$\prod_{h=1}^{H}(\sigma^{(h)}_{-h})^2 \ge \bigg(1 + \frac{\Delta\widetilde{I}_{\chi^2}}{14\widetilde{I}_{\chi^2}(\mathcal{S}^\star)}\bigg)^{-1} > \frac{2\widetilde{I}_{\chi^2}(\mathcal{S}^\star) + 2/3\cdot\Delta\widetilde{I}_{\chi^2}}{2\widetilde{I}_{\chi^2}(\mathcal{S}^\star) + \Delta\widetilde{I}_{\chi^2}},$$

which implies that

$$2\prod_{h=1}^{H}(\sigma^{(h)}_{-h})^2\cdot\widetilde{I}_{\chi^2}(\mathcal{S}^\star) + \prod_{h=1}^{H}(\sigma^{(h)}_{-h})^2\cdot\Delta\widetilde{I}_{\chi^2} - 2\widetilde{I}_{\chi^2}(\mathcal{S}^\star) \ge \frac{2}{3}\Delta\widetilde{I}_{\chi^2}. \tag{E.11}$$

Moreover, when L is sufficiently large such that $L \ge \Omega((\Delta\widetilde{I}^2_{\chi^2}(1-\lambda)\gamma^{r_n+2})^{-1})$, $\mathcal{E}$ in (E.8) satisfy $|\mathcal{E}| \le 13a\Delta\widetilde{I}_{\chi^2}/6C_D$, where Ω hides a universal constant that does not depend on the model parameters. Therefore, combining (E.8) and (E.11), we conclude that

$$\partial_t \log c^2_{\mathcal{S}^\star} - \partial_t \log c^2_{\mathcal{S}} \ge \frac{8a\Delta\widetilde{I}_{\chi^2}}{3C_D} - 2|\mathcal{E}| \ge \frac{a\Delta\widetilde{I}_{\chi^2}}{2C_D}. \tag{E.12}$$

This implies that $c^2_{\mathcal{S}^\star}$ grows exponentially fast and becomes dominant.

Convergence of $p_{\mathcal{S}^\star}$. In this part, we treat all the model parameters as a function of time t. For simplicity, we omit the dependence on t when it is clear from the context. It remains to derive the convergence of $p_{\mathcal{S}^\star} = c_{\mathcal{S}^\star}^2/C_D$. Expanding $C_D = \sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2$, we can directly calculate the derivative of $p_{\mathcal{S}^\star}$ as follows:

$$
\begin{aligned}
\partial_t \log(1-p_{\mathcal{S}^\star}) &= \partial_t \log\left(1-\frac{c_{\mathcal{S}^\star}^2}{\sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2}\right) = \frac{C_D}{C_D - c_{\mathcal{S}^\star}^2}\cdot \partial_t\left(1-\frac{c_{\mathcal{S}^\star}^2}{\sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2}\right)\\
&= \frac{C_D}{C_D - c_{\mathcal{S}^\star}^2}\cdot \frac{-(\sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2)\cdot \partial_t c_{\mathcal{S}^\star}^2 + c_{\mathcal{S}^\star}^2\cdot \sum_{\mathcal{S}\in[H]_{\le D}}\partial_t c_{\mathcal{S}}^2}{(\sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2)^2}\\
&= \frac{1}{C_D(C_D - c_{\mathcal{S}^\star}^2)}\sum_{\mathcal{S}\in[H]_{\le D}} (-c_{\mathcal{S}}^2\cdot\partial_t c_{\mathcal{S}^\star}^2 + c_{\mathcal{S}^\star}^2\cdot \partial_t c_{\mathcal{S}}^2)\\
&= \frac{1}{C_D(C_D - c_{\mathcal{S}^\star}^2)}\sum_{\mathcal{S}\in[H]_{\le D}\setminus\{\mathcal{S}^\star\}} c_{\mathcal{S}^\star}^2\cdot c_{\mathcal{S}}^2\cdot(-\partial_t \log c_{\mathcal{S}^\star}^2 + \partial_t \log c_{\mathcal{S}}^2)
\end{aligned}
$$

where in the last equality we use the fact that $\partial_t \log c_{\mathcal{S}}^2 = (\partial_t c_{\mathcal{S}}^2)/c_{\mathcal{S}}^2$. Applying (E.12) to each $\mathcal{S}\ne\mathcal{S}^\star$, we further have

$$
\begin{aligned}
\partial_t \log(1-p_{\mathcal{S}^\star}) &\le \frac{1}{C_D(C_D - c_{\mathcal{S}^\star}^2)}\sum_{\mathcal{S}\in[H]_{\le D}\setminus\{\mathcal{S}^\star\}} c_{\mathcal{S}^\star}^2\cdot c_{\mathcal{S}}^2\cdot\left(-\frac{a\Delta\widetilde{I}_{\chi^2}}{2C_D}\right)\\
&= \frac{1}{C_D(C_D - c_{\mathcal{S}^\star}^2)}\cdot c_{\mathcal{S}^\star}^2\cdot(C_D - c_{\mathcal{S}^\star}^2)\cdot\left(-\frac{a\Delta\widetilde{I}_{\chi^2}}{2C_D}\right) = -\frac{c_{\mathcal{S}^\star}^2\cdot a\Delta\widetilde{I}_{\chi^2}}{2C_D^2} < 0.
\end{aligned}
$$

This implies that $p_{\mathcal{S}^\star} = c_{\mathcal{S}^\star}^2/C_D$ monotonically increases, and thus $c_{\mathcal{S}^\star}^2(t)\ge c_{\mathcal{S}^\star}^2(0)$ for any $t\ge 0$ because C_D is constant by Lemma E.1 and $c_{\mathcal{S}^\star}^2(0)$ is the initial value for $c_{\mathcal{S}^\star}^2$ at time $t=0$. Therefore, we can further replace $c_{\mathcal{S}^\star}^2$ by its initial value in the above inequality, which yields

$$
\partial_t \log(1-p_{\mathcal{S}^\star}) \le -\frac{c_{\mathcal{S}^\star}^2(0) a\Delta\widetilde{I}_{\chi^2}}{2C_D^2} = -\frac{p_{\mathcal{S}^\star}(0) a\Delta\widetilde{I}_{\chi^2}}{2C_D}
$$

We remark that the above upper bound is independent of t. Finally, applying the Grönwall's inequality to $\log(1-p_{\mathcal{S}^\star})$, we obtain

$$
1-p_{\mathcal{S}^\star}(t) \le (1-p_{\mathcal{S}^\star}(0))\cdot\exp\left(-\frac{p_{\mathcal{S}^\star}(0)a\Delta\widetilde{I}_{\chi^2}}{2C_D}\cdot t\right).
$$

With training time $t_1 \ge (2C_D(0)\log L)/(a\cdot p_{\mathcal{S}^\star}(0)\Delta\widetilde{I}_{\chi^2})$, we can guarantee that

$$
1-p_{\mathcal{S}^\star}(t_1)\le L^{-1}.
$$

This concludes the proof for the first stage of the training. □

E.2.1 Additional Proofs for the Stage I

We conclude this subsection with the proof of Lemma E.1.

Proof of Lemma E.1. By (E.5), we have

$$
\partial_t c_{\mathcal{S}}^2 = \mathbb{E}_{\pi\sim\mathcal{P}}[4a\cdot p_{\mathcal{S}}(g_{0,\mathcal{S}} - f)].
$$

Moreover, by the definition of $g_{0,\mathcal{S}}$ and f, it holds that $\sum_{\mathcal{S}\in[H]_{\le D}} p_{\mathcal{S}} g_{0,\mathcal{S}} = f$. Then,

$$
\partial_t C_D = \sum_{\mathcal{S}\in[H]_{\le D}} \partial_t c_{\mathcal{S}}^2 = 4a\cdot\mathbb{E}_{\pi\sim\mathcal{P}}\Bigg[\sum_{\mathcal{S}\in[H]_{\le D}} p_{\mathcal{S}} g_{0,\mathcal{S}} - f\Bigg] \equiv 0.
$$

Thus, the quantity C_D is preserved under the dynamics. □

https://doi.org/10.52202/079017-2127

E.3 Analysis for Stage II

In this section, we provide the analysis of the dynamics of $\sigma^{(h)} \equiv \sigma(w^{(h)})$ for head $h \in \mathcal{S}^\star$. For head $h \notin \mathcal{S}^\star$, the results from Stage I imply that $p_\mathcal{S} \to 0$ for any $\mathcal{S} \neq \mathcal{S}^\star$. Consequently, any head $h \notin \mathcal{S}^\star$ will be ignored when producing the output features of FFN. Conversely, for $h \in \mathcal{S}^\star$, we establish the dominance of $w_{-h}^{(h)}$ over $w_{-i}^{(h)}$ for all $i \neq h$, yielding $\sigma_{-h}^{(h)} \to 1$ as $t \to \infty$. In this limiting case, head h exactly copies the $(-h)$-th parent. We also provide the corresponding convergence rate.

Proof Strategy. Similar to the proof for Stage I, our analysis for Stage II characterizes the dynamics of the difference between the positional embedding weights, $\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)}$ for all $i \neq h$, via the following steps:

1. **Dynamics Calculation.** We initiate the analysis by deriving the dynamics of $w_{-i}^{(h)}$ for any fixed i and h.
2. **Dynamics Approximation** Then we approximate the dynamics by identifying the dominant term controlled by the modified χ^2 mutual information $\widetilde{I}_{\chi^2}(\mathcal{S}^\star)$.
3. **Lower Bound for The Growth Rate** By comparing the corresponding modified χ^2 mutual information, we establish a lower bound on $\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)}$ for all $i \neq h$.
4. **Convergence.** Finally, we derive the convergence rate of $\sigma_{-h}^{(h)}$ using the above lower bound.

Again, before proceeding with the detailed proof, we review the notations related to the dynamics of the positional embedding weights $\{w^{(h)}\}_{h=1}^H$. For the h-th head of the first attention layer, the positional embedding vector $w^{(h)}$ induces the attention probability over a window of size M, i.e.,

$$\sigma(w^{(h)}) =: \sigma^{(h)} = (\sigma_{-M}^{(h)}, \ldots, \sigma_{-1}^{(h)}) \in \mathbb{R}^{1\times M}.$$

Further recall the attention scores for the second attention layer, as, where $s = u_{L+1}^\top U_{M+1:L}^\top$. Then for each $l \in [L]$, the l-th coordinate of s is given by

$$s_l = \sum_{\mathcal{S}\in[H]_{\leq D}} p_\mathcal{S} \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle, \quad \text{where each } v_l^{(h)} = \sum_{i=1}^{M} \sigma_{-i}^{(h)} x_{l-i} = \sigma^{(h)} X_{(l-M):(l-1)}.$$

Here $p_\mathcal{S}$ is defined as in the analysis of Stage 1, and $X_{(l-M):(l-1)} \in \mathbb{R}^{M\times d}$ is the submatrix of X with rows $l-M, \ldots, l-1$.

By direct calculation, we have

$$\frac{\partial \sigma^{(h)}}{\partial w^{(h)}} = \mathrm{diag}(\sigma^{(h)}) - (\sigma^{(h)})^\top \sigma^{(h)} \in \mathbb{R}^{M\times M}, \qquad \frac{\partial v_l^{(h)}}{\partial \sigma^{(h)}} = X_{(l-M):(l-1)}^\top \in \mathbb{R}^{d\times M},$$

Then by chain rule,

$$\frac{\partial v_l^{(h)}}{\partial w^{(h)}} = \frac{\partial v_l^{(h)}}{\partial \sigma^{(h)}} \frac{\partial \sigma^{(h)}}{\partial w^{(h)}} = X_{l-M:l-1}^\top \left(\mathrm{diag}(\sigma^{(h)}) - (\sigma^{(h)})^\top \sigma^{(h)}\right) \in \mathbb{R}^{d\times M}.$$

Moreover, we can view each s_l as a function of $\{v_1^{(h)}, \ldots, v_{L+1}^{(h)}\}_{h\in[H]}$. Differentiating s_l with respect to $v_l^{(h)}$ and $v_{L+1}^{(h)}$, we have

$$\frac{\partial s_l}{\partial v_l^{(h)}} = \sum_{\mathcal{S}\in[H]_{\leq D}\ \text{s.t}\ h\in\mathcal{S}} p_\mathcal{S} \prod_{h'\in\mathcal{S}\setminus\{h\}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle v_{L+1}^{(h)} \in \mathbb{R}^d,$$

$$\frac{\partial s_l}{\partial v_{L+1}^{(h)}} = \sum_{\mathcal{S}\in[H]_{\leq D}\ \text{s.t}\ h\in\mathcal{S}} p_\mathcal{S} \prod_{h'\in\mathcal{S}\setminus\{h\}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle v_l^{(h)} \in \mathbb{R}^{d\times 1}.$$

In the summation, we only add those $\mathcal{S}$'s in $[H]_{\leq D}$ containing h. Also, recall from (E.3) that

$$\frac{\partial \ell}{\partial s_l} = -a \left(\frac{x_{L+1}}{y + \varepsilon \mathbf{1}}\right)^\top (x_l - y) \cdot \sigma_l(as)\,. \tag{E.13}$$

Now we are ready to proceed with the analysis for Stage II.

https://doi.org/10.52202/079017-2127

Proof of Theorem 3.6: Stage II. We start by calculating the explicit expression of the dynamics of $\partial_t w_{-i}^{(h)}$, and then derive approximation of the dynamics, which allows us to further show the convergence of $\sigma^{(h)}$.

Calculation of The Dynamics of $\partial_t w^{(h)}$. First fix an $h \in [H]$. To simplify the notation, for each $l \in [L]$ we define

$$b_l := X_{(l-M):(l-1)} \cdot v_{L+1}^{(h)} + X_{(L+1-M):L} \cdot v_l^{(h)} \in \mathbb{R}^M. \tag{E.14}$$

Note that $w^{(h)}$ is the parameters of the h-th head and only enters each $v_l^{(h)}$, $l = 1, \ldots, L+1$. Recall that $s_l = \sum_{\mathcal{S}\in[H]_{\le D}} p_{\mathcal{S}} \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle$, and the RPE weight $w^{(h)}$ for attention head h only influences its outputs $v_l^{(h)}$ and $v_{L+1}^{(h)}$ in the sum. It thus follows from the chain rule that for each $i \in [M]$, we have

$$\begin{aligned}
\frac{\partial s_l}{\partial w_{-i}^{(h)}} &= \left(\frac{\partial s_l}{\partial v_{L+1}^{(h)}}\right)^{\top} \frac{\partial v_{L+1}^{(h)}}{\partial w_{-i}^{(h)}} + \left(\frac{\partial s_l}{\partial v_l^{(h)}}\right)^{\top} \frac{\partial v_l^{(h)}}{\partial w_{-i}^{(h)}} \\
&= \sum_{\mathcal{S}\in[H]_{\le D} \text{ s.t } h\in\mathcal{S}} p_{\mathcal{S}} \prod_{h'\in\mathcal{S}\setminus\{h\}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle \cdot v_{L+1}^{(h)\top} X_{(l-M):(l-1)}^{\top} \left(\operatorname{diag}(\sigma^{(h)}) - (\sigma^{(h)})^{\top}\sigma^{(h)}\right) e_{M+1-i} \\
&\quad + \sum_{\mathcal{S}\in[H]_{\le D} \text{ s.t } h\in\mathcal{S}} p_{\mathcal{S}} \prod_{h'\in\mathcal{S}\setminus\{h\}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle \cdot v_l^{(h)\top} X_{(L+1-M):L}^{\top} \left(\operatorname{diag}(\sigma^{(h)}) - (\sigma^{(h)})^{\top}\sigma^{(h)}\right) e_{M+1-i} \\
&= \sum_{\mathcal{S}\in[H]_{\le D} \text{ s.t } h\in\mathcal{S}} p_{\mathcal{S}} \prod_{h'\in\mathcal{S}\setminus\{h\}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle \cdot b_l^{\top} \left(e_{M+1-i} - (\sigma^{(h)})^{\top}\right) \cdot \sigma_{-i}^{(h)},
\end{aligned}$$

where we remind readers that $e_i \in \mathbb{R}^{M\times 1}$ is the i-th standard basis vector.

Furthermore, along the gradient flow $\partial_t w_{-i}^{(h)} = -\partial\mathcal{L}/\partial w_{-i}^{(h)}$, it follows from (E.13) that

$$\begin{aligned}
\partial_t w_{-i}^{(h)} &= -\mathbb{E}_{\pi,X}\left[\sum_{l=M+1}^{L} \frac{\partial \ell}{\partial s_l} \frac{\partial s_l}{\partial w_{-i}^{(h)}}\right] = a \sum_{l=M+1}^{L} \mathbb{E}_{\pi,X}\left[\sigma_l\,(as)\left(\frac{x_{L+1}}{y+\varepsilon\mathbf{1}}\right)^{\top}(x_l - y)\,\frac{\partial s_l}{\partial w_{-i}^{(h)}}\right] \\
&= a \sum_{l=M+1}^{L} \mathbb{E}_{\pi,X}\left[\sigma_l\,(as) \sum_{k=1}^{d}\left(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon}\right)\frac{\partial s_l}{\partial w_{-i}^{(h)}}\right] \\
&= a \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\left[g_{h,0}^{\top}\left(e_{M+1-i} - (\sigma^{(h)})^{\top}\right)\sigma_{-i}^{(h)}\right],
\end{aligned}$$

where we plug in the expression of $\partial s_l/\partial w_{-i}^{(h)}$ above in the last equality. Here the vector $g_{h,0}$ is defined as

$$g_{h,0} := \sum_{l=M+1}^{L} \sum_{\substack{\mathcal{S}\in[H]_{\le D} \\ \text{s.t } h\in\mathcal{S}}} \mathbb{E}_{X|\pi}\left[p_{\mathcal{S}}\sigma_l \cdot \sum_{k=1}^{d}\left(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon}\right) \cdot \prod_{h'\in\mathcal{S}\setminus\{h\}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle b_l\right],$$

where σ_l is the softmax probability for the l-th token in the second attention layer. Comparing $\partial_t w_{-i}^{(h)}$ and $\partial_t w_{-h}^{(h)}$, we have

$$\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} = a \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\left[g_{h,0}^{\top}\left(e_{M+1-h} - (\sigma^{(h)})^{\top}\right)\sigma_{-h}^{(h)} - g_{h,0}^{\top}\left(e_{M+1-i} - (\sigma^{(h)})^{\top}\right)\sigma_{-i}^{(h)}\right]. \tag{E.15}$$

Using the fact that $\sum_{j=1}^{M} \sigma_{-j}^{(h)} = 1$, we can rewrite

$$\begin{aligned}
&\left(e_{M+1-h} - (\sigma^{(h)})^{\top}\right)\sigma_{-h}^{(h)} - \left(e_{M+1-i} - (\sigma^{(h)})^{\top}\right)\sigma_{-i}^{(h)} \\
&\quad = \sigma_{-i}^{(h)}(e_{M+1-h} - e_{M+1-i}) + (\sigma_{-h}^{(h)} - \sigma_{-i}^{(h)})(e_{M+1-h} - (\sigma^{(h)})^{\top}). \\
&\quad = \sigma_{-i}^{(h)}(e_{M+1-h} - e_{M+1-i}) + (\sigma_{-h}^{(h)} - \sigma_{-i}^{(h)}) \sum_{j=1}^{M} \sigma_{-j}^{(h)}(e_{M+1-h} - e_{M+1-j}),
\end{aligned} \tag{E.16}$$

https://doi.org/10.52202/079017-2127

where in the first identity, we add and then subtract term $\sigma_{-i}^{(h)} e_{M+1-h}$. Combining (E.15) and (E.16) yields for each $i \in [M]$ that

$$
\begin{aligned}
&\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} \qquad\qquad \text{(E.17)}\\
&\quad = a \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\biggl[g_{h,0}^\top\biggl(\sigma_{-i}^{(h)}\left(e_{M+1-h} - e_{M+1-i}\right) + \left(\sigma_{-h}^{(h)} - \sigma_{-i}^{(h)}\right)\sum_{j=1}^{M}\sigma_{-j}^{(h)}(e_{M+1-h} - e_{M+1-j})\biggr)\biggr].
\end{aligned}
$$

Simplification of $\partial_t w_{-i}^{(h)}$. We proceed by deriving approximations to the vector $g_{h,0}$, which will help us identify the dominant term in the dynamics $\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)}$. Specifically, we define

$$
\begin{aligned}
g_{h,1} &:= \sum_{l=M+1}^{L}\mathbb{E}_{X|\pi}\biggl[\sigma_l(as)\sum_{k=1}^{d}\biggl(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon}\biggr)\prod_{h'\in\mathcal{S}^\star\setminus\{h\}}\langle v_l^{(h')}, v_{L+1}^{(h')}\rangle b_l\biggr],\\
g_{h,2} &:= \frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}_{X|\pi}\biggl[\sum_{k=1}^{d}\biggl(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\bar{y}(k)+\varepsilon} - \frac{\bar{y}(k)\,\mathbb{1}(x_{L+1} = e_k)}{\bar{y}(k)+\varepsilon}\biggr)\prod_{h'\in\mathcal{S}^\star\setminus\{h\}}\langle v_l^{(h')}, v_{L+1}^{(h')}\rangle b_l\biggr],\\
g_{h,3} &:= \frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}_{X|\pi}\biggl[\biggl(\sum_{k=1}^{d}\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\mu^\pi(e_k)} - 1\biggr)\prod_{h'\in\mathcal{S}^\star\setminus\{h\}}\langle v_l^{(h')}, v_{L+1}^{(h')}\rangle b_l\biggr],\\
g_{h,4} &:= \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\biggl[\biggl(\sum_{k=1}^{d}\frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(e_k)} - 1\biggr)\prod_{h'\in\mathcal{S}^\star\setminus\{h\}}\langle v^{(h')}(Z), v^{(h')}(X)\rangle b(X,Z)\biggr],
\end{aligned}
$$

where $Z = [z_{-M}, \ldots, z_{-1}]^\top \in \mathbb{R}^{M\times d}$ is an independent copy of $X = [x_{-M}, \ldots, x_{-1}]^\top \in \mathbb{R}^{M\times d}$, and

$$
v^{(h)}(X) := \sum_{i=1}^{M}\sigma_{-i}^{(h)}x_{-i}, \quad v^{(h)}(Z) := \sum_{i=1}^{M}\sigma_{-i}^{(h)}z_{-i},
$$

$$
b(X,Z) := Z(v^{(h)}(X)) + X(v^{(h)}(Z)), \quad \bar{y} := \frac{1}{L-M}\sum_{l=M+1}^{L}x_l.
$$

The strategy of gradually approximating $g_{h,0}$ by $g_{h,1}, g_{h,2}, g_{h,3}$ and $g_{h,4}$ is similar to the analysis in Stage I. To see the intuition, from $g_{h,0}$ to $g_{h,1}$, we use the fact that $p_{\mathcal{S}^\star} \approx 1$ and $p_{\mathcal{S}} \approx 0$ for any other $\mathcal{S}$, which is a result of Stage 1. From $g_{h,1}$ to $g_{h,2}$, we replace y by the empirical mean $\bar{y}$, thanks to the fact that $\sigma_l(a) \approx 1/L$ when a is small. Then, from $g_{h,2}$ to $g_{h,3}$, we replace the empirical distribution $\bar{y}$ with the stationary distribution of the Markov chain. These two steps also appear in the analysis of Stage 1. Finally, to go from $g_{h,3}$ to $g_{h,4}$, we leverage the rapid mixing of the Markov chain.

Note that the common structures in (E.15) are $g_{h,0}^\top(e_{M+1-h} - e_{M+1-i})$ for $i \neq h$. Hence, we only need to understand the approximation error in each step for $g_{h,0}^\top(e_{M+1-h} - e_{M+1-i})$. Recall that we are focusing on $h \in \mathcal{S}^\star$ in this stage.

- From $g_{h,0}$ to $g_{h,1}$, we remove the terms in the summation that are weighted down by $p_{\mathcal{S}}$ for any $\mathcal{S} \neq \mathcal{S}^\star$ due to the rapid dominance of $p_{\mathcal{S}^\star}$ from Stage I. Recall that $p_{\mathcal{S}^\star}$ converges to one at an exponential rate while all other $p_{\mathcal{S}}$'s converge to zero. For simplicity, let us define

$$
\begin{aligned}
\rho(\mathcal{S}) := \sum_{l=M+1}^{L}\mathbb{E}_{X|\pi}\biggl[\sigma_l(as)\sum_{k=1}^{d}\biggl(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon}\biggr)&\\
\cdot\prod_{h'\in\mathcal{S}\setminus\{h\}}\langle v_l^{(h')}, v_{L+1}^{(h')}\rangle b_l\biggr](e_{M+1-h} - e_{M+1-i})&.
\end{aligned}
$$

By the triangular inequality, we have

$$\left|(g_{h,0}-g_{h,1})^\top(e_{M+1-h}-e_{M+1-i})\right| = \Bigg|\sum_{\substack{\mathcal{S}\in[H]_{\le D}\setminus\{\mathcal{S}^\star\}\\ \text{s.t } h\in\mathcal{S}}} p_{\mathcal{S}}\cdot\rho(\mathcal{S})-\rho(\mathcal{S}^\star)\Bigg|$$

$$\le (1-p_{\mathcal{S}^\star})\cdot\left|\rho(\mathcal{S}^\star)\right| + \sum_{\mathcal{S}\in[H]_{\le D}\setminus\{\mathcal{S}^\star\}} p_{\mathcal{S}}\cdot\left|\rho(\mathcal{S})\right| \le 16(1-p_{\mathcal{S}^\star}),$$

where in the last line we use the claim that $|\rho(\mathcal{S})|\le 8$ for all $\mathcal{S}$. To see this point, note that by definition of b_l in (E.14), we have

$$|b_l^\top(e_{M+1-h}-e_{M+1-i})| = \left|\langle v_{L+1}^{(h)}, x_{l-h}-x_{l-i}\rangle - \langle v_l^{(h)}, x_{L+1-h}-x_{L+1-i}\rangle\right| \le 4,$$
$$\Bigg|\prod_{h'\in\mathcal{S}\setminus\{h\}}\langle v_l^{(h')}, v_{L+1}^{(h')}\rangle\Bigg| \le 1,$$

since $v_l^{(h)}$ and x_l have norm at most 1. Then, by Lemma F.2 where we plug in the upper bound 4 for the function $f(\cdot)$ in the lemma, we conclude that $\rho(\mathcal{S})\le 8, \quad \forall\mathcal{S}\in[H]_{\le D}\setminus\{\mathcal{S}^\star\}$. Define $\Delta_1 := 1-p_{\mathcal{S}^\star}(t_1)$, and $\Delta_1\le 1/L$ by the results from Stage I. Thus, we obtain

$$\left|(g_{h,0}-g_{h,1})^\top(e_{M+1-h}-e_{M+1-i})\right| \le 16\Delta_1 \le 16/L.$$

- For the approximation of $g_{h,1}$ by $g_{h,2}$, we use the fact that $\sigma_l(as)\approx 1/L$ when a is sufficiently small. Specifically, we also take the absolute bound for $f(\cdot)$ as 4 in Lemma F.3 and obtain

$$\left|(g_{h,1}-g_{h,2})^\top(e_{M+1-h}-e_{M+1-i})\right| \le \frac{32ad}{\varepsilon^2}.$$

- For the approximation of $g_{h,2}$ by $g_{h,3}$, we use the fact that $\overline{y}(k)\approx\mu^\pi(e_k)$ for large L. More precisely, it follows from Lemma F.4 with the upper bound 4 for $f(\cdot)$ in the lemma that

$$\left|(g_{h,2}-g_{h,3})^\top(e_{M+1-h}-e_{M+1-i})\right| \le 16\cdot\frac{(1-\lambda)^{-1/2}(D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1)^{1/4}+2\sqrt{M}}{L^{1/2}\gamma} + 4\gamma^{-1}\varepsilon.$$

- Finally, to go from $g_{h,3}$ to $g_{h,4}$, we leverage the rapid mixing of the Markov chain. Intuitively, when l and $L+1$ are far apart, x_l and its parents in $\mathcal{S}^\star$ are independent of x_{L+1} and its parents in $\mathcal{S}^\star$. This observation yields the approximation of $g_{h,3}^\top(e_{M+1-h}-e_{M+1-i})$ by $g_{h,4}^\top(e_{M+1-h}-e_{M+1-i})$. To simplify the notation, define two scalars

$$\widetilde{g}_{h,l} := \left(\sum_{k=1}^{d}\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\mu^\pi(e_k)}-1\right)\prod_{h'\in\mathcal{S}^\star\setminus\{h\}}\langle v_l^{(h')}, v_{L+1}^{(h')}\rangle,$$
$$\widetilde{g}_{h,4} := \left(\sum_{k=1}^{d}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(e_k)}-1\right)\prod_{h'\in\mathcal{S}^\star\setminus\{h\}}\langle v^{(h')}(Z), v^{(h')}(X)\rangle.$$

Using the notation above, we have

$$\left|(g_{h,3}-g_{h,4})^\top(e_{M+1-h}-e_{M+1-i})\right|$$
$$= \left|\left(\sum_{l=M+1}^{L}\frac{\mathbb{E}_{X|\pi}\left[\widetilde{g}_{h,l}b_l^\top\right]}{L-M} - \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\widetilde{g}_{h,4}b(X,Z)^\top\right]\right)(e_{M+1-h}-e_{M+1-i})\right|.$$

Recall that

$$b_l^\top(e_{M+1-h}-e_{M+1-i}) = \langle v_{L+1}^{(h)}, x_{l-h}-x_{l-i}\rangle - \langle v_l^{(h)}, x_{L+1-h}-x_{L+1-i}\rangle,$$
$$b(X,Z)^\top(e_{M+1-h}-e_{M+1-i}) = \langle v^{(h)}(X), z_{-h}-z_{-i}\rangle - \langle v^{(h)}(Z), x_{-h}-x_{-i}\rangle.$$

We apply the triangular inequality to obtain that

$$\begin{aligned}
&\left|(g_{h,3}-g_{h,4})^\top\left(e_{M+1-h}-e_{M+1-i}\right)\right| \\
&\quad\le \left|\frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}_{X|\pi}\big[\widetilde{g}_{h,l}\langle v_{L+1}^{(h)}, x_{l-h}\rangle\big] - \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\big[\widetilde{g}_{h,4}\langle v^{(h)}(Z), x_{-h}\rangle\big]\right| \\
&\qquad + \left|\frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}_{X|\pi}\big[\widetilde{g}_{h,l}\langle v_{L+1}^{(h)}, x_{l-i}\rangle\big] - \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\big[\widetilde{g}_{h,4}\langle v^{(h)}(Z), x_{-i}\rangle\big]\right| \\
&\qquad + \left|\frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}_{X|\pi}\big[\widetilde{g}_{h,l}\langle v_{l}^{(h)}, x_{L+1-h}\rangle\big] - \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\big[\widetilde{g}_{h,4}\langle v^{(h)}(X), z_{-h}\rangle\big]\right| \\
&\qquad + \left|\frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}_{X|\pi}\big[\widetilde{g}_{h,l}\langle v_{l}^{(h)}, x_{L+1-i}\rangle\big] - \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\big[\widetilde{g}_{h,4}\langle v^{(h)}(X), z_{-h}\rangle\big]\right|.
\end{aligned}$$

Each term on the right-hand side can be bounded by Lemma F.5, where in the lemma we take $(\sigma^{(h)})_{h'\in\mathcal{S}^\star}\in\mathbb{R}^{M\times|\mathcal{S}^\star|}$ and $((\sigma^{(h')})_{h'\in\mathcal{S}^\star\backslash\{h\}}, e_h)\in\mathbb{R}^{M\times|\mathcal{S}^\star|}$ as the two lists of vectors on the M-dimensional probability simplex for $\widetilde{\sigma}$ and σ respectively. Consequently, we have

$$\left|(g_{h,3}-g_{h,4})^\top\left(e_{M+1-h}-e_{M+1-i}\right)\right| \le \frac{8M}{L\gamma} + \frac{16}{L(1-\lambda)\gamma^{|\mathcal{S}|/2+r_n/2+1}}.$$

Combining the above results and setting $\varepsilon = 1/\sqrt{L}$, $a = a(0) \le O(1/L^{3/2})$ and together with the conditions in (E.1), we have

$$\left|(g_{h,0}-g_{h,4})^\top\left(e_{M+1-h}-e_{M+1-i}\right)\right| = |\mathcal{E}| = O\left(\frac{1}{\sqrt{L(1-\lambda)\gamma^{r_n+2}}}\right),$$

where $O(\cdot)$ hides universal constants independent of the parameters of the model. We remark that while the left hand side is a function of t, the upper bound is independent of t. Then, we can rewrite (E.17) as

$$\begin{aligned}
&\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} \\
&\quad = a\cdot\mathbb{E}_{\pi\sim\mathcal{P}}\left[\sigma_{-i}^{(h)}\cdot g_{h,4}^\top\left(e_{M+1-h}-e_{M+1-i}\right) + (\sigma_{-h}^{(h)}-\sigma_{-i}^{(h)})\cdot\sum_{j=1}^{M}\sigma_{-j}^{(h)}\cdot g_{h,4}^\top(e_{M+1-h}-e_{M+1-j})\right] \\
&\qquad \pm a\left(\sigma_{-i}^{(h)} + (\sigma_{-h}^{(h)}-\sigma_{-i}^{(h)})\sum_{j=1,j\neq h}^{M}\sigma_{-j}^{(h)}\right)\cdot|\mathcal{E}|. \qquad \text{(E.18)}
\end{aligned}$$

Lower Bound for The Difference $\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)}$. To show $\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} > 0$, we first derive the lower bound of $\mathbb{E}_{\pi\sim\mathcal{P}}[g_{h,4}^\top\left(e_{M+1-h}-e_{M+1-i}\right)]$ for any $i\neq h$. Since (x,X) and (z,Z) are independent and identically distributed, by the definition of $b(X,Z)$,

$$\begin{aligned}
&\mathbb{E}_{\pi\sim\mathcal{P}}\left[g_{h,4}^\top\left(e_{M+1-h}-e_{M+1-i}\right)\right] \\
&\quad = 2\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\sum_{k=1}^{d}\left(\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(e_k)}-1\right)\prod_{h'\in\mathcal{S}^\star\backslash\{h\}}\langle v^{(h')}(Z), v^{(h')}(X)\rangle\cdot\langle v^{(h)}(X), z_{-h}\rangle\right] \\
&\qquad - 2\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\sum_{k=1}^{d}\left(\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(e_k)}-1\right)\prod_{h'\in\mathcal{S}^\star\backslash\{h\}}\langle v^{(h')}(Z), v^{(h')}(X)\rangle\cdot\langle v^{(h)}(X), z_{-i}\rangle\right] \\
&\quad = 2\tau_{h,1} - 2\tau_{h,2},
\end{aligned}$$

where we introduce the following quantities for convenience:

$$\tau_{h,1} := \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\sum_{k=1}^{d}\left(\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(e_k)}-1\right)\prod_{h'\in\mathcal{S}^\star\backslash\{h\}}\langle v^{(h')}(Z), v^{(h')}(X)\rangle\cdot\langle v^{(h)}(X), z_{-h}\rangle\right],$$

$$\tau_{h,2} := \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\sum_{k=1}^{d}\left(\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(e_k)}-1\right)\prod_{h'\in\mathcal{S}^\star\backslash\{h\}}\langle v^{(h')}(Z), v^{(h')}(X)\rangle\cdot\langle v^{(h)}(X), z_{-i}\rangle\right].$$

The quantities $\tau_{h,1}$ and $\tau_{h,2}$ can be further approximated. Specifically, by applying Lemma F.6 to $\tau_{h,1}$, , where in the lemma we take $(\sigma^{(h)})_{h'\in\mathcal{S}^\star} \in \mathbb{R}^{M\times|\mathcal{S}^\star|}$ and $((\sigma^{(h')})_{h'\in\mathcal{S}^\star\backslash\{h\}}, e_h) \in \mathbb{R}^{M\times|\mathcal{S}^\star|}$ as the two lists of vectors on the M-dimensional probability simplex for σ and $\widetilde{\sigma}$ respectively, and we obtain

$$\left|\tau_{h,1} - \prod_{h'\in\mathcal{S}^\star\backslash\{h\}}(\sigma_{-h'}^{(h')})^2\cdot\sigma_{-h}^{(h)}\cdot\widetilde{I}_{\chi^2}(\mathcal{S}^\star)\right| \le \left(1-\prod_{h'\in\mathcal{S}^\star\backslash\{h\}}(\sigma_{-h'}^{(h')})^2\cdot\sigma_{-h}^{(h)}\right)\widetilde{I}_{\chi^2}(\mathcal{S}^\star). \tag{E.19}$$

Drawing on the analagous reasoning as in the proof of Lemma F.6, we can approximate $\tau_{h,2}$ as follows:

$$\left|\tau_{h,2} - \prod_{h'\in\mathcal{S}^\star\backslash\{h\}}(\sigma_{-h'}^{(h')})^2\cdot\sigma_{-h}^{(h)}\cdot\psi\right| \le \left(1-\prod_{h'\in\mathcal{S}^\star\backslash\{h\}}(\sigma_{-h'}^{(h')})^2\cdot\sigma_{-h}^{(h)}\right)\widetilde{I}_{\chi^2}(\mathcal{S}^\star), \tag{E.20}$$

where

$$\psi := \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\prod_{h'\in\mathcal{S}^\star\backslash\{h\}}\mathbb{1}(x_{-h'}=z_{-h'})\cdot\mathbb{1}(x_{-h}=z_{-i})\cdot\left(\sum_{k=1}^{d}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(e_k)}-1\right)\right].$$

To establish the lower bound for $\tau_{h,1}-\tau_{h,2}$, let us begin by establishing an upper bound for ψ, which is approximately equal to $\tau_{h,2}$. We invoke Lemma F.7 with $\mathcal{S}=\mathcal{S}^\star$ and $\mathcal{S}'=\mathcal{S}^\star\backslash\{h\}\cup\{i\}$ in the lemma to obtain

$$\psi \le \frac{1}{2}\widetilde{I}_{\chi^2}(\mathcal{S}^\star)+\frac{1}{2}\widetilde{I}_{\chi^2}(\mathcal{S}^\star\backslash\{h\}\cup\{i\}) \le \widetilde{I}_{\chi^2}(\mathcal{S}^\star)-\frac{1}{2}\cdot\Delta\widetilde{I}_{\chi^2}, \quad \forall i\neq h$$

Leveraging this for (E.19) and (E.20),

$$\begin{aligned}2\tau_{h,1}-2\tau_{h,2} &\ge \prod_{h'\in\mathcal{S}^\star\backslash\{h\}}(\sigma_{-h'}^{(h')})^2\cdot\sigma_{-h}^{(h)}\cdot\Delta\widetilde{I}_{\chi^2}-4\left(1-\prod_{h'\in\mathcal{S}^\star\backslash\{h\}}(\sigma_{-h'}^{(h')})^2\cdot\sigma_{-h}^{(h)}\right)\widetilde{I}_{\chi^2}(\mathcal{S}^\star)\\ &\ge \prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)})^2\cdot\Delta\widetilde{I}_{\chi^2}-4\left(1-\prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)})^2\right)\widetilde{I}_{\chi^2}(\mathcal{S}^\star), \end{aligned}\tag{E.21}$$

where in the second line we multiply an additional $\sigma_{-h}^{(h)}$ to the product as $\sigma_{-h}^{(h)}\in[0,1]$.

Next, we provide a lemma showing that $\partial_t\sigma_{-h}^{(h)}$ is growing for all time $t\ge t_1$, where t_1 is the starting time of the second stage.

Lemma E.2 (Reinforced Growth of $\sigma_{-h}^{(h)}$)**.** *For all $h\in\mathcal{S}^\star$, we have for all $i\neq h$ at any $t\ge t_1$:*

$$\partial_t\sigma_{-h}^{(h)}>0, \quad \textit{and} \quad \partial_t\log\sigma_{-h}^{(h)}-\partial_t\log\sigma_{-i}^{(h)}=\partial_t w_{-h}^{(h)}-\partial_t w_{-i}^{(h)}>0. \tag{E.22}$$

Proof. See §E.3.1 for the proof. □

In the proof of Lemma E.2, we will use the following useful proposition.

Proposition E.3. *Suppose $\prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)})^2\ge 1/(1+(M-1)\exp(-\Delta w))^{2|\mathcal{S}^\star|}$ with Δw satisfying (3.6), and $\sigma_{-h}^{(h)}>\sigma_{-i}^{(h)}$ for any $i\neq h, h\in\mathcal{S}^\star$ at a given time t. Suppose Assumption 3.5 holds and L satisfies (E.1). It holds that*

$$\begin{gathered}\partial_t\log\sigma_{-h}^{(h)}-\partial_t\log\sigma_{-i}^{(h)}=\partial_t w_{-h}^{(h)}-\partial_t w_{-i}^{(h)}\ge\frac{a\Delta\widetilde{I}_{\chi^2}}{2}\left(\sigma_{-i}^{(h)}+(\sigma_{-h}^{(h)}-\sigma_{-i}^{(h)})\sum_{j=1,j\neq h}^{M}\sigma_{-j}^{(h)}\right)>0,\\ \partial_t\sigma_{-h}^{(h)}>0, \qquad \forall i\neq h, \quad \forall h\in\mathcal{S}^\star.\end{gathered}\tag{E.23}$$

Proof. See §E.3.1 for the proof. □

Lemma E.2 implies that during Stage II, for all $i \neq h$ and $h \in \mathcal{S}^\star$, we have $w_{-h}^{(h)} > w_{-i}^{(h)}$ and $\sigma_{-h}^{(h)} > \sigma_{-i}^{(h)}$ for all $t \geq t_1$. In addition, as $\sigma_{-h}^{(h)}$ is growing, all the conditions in Proposition E.3 are satisfied for any $t \geq t_1$, and hence all the conclusions in (E.23).

Convergence of $\sigma^{(h)}$. Finally, we characterize the convergence rate of $\sigma^{(h)}$. For the convergence analysis, we adhere to the convention used in the previous stage, treating all model parameters as functions of the training time t, where $t = t_1$ marks the start of the second stage. With a slight abuse of notation, we denote by $\sigma_{-i}^{(h)}(t)$ the value of $\sigma_{-i}(w^{(h)}(t))$ at time t, where $w^{(h)}(t)$ is the input to the softmax function, and $\sigma_{-i}(\cdot)$ refers to the $(M+1-i)$-th element of the softmax probability. For simplicity, we sometimes omit the time index t when the context makes it clear.

Note that $\partial_t \sigma_{-h}^{(h)} > 0$ for all $h \in \mathcal{S}^\star$. Hence by the definition of the softmax operation, we have

$$\begin{aligned} \sigma_{-i}^{(h)} &= \sigma_{-h}^{(h)} \cdot \exp(-(w_{-h}^{(h)} - w_{-i}^{(h)})) \geq \sigma_{-h}^{(h)}(t_1) \cdot \exp(-(w_{-h}^{(h)} - w_{-i}^{(h)}) \\ &= \sigma_{-h}^{(h)}(0) \cdot \exp(-(w_{-h}^{(h)} - w_{-i}^{(h)})), \end{aligned} \tag{E.24}$$

where the first inequality follows from the monotone growth of $\sigma_h^{(h)}$, and the second line follows from the fact that the first attention layer is untouched during the first stage. Note that here in (E.24), $\sigma^{(h)}$ and $w^{(h)}$ are functions of t. Now, putting together (E.23) and (E.24), and also noting that $\sigma_{-h}^{(h)} > \sigma_{-i}^{(h)}$ for all $i \neq h$ and $h \in \mathcal{S}^\star$, it follows that

$$\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} \geq \frac{a\Delta\widetilde{I}_{\chi^2}}{2} \sigma_{-i}^{(h)} \geq \frac{a\Delta\widetilde{I}_{\chi^2}}{2} \cdot \sigma_{-h}^{(h)}(0) \cdot \exp(-(w_{-h}^{(h)} - w_{-i}^{(h)})).$$

Rearranging the terms, and using the fact that $w_{-h}^{(h)}(t_1) - w_{-i}^{(h)}(t_1) \geq \Delta w$ by Assumption 3.3, we get

$$\exp\left(w_{-h}^{(h)}(t) - w_{-i}^{(h)}(t)\right) \geq \frac{a\Delta\widetilde{I}_{\chi^2} \cdot \sigma_{-h}^{(h)}(0)}{2} \cdot (t - t_1) + \exp(\Delta w).$$

This yields a lower bound for $\sigma_{-h}^{(h)}(t)$ as follows:

$$\sigma_{-h}^{(h)}(t) = \frac{1}{1 + \sum_{i \neq h} \exp(w_{-i}^{(h)}(t) - w_{-h}^{(h)}(t))} \geq \frac{1}{1 + (M-1) \cdot (a\Delta\widetilde{I}_{\chi^2} \cdot \sigma_{\min}(0) \cdot (t - t_1)/2 + \exp(\Delta w))^{-1}},$$

where we define $\sigma_{\min}(0) := \min_{h \in \mathcal{S}^\star} \sigma_{-h}^{(h)}(0)$. Consequently, we have

$$\begin{aligned} 1 - \prod_{h \in \mathcal{S}^\star} (\sigma_{-h}^{(h)}(t))^2 &\leq 1 - \left(\frac{1}{1 + (M-1) \cdot (a\Delta\widetilde{I}_{\chi^2} \cdot \sigma_{\min}(0) \cdot (t - t_1)/2 + \exp(\Delta w))^{-1}}\right)^{2|\mathcal{S}^\star|} \\ &= 1 - \left(1 - \frac{(M-1)}{(a\Delta\widetilde{I}_{\chi^2} \cdot \sigma_{\min}(0) \cdot (t - t_1)/2 + \exp(\Delta w)) + (M-1)}\right)^{2|\mathcal{S}^\star|}. \end{aligned}$$

Now, we consider large t such that

$$\frac{(M-1)}{(a\Delta\widetilde{I}_{\chi^2} \cdot \sigma_{\min}(0) \cdot (t - t_1)/2 + \exp(\Delta w)) + (M-1)} < \frac{1}{2|\mathcal{S}^\star|}.$$

Then, we can apply the inequality $(1-x)^n \geq 1 - nx$ for $x \in [0, 1/n]$ and $n \geq 1$ to obtain

$$1 - \prod_{h \in \mathcal{S}^\star} (\sigma_{-h}^{(h)}(t))^2 \leq \frac{2|\mathcal{S}^\star| \cdot (M-1)}{a\Delta\widetilde{I}_{\chi^2} \cdot \sigma_{\min}(0) \cdot (t - t_1)/2 + \exp(\Delta w) + (M-1)}.$$

Therefore, with training time $t_2 = 4L|\mathcal{S}^\star| \cdot (M-1)/a\Delta\widetilde{I}_{\chi^2} \cdot \sigma_{\min}(0) + t_1$, we can ensure that

$$1 - \prod_{h \in \mathcal{S}^\star} (\sigma_{-h}^{(h)}(t_2))^2 \leq L^{-1}.$$

This completes the proof for Stage II. □

E.3.1 Additional Proofs for Stage II

We conclude this subsection with the proof of Lemma E.2 and Proposition E.3.

Proof of Proposition E.3. The condition $\prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)})^2 \geq 1/(1+(M-1)\exp(-\Delta w))^{2|\mathcal{S}^\star|}$ with Δw in (3.6) implies that

$$\prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)})^2 \geq \left(1+\frac{\Delta\widetilde{I}_{\chi^2}}{14\widetilde{I}_{\chi^2}(\mathcal{S}^\star)}\right)^{-1} \geq \frac{4\widetilde{I}_{\chi^2}(\mathcal{S}^\star)+\frac{2}{3}\Delta\widetilde{I}_{\chi^2}}{4\widetilde{I}_{\chi^2}(\mathcal{S}^\star)+\Delta\widetilde{I}_{\chi^2}}. \tag{E.25}$$

Combining (E.21) and (E.25) yields

$$\mathbb{E}_{\pi\sim\mathcal{P}}\left[g_{h,4}^\top(e_{M+1-h}-e_{M+1-i})\right] = 2\tau_{h,1}-2\tau_{h,2} \geq \frac{2}{3}\Delta\widetilde{I}_{\chi^2} \tag{E.26}$$

for any $i\neq h$. Applying (E.26) to (E.18), since each $\sigma_{-i}^{(h)}>0$ and $\sigma_{-h}^{(h)}>\sigma_{-i}^{(h)}$ at time t for all $i\neq h, h\in\mathcal{S}^\star$, it holds that

$$\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} \geq a\Bigg(\sigma_{-i}^{(h)} + (\sigma_{-h}^{(h)}-\sigma_{-i}^{(h)})\cdot\sum_{j=1,j\neq h}^{M}\sigma_{-j}^{(h)}\Bigg)\cdot\Bigg(\frac{2}{3}\Delta\widetilde{I}_{\chi^2}-|\mathcal{E}|\Bigg).$$

Then since we assume a sufficiently large $L\geq\Omega((\Delta\widetilde{I}_{\chi^2}^2(1-\lambda)\gamma^{r_n+2})^{-1})$, it holds that $|\mathcal{E}|\leq\Delta\widetilde{I}_{\chi^2}/6$, we further have

$$\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} \geq \frac{a\Delta\widetilde{I}_{\chi^2}}{2}\Bigg(\sigma_{-i}^{(h)} + (\sigma_{-h}^{(h)}-\sigma_{-i}^{(h)})\sum_{j=1,j\neq h}^{M}\sigma_{-j}^{(h)}\Bigg) > 0, \quad \forall i\neq h, \quad \forall h\in\mathcal{S}^\star.$$

As $\partial_t\log\sigma_{-h}^{(h)}-\partial_t\log\sigma_{-i}^{(h)} = \partial_t w_{-h}^{(h)}-\partial_t w_{-i}^{(h)} > 0$ by property of the softmax function, and $\sum_{i=1}^M\partial_t\sigma_{-i}^{(h)}=0$, we have $\partial_t\sigma_{-h}^{(h)}>0$ for all $h\in\mathcal{S}^\star$. This completes the proof of Proposition E.3. □

Proof of Lemma E.2. We give a proof to Lemma E.2 by contradiction. Note that at the beginning of the second stage $t=t_1$, we have all the conditions for Proposition E.3 satisfied by the initialization conditions in Assumption 3.3. Then, by (E.23) in Proposition E.3, we have $\partial_t\log\sigma_{-h}^{(h)}-\partial_t\log\sigma_{-i}^{(h)}>0$ and $\partial_t\sigma_{-h}^{(h)}>0$ for all $i\neq h$ and $h\in\mathcal{S}^\star$ at $t=t_1$.

Next, assume that $\tau>t_1$ is the smallest time such that at least $\partial_t\sigma_{-h}^{(h)}\leq 0$ or $\partial_t\log\sigma_{-h}^{(h)}-\partial_t\log\sigma_{-i}^{(h)}\leq 0$ for some $i\neq h$ and $h\in\mathcal{S}^\star$. By definition of τ, we have (E.22) holds for any moment $t\in[t_1,\tau)$. As $\sigma_{-h}^{(h)}$ and the gap $\sigma_{-h}^{(h)}-\sigma_{-i}^{(h)}$ are monotonically increasing, we have by the initialization condition and the boundedness of the gradient that at time τ:

$$\prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)})^2 \geq 1/(1+(M-1)\exp(-\Delta w))^{2|\mathcal{S}^\star|}, \quad \text{and} \quad \sigma_{-h}^{(h)}>\sigma_{-i}^{(h)}, \quad \forall i\neq h, \quad \forall h\in\mathcal{S}^\star.$$

Hence, by Proposition E.3, we have $\partial_t\log\sigma_{-h}^{(h)}-\partial_t\log\sigma_{-i}^{(h)}>0$ and $\partial_t\sigma_{-h}^{(h)}>0$ for all $i\neq h$ and $h\in\mathcal{S}^\star$ at time τ, which contradicts the definition of τ. This completes the proof of Lemma E.2. □

E.4 Analysis for Stage III

In this section, we derive the dynamics of the second attention layer's weights a in Stage III. We characterize the dynamics of a when $a<O(\log L)$, where the signal term of the dynamics dominates the approximation error. We provide the growth rate of the weights for two regimes: when a is either sufficiently small or large.

https://doi.org/10.52202/079017-2127

Proof Strategy. We analyze the dynamics of a via the following steps:

1. **Dynamics Calculation.** First, we derive the explicit expression for the dynamics of a.
2. **Dynamics Approximation.** We approximate the dynamics by exploiting the mixing properties of the Markov chain and the convergence of the weights from Stage I and II.
3. **Lower and Upper Bound for The Growth Rate.** Finally, we establish the upper and lower bounds for the growth rate of a when a is either sufficiently small or large.

For a set $\mathcal{S} \subseteq [M]$, we denote $X_{l-\mathcal{S}} := (x_{l-s} : s \in \mathcal{S})$. If $l = 0$, we will ignore l in the subscript and simply use $X_{-\mathcal{S}}$. In this section, we abbreviate $p_{\mathcal{S}}(t_1)$ after the first stage's training as $p_{\mathcal{S}}$, and $\sigma_{-i}^{(h)}(t_2)$ after the second stage's training as $\sigma_{-i}^{(h)}$.

Proof of Theorem 3.6: Stage III. We start with the explicit expression of the dynamics of a.

Calculation of The Dynamics of a. First by the chain rule,

$$\frac{\partial \ell}{\partial a} = \sum_{l=M+1}^{L} \frac{\partial \ell}{\partial (as_l)} \frac{\partial (as_l)}{\partial a} = -\sum_{l=M+1}^{L} \left(\frac{x_{L+1}}{y + \varepsilon \mathbf{1}}\right)^{\top} (x_l - y) \cdot \sigma_l(as) \cdot s_l.$$

where in the last equality we remind readers of the same procedure as we have used in the derivation of (E.3) in Stage I. Then, taking expectation with respect to X and π and expanding $s_l = a \sum_{\mathcal{S}\in[H]_{\leq D}} p_{\mathcal{S}} \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle$, we have

$$\begin{aligned}\partial_t a = -\frac{\partial \mathcal{L}}{\partial a} &= \mathbb{E}\Bigg[\sum_{l=M+1}^{L} \left(\frac{x_{L+1}}{y + \varepsilon \mathbf{1}}\right)^{\top} (x_l - y) \cdot \sigma_l\,(as) \cdot s_l\Bigg] \\ &= \mathbb{E}\Bigg[\sum_{\mathcal{S}\in[H]_{\leq D}} p_{\mathcal{S}} \sum_{l=M+1}^{L} \sigma_l \sum_{k=1}^{d} \bigg(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon}\bigg) \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Bigg] =: f_0\end{aligned}$$

We remind readers the shorthand $\sigma \equiv \sigma(as)$. We denote the above quantity by f_0.

Approximation of $\partial_t a$. Similar to the analysis for the previous two stages, we develop a sequence of approximation steps that transforms $\partial_t a$ into a tractable quantity. We aim to decouple x_{L+1} and x_l, approximate s_l by a population version, and transform the expectation to one under the stationary distribution of the Markov chain. Specifically, the approximation involves the following steps:

- Our first step is to remove the summation over $[H]_{\leq D} \backslash \{\mathcal{S}^\star\}$ where $\mathcal{S}^\star$ is the optimal set that maximizes the modified mutual information defined in (3.1). This is because $c_{\mathcal{S}^\star}$ dominates by the analysis of Stage I. Specifically, we define

  $$f_1 := \mathbb{E}\Bigg[\sum_{l=M+1}^{L} \sigma_l \sum_{k=1}^{d} \bigg(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon}\bigg) \prod_{h\in\mathcal{S}^\star} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Bigg].$$

  To bound $|f_0 - f_1|$, note that for any $\mathcal{S} \in [H]_{\leq D}$, since each $v_l^{(h)}$ has norm at most 1, we can invoke Lemma F.2 with $C = 1$ and obtain

  $$\left|\sum_{l=M+1}^{L} \sigma_l \sum_{k=1}^{d} \bigg(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon}\bigg) \prod_{h\in\mathcal{S}^\star} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\right| \leq 2.$$

  It follows that

$$\begin{aligned}|f_0 - f_1| = \mathbb{E}\Bigg[&\sum_{\mathcal{S}\in[H]_{\leq D}\backslash\{\mathcal{S}^\star\}} p_{\mathcal{S}} \sum_{l=M+1}^{L} \sigma_l \sum_{k=1}^{d} \mathbb{1}(x_{L+1} = e_k) \bigg(\frac{\mathbb{1}(x_l = e_k)}{y(k)+\varepsilon} - \frac{y(k)}{y(k)+\varepsilon}\bigg) \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Bigg] \\ &+ (1 - p_{\mathcal{S}^\star}) \left|\mathbb{E}\Bigg[\sum_{l=M+1}^{L} \sigma_l \sum_{k=1}^{d} \mathbb{1}(x_{L+1} = e_k) \bigg(\frac{\mathbb{1}(x_l = e_k)}{y(k)+\varepsilon} - \frac{y(k)}{y(k)+\varepsilon}\bigg) \prod_{h\in\mathcal{S}^\star} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Bigg]\right| \\ &\leq 4(1 - p_{\mathcal{S}^\star}(t_1)) = 2\Delta_1, \qquad \text{where} \quad \Delta_1 := (1 - p_{\mathcal{S}^\star}(t_1)).\end{aligned}$$

In summary, the difference between f_0 and f_1 is controlled by the convergence results from Stage I.

- Our second step is to characterize the approximation error incurred by the difference between the ideal attention scores and the actual attention scores in the second attention layer. Let us define $s_l^\star = \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})$ as the ideal attention score for the second attention layer. We invoke Lemma F.1 to have for all $l \in [L]$,

$$|s_l - s_l^\star| \le \Delta_1 + \Delta_2, \quad \text{where} \quad \Delta_2 := 1 - \prod_{h\in\mathcal{S}^\star} (\sigma_{-h}^{(h)}(t_2))^2.$$

Corresponding to $\{s_l^\star\}_{l=M+1}^L$, we define

$$\sigma_l^\star := \frac{\exp\left(a \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})\right)}{\sum_{l'=M+1}^L \exp\left(a\prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l'-h} = x_{L+1-h})\right)}, \quad y^\star(k) := \sum_{l=M+1}^L \sigma_l^\star \, \mathbb{1}(x_l = e_k), \quad \forall k \in [d].$$

In the vector form, we have $y^\star = \sum_{l=M+1}^L \sigma_l^\star x_l$. Leveraging the above approximations, we define an approximation of f_1 as

$$f_2 := \mathbb{E}\Bigg[\sum_{l=M+1}^L \sigma_l^\star \sum_{k=1}^d \left(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y^\star(k) + \varepsilon} - \frac{y^\star(k)\,\mathbb{1}(x_{L+1} = e_k)}{y^\star(k) + \varepsilon}\right) \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})\Bigg].$$

Applying Lemma F.9, it holds that

$$|f_1 - f_2| \le 12 \cdot (1 + a(t)\cdot \varepsilon^{-1}) \cdot (\Delta_1 + \Delta_2)$$

In summary, this error terms captures the difference between the ideal weights and the actual weights obtained by gradient flow at the end of Stage II.

- Note that $y^\star(k)$ is also random due to the randomness in $\sigma_l^\star$, and as L is sufficiently large, we want to replace $y^\star(k)$ with its population counterpart. Let $z \in \mathcal{X}$ and $Z = (z_{-M}, \ldots, z_{-1}) \in \mathcal{X}^M$ be two random variables and we define similarly for $x \in \mathcal{X}$ and $X = (x_{-M}, \ldots, x_{-1}) \in \mathcal{X}^M$. To this end, we define a reweighed distribution

$$\widetilde{\mu}^\pi(z, Z \,|\, X_{-\mathcal{S}^\star}) = \frac{\mu^\pi(z, Z)\exp\left(a\prod_{h\in\mathcal{S}^\star}\mathbb{1}(z_{-h} = x_{-h})\right)}{\sum_{z', Z'} \mu^\pi(z', Z')\exp\left(a\prod_{h\in\mathcal{S}^\star}\mathbb{1}(z'_{-h} = x_{-h})\right)}, \tag{E.27}$$

where μ^π is the stationary distribution of the Markov chain over a window of size $M+1$. This can be viewed as a reweighting of the stationary distribution over (z, Z) by an exponential term that depends on the sequence $X_{-\mathcal{S}^\star}$. We use $\widetilde{\mu}^\pi(z = e_k \,|\, X_{L+1-\mathcal{S}^\star})$ to replace $y^\star(k)$ and define f_3 as

$$f_3 := \mathbb{E}\Bigg[\sum_{l=M+1}^L \sigma_l^\star \sum_{k=1}^d \left(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\widetilde{\mu}^\pi(z = e_k \,|\, X_{L+1-\mathcal{S}^\star})} - \mathbb{1}(x_{L+1} = e_k)\right) \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})\Bigg].$$

One can immediately draw a connection to Lemma F.4 as both targets characterize the gap between the empirical and population distributions. The only difference is that this time we have the distribution reweighed by some exponential term. For completeness, we provide the approximation result in Lemma F.10, which bounds the difference between f_2 and f_3 as

$$|f_2 - f_3| \le \frac{8(1-\lambda)^{-1/2}(D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1)^{1/4} + 8\sqrt{M}}{L^{1/2}\cdot\gamma^{|\mathcal{S}^\star|+1}} + \frac{2d\varepsilon}{\gamma} \lesssim \frac{\sqrt{M} + d}{L^{1/2}(1-\lambda)^{1/2}\gamma^{|\mathcal{S}^\star|+1+r_n/4}}.$$

where $\mu_0(\cdot)$ is the initial distribution for the first r_n tokens in the Markov chain. Here and in the sequel, we simply use $D_{\chi^2}(\mu_0 \,\|\, \mu^\pi)$ to denote $D_{\chi^2}(\mu_0(X_{1:r_n} = \cdot) \,\|\, \mu^\pi(X_{1:r_n} = \cdot))$ when it is clear from the context. In the last inequality, we use the fact that $D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) \le \gamma^{-r_n}$ by (E.6) and the condition $\varepsilon = L^{-1/2}$.

- Note that in the expression of f_3, each $\sigma_l^\star$ still implicitly depends on the actual value of the sequence X. Since L is large and the Markov chain is well-mixed, we can approximate

$\sum_{l=M+1}^{L} \sigma_l^\star \, \mathbb{1}((x_l, X_{l-\mathcal{S}^\star}) = (\cdot,\cdot))$ by $\widetilde{\mu}^\pi(\cdot,\cdot \mid X_{L+1-\mathcal{S}^\star})$. This gives rise to the following approximation of f_3:

$$
\begin{aligned}
f_4 &:= \mathbb{E}_{\pi, X, Z \sim \widetilde{\mu}^\pi(\cdot \mid X_{L+1-\mathcal{S}^\star})}\left[\sum_{k=1}^{d}\left(\frac{\mathbb{1}(x_{L+1} = z = e_k)}{\widetilde{\mu}^\pi(z = e_k \mid X_{L+1-\mathcal{S}^\star})} - \mathbb{1}(x_{L+1} = e_k)\right)\cdot \mathbb{1}(Z_{l-\mathcal{S}^\star} = x_{L+1-\mathcal{S}^\star})\right] \\
&= \mathbb{E}_{\pi, X, Z \sim \widetilde{\mu}^\pi(\cdot \mid X_{L+1-\mathcal{S}^\star})}\left[\sum_{k=1}^{d}\frac{\mu^\pi(x = e_k \mid X_{-\mathcal{S}^\star})\widetilde{\mu}^\pi(z = e_k, Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star} \mid X_{-\mathcal{S}^\star})}{\widetilde{\mu}^\pi(z = e_k \mid X_{-\mathcal{S}^\star})} \right. \\
&\qquad \left. - \widetilde{\mu}^\pi(Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star} \mid X_{-\mathcal{S}^\star})\right]
\end{aligned}
$$

Applying Lemma F.11 yields

$$
|f_3 - f_4| \le \sup_{\pi \in \mathrm{supp}(\mathcal{P})} \frac{8\gamma^{-1}(1-\lambda)^{-1/2}(D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1)^{1/4} + 16\gamma^{-1}\sqrt{M}}{L^{1/2}\cdot \gamma^{|\mathcal{S}^\star|+1}} \lesssim \frac{\sqrt{M} + d}{L^{1/2}(1-\lambda)^{1/2}\gamma^{|\mathcal{S}^\star|+2+r_n/4}},
$$

where we use the fact that $D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) \le \gamma^{-r_n}$ by (E.6).

- Let $(z, Z) \sim \widetilde{\mu}^\pi(\cdot \mid X_{L+1-\mathcal{S}^\star})$. Since L is large, the distribution of $(x_{L+1}, X_{L+1-\mathcal{S}^\star})$ is close to the stationary distribution μ^π. Thus, we introduce the following approximation of f_4:

$$
\begin{aligned}
f_5 &:= \mathbb{E}_{\pi, (x, X_{-\mathcal{S}^\star}) \sim \mu^\pi, (z,Z) \sim \widetilde{\mu}^\pi(\cdot \mid X_{-\mathcal{S}^\star})}\left[\sum_{k=1}^{d}\left(\frac{\mathbb{1}(x = z = e_k)}{\widetilde{\mu}^\pi(e_k \mid X_{-\mathcal{S}^\star})} - \mathbb{1}(x = e_k)\right)\prod_{h \in \mathcal{S}^\star}\mathbb{1}(z_{-h} = x_{-h})\right] \\
&= \mathbb{E}_{\pi, (x, X_{-\mathcal{S}^\star}) \sim \mu^\pi}\left[\sum_{k=1}^{d}\frac{\mu^\pi(x = e_k \mid X_{-\mathcal{S}^\star})\widetilde{\mu}^\pi(z = e_k, Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star} \mid X_{-\mathcal{S}^\star})}{\widetilde{\mu}^\pi(z = e_k \mid X_{-\mathcal{S}^\star})} \right. \\
&\qquad \left. - \widetilde{\mu}^\pi(Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star} \mid X_{-\mathcal{S}^\star})\right].
\end{aligned} \tag{E.28}
$$

Note that

$$
\begin{aligned}
&\left|\sum_{k=1}^{d}\frac{\mu^\pi(x = e_k \mid X_{-\mathcal{S}^\star})\widetilde{\mu}^\pi(z = e_k, Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star} \mid X_{-\mathcal{S}^\star})}{\widetilde{\mu}^\pi(z = e_k \mid X_{-\mathcal{S}^\star})}\right| \\
&= \left|\sum_{k=1}^{d}\mu^\pi(x = e_k \mid X_{-\mathcal{S}^\star})\cdot\widetilde{\mu}^\pi(Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star} \mid X_{-\mathcal{S}^\star}, z = e_k)\right| \le \left|\sum_{k=1}^{d}\mu^\pi(x = e_k \mid X_{-\mathcal{S}^\star})\right| = 1,
\end{aligned}
$$

and so is $|\widetilde{\mu}^\pi(Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star} \mid X_{-\mathcal{S}^\star})| \le 1$. The difference between f_4 and f_5 is thus bounded by $2\|p^\pi(x_{L+1}, X_{L+1-\mathcal{S}^\star} = \cdot,\cdot) - \mu^\pi(x_{L+1}, X_{L+1-\mathcal{S}^\star} == \cdot)\|_{\mathrm{TV}}$ and by the results in (F.29) of Lemma F.16:

$$
|f_4 - f_5| \le 2\cdot\sup_{\pi\in\mathrm{supp}(\mathcal{P})}\lambda^{L-M}\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1} \lesssim \frac{\lambda^{L-M}}{\gamma^{r_n/2}} \le L^{-1},
$$

where we use $D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) \le \gamma^{-r_n}$ and the condition on L in (E.2).

Collecting all the above approximation steps, we obtain (where we use $\lesssim$ to hide absolute constants)

$$
\begin{aligned}
|f_0 - f_5| &\lesssim \Delta_1 + (1 + a\cdot\varepsilon^{-1})\cdot(\Delta_1 + \Delta_2) + L^{-1} + \frac{\sqrt{M} + d}{L^{1/2}(1-\lambda)^{1/2}\gamma^{|\mathcal{S}^\star|+2+r_n/4}} \\
&\lesssim a\cdot L^{-1/2} + \frac{\sqrt{M} + d}{L^{1/2}(1-\lambda)^{1/2}\gamma^{|\mathcal{S}^\star|+2+r_n/4}}.
\end{aligned}
$$

where the last line holds by moting that with sufficiently large t_1 and t_2 we have $\Delta_1 + \Delta_2 \le L^{-1}$, and $\varepsilon = L^{-1/2}$. Here, express the error in terms of the trainable parameter a and define

$$\xi(a) \asymp \frac{\sqrt{M} + d}{L^{1/2}(1-\lambda)^{1/2}\gamma^{|\mathcal{S}^\star|+2+r_n/4}} + a \cdot L^{-1/2}.$$

In particular, we have for $a = O(\log L)$ that

$$\xi(a) = O\bigg(\frac{\sqrt{M} + d}{L^{1/2}(1-\lambda)^{1/2}\gamma^{|\mathcal{S}^\star|+2+r_n/4}} + \frac{\log L}{L^{1/2}}\bigg). \tag{E.29}$$

In a nutshell, we conclude that when the weight a satisfies $a < O(\log L)$, the dynamics of a can be approximated by

$$\partial_t a = f_5 \pm \xi(a). \tag{E.30}$$

The following proposition helps us reformulate f_5 in a form that facilitates the analysis of the dynamics of a.

Proposition E.4. *The term f_5 can be reformulated as*

$$f_5 = \mathbb{E}_{\pi, X_{-\mathcal{S}^\star} \sim \mu^\pi}\Big[J(X_{-\mathcal{S}^\star}; a, \pi) \cdot e^a \cdot \big(r^\pi(X_{-\mathcal{S}^\star})\big)^3 \cdot \mu^\pi(X_{-\mathcal{S}^\star})\Big],$$

where $r^\pi(X_{-\mathcal{S}^\star}; a) = (1 + \mu^\pi(X_{-\mathcal{S}^\star}) \cdot (e^a - 1))^{-1}$ is the inverse of the normalization factor of $\widetilde{\mu}^\pi$ in (E.27) and

$$J(X_{-\mathcal{S}^\star}; a, \pi) = \sum_{k \in [d]} \frac{(\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) - \mu^\pi(x = e_k))^2}{(1 - r^\pi(X_{-\mathcal{S}^\star}; a)) \cdot \mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) + r^\pi(X_{-\mathcal{S}^\star}; a) \cdot \mu^\pi(x = e_k)}.$$

Proof. See §E.4.1 for the proof. □

Inspired by this form, we define an alternative function $\widetilde{J}(\cdot; r, \pi)$ as

$$\widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi) := \sum_{k \in [d]} \frac{(\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) - \mu^\pi(x = e_k))^2}{(1 - r) \cdot \mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) + r \cdot \mu^\pi(x = e_k)}, \quad r \in [0, 1] \tag{E.31}$$

where we replace $r^\pi(X_{-\mathcal{S}^\star}; a)$ by a parameter $r \in [0, 1]$. As exactly calculating the inverse normalization factor $r^\pi(X_{-\mathcal{S}^\star}; a)$ is intractable, we instead seek to find an upper and lower bound for $r^\pi(X_{-\mathcal{S}^\star}; a)$ and plug them into $\widetilde{J}(\cdot; r, \pi)$ to bound f_5 Suppose that $r^\pi(X_{-\mathcal{S}^\star}; a)$ enjoys the following parameter-dependent upper and lower bounds:

$$r_-(a) \le r^\pi(X_{-\mathcal{S}^\star}; a) \le r_+(a), \quad \forall X_{-\mathcal{S}^\star} \in \mathcal{X}^{|\mathcal{S}^\star|}, \quad \forall \pi \in \operatorname{supp}(\mathcal{P}).$$

Thus, an upper and lower bound to $J(X_{-\mathcal{S}^\star}; a, \pi)$ can be given by

$$\inf_{r \in [r_-(a), r_+(a)]} \widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi) \le J(X_{-\mathcal{S}^\star}; a, \pi) \le \sup_{r \in [r_-(a), r_+(a)]} \widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi).$$

In order to effectively tackle these bounds, we then study the properties of $\widetilde{J}(\cdot; r, \pi)$ next.

Proposition E.5. *Define*

$$D_+(X_{-\mathcal{S}^\star}, \pi) = \max\big\{D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})), D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot))\big\}.$$

The function $\widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi)$ with $r \in [0, 1]$ defined in (E.31) satisfies the following properties:

1. *$\widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi)$ is convex in r.*
2. *$\widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi) \le D_+(X_{-\mathcal{S}^\star}, \pi)$.*
3. *$\widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi)$ is Lipschitz continuous in r with Lipschitz constant $\gamma^{-1} D_+(X_{-\mathcal{S}^\star}, \pi)$.*

Proof. See §E.4.1 for the proof. □

Upper and Lower Bounding $J(X_{-\mathcal{S}^\star}; a, \pi)$. Previously, we show via a reformulation of f_5 that it suffices to bound $J(X_{-\mathcal{S}^\star}; a, \pi)$. In the sequel, we let

$$D_+(X_{-\mathcal{S}^\star}, \pi) = \max\left\{D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})), D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot))\right\},$$

$$\rho = \max\left\{\max_{X_{-\mathcal{S}^\star},\pi} \frac{D_+(X_{-\mathcal{S}^\star}, \pi)}{D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}))}, \max_{X_{-\mathcal{S}^\star},\pi} \frac{D_+(X_{-\mathcal{S}^\star}, \pi)}{D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot))}\right\}.$$

It can be noticed that

$$\rho \le \max\left\{\max_{X_{-\mathcal{S}^\star},\pi} \frac{D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}))}{D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot))}, \max_{X_{-\mathcal{S}^\star},\pi} \frac{D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot))}{D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}))}\right\}$$
$$\le \max\left\{\max_{X_{-\mathcal{S}^\star},\pi} \frac{\mu^\pi(\cdot)}{\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})}, \max_{X_{-\mathcal{S}^\star},\pi} \frac{\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})}{\mu^\pi(\cdot)}\right\} \le \gamma^{-1},$$

where the second inequality follows from noting that the χ^2-divergence defined as $D_{\chi^2}(\mu \,\|\, \nu) = \sum_x (\mu(x) - \nu(x))^2/\nu(x)$, and $D_{\chi^2}(\mu \,\|\, \nu)/D_{\chi^2}(\nu \,\|\, \mu) \le \sup_x \mu(x)/\nu(x)$.

Apparently, $r^\pi(X_{-\mathcal{S}^\star}; a)$ is a function of a and enjoys the following parameter-dependent upper and lower bounds:

$$r_+(a) = (1 + \min_{X_{-\mathcal{S}^\star},\pi} \mu^\pi(X_{-\mathcal{S}^\star})(e^a - 1))^{-1},$$
$$r_-(a) = (1 + \max_{X_{-\mathcal{S}^\star},\pi} \mu^\pi(X_{-\mathcal{S}^\star})(e^a - 1))^{-1}.$$

If a is small, we see that both $r_+(a)$ and $r_-(a)$ are close to 1, and we directly have

$$r_-(a) \le r^\pi(X_{-\mathcal{S}^\star}; a) \le 1, \quad \text{where} \quad 1 - \max_{X_{-\mathcal{S}^\star},\pi} \mu^\pi(X_{-\mathcal{S}^\star})(e^a - 1) \le r_-(a) < 1.$$

This suggests an upper bound of $J(X_{-\mathcal{S}^\star}; a, \pi)$ as

$$J(X_{-\mathcal{S}^\star}; a, \pi) \le \sup_{r \in [r_-(a), 1]} \widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi) \le \widetilde{J}(X_{-\mathcal{S}^\star}; 1, \pi) + \gamma^{-1} \cdot D_+(X_{-\mathcal{S}^\star}, \pi) \cdot (1 - r_-(a))$$
$$\le D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot)) + \gamma^{-1} \cdot D_+(X_{-\mathcal{S}^\star}, \pi) \cdot \max_{X_{-\mathcal{S}^\star},\pi} \mu^\pi(X_{-\mathcal{S}^\star}) \cdot (e^a - 1)$$
$$\le D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot)) \cdot \left(1 + \gamma^{-2} \cdot \max_{X_{-\mathcal{S}^\star},\pi} \mu^\pi(X_{-\mathcal{S}^\star}) \cdot (e^a - 1)\right),$$

where the second line follows from the Lipschitz continuity property, and the last line holds because the ratio $D_+(X_{-\mathcal{S}^\star}, \pi)/D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot))$ is upper bounded by ρ, and further by γ^{-1}. A similar lower bound can be obtained by changing the sign of $\gamma^{-2} \cdot \max_{X_{-\mathcal{S}^\star},\pi} \mu^\pi(X_{-\mathcal{S}^\star}) \cdot (e^a - 1)$. Hence, we h

$$J(X_{-\mathcal{S}^\star}; a, \pi) = D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot)) \cdot \left(1 \pm \gamma^{-2} \cdot \max_{X_{-\mathcal{S}^\star},\pi} \mu^\pi(X_{-\mathcal{S}^\star}) \cdot (e^a - 1)\right). \tag{E.32}$$

On the other hand, when a becomes large, we have both $r_+(a)$ and $r_-(a)$ close to 0, and we have

$$0 \le r^\pi(X_{-\mathcal{S}^\star}; a) \le r_+(a), \quad \text{where} \quad 0 < r_+(a) \le \frac{1}{\min_{X_{-\mathcal{S}^\star},\pi} \mu^\pi(X_{-\mathcal{S}^\star})(e^a - 1)}.$$

In a similar fashion, we have the following upper bound:

$$J(X_{-\mathcal{S}^\star}; a, \pi) \le \sup_{r \in [0, r_+(a)]} \widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi) \le \widetilde{J}(X_{-\mathcal{S}^\star}; 0, \pi) + \gamma^{-1} \cdot D_+(X_{-\mathcal{S}^\star}, \pi) \cdot r_+(a)$$
$$= D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})) + \gamma^{-1} \cdot \frac{D_+(X_{-\mathcal{S}^\star}, \pi)}{\min_{X_{-\mathcal{S}^\star},\pi} \mu^\pi(X_{-\mathcal{S}^\star})(e^a - 1)}$$
$$\le D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})) \cdot \left(1 + \frac{\gamma^{-2}}{\min_{X_{-\mathcal{S}^\star},\pi} \mu^\pi(X_{-\mathcal{S}^\star})(e^a - 1)}\right).$$

We can similarly obtain a lower bound by changing the sign of the second term inside the bracket. Hence, we have

$$J(X_{-\mathcal{S}^\star}; a, \pi) = D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})) \cdot \left(1 \pm \frac{\gamma^{-2}}{\min_{X_{-\mathcal{S}^\star},\pi} \mu^\pi(X_{-\mathcal{S}^\star})(e^a - 1)}\right). \tag{E.33}$$

https://doi.org/10.52202/079017-2127

Divergence of a. Recall that we have shown the dynamics of a in (E.30), where $\xi(a)$ is negligible when L goes to infinity. Thus, when L is sufficiently large, we see by the nonnegativity of f_5 that $a(t)$ continues to increase as t increases until it reaches a point where f_5 no longer dominates the approximation error. To characterize the regime where $f_5 \geq \xi(a)$, we first note that for $a \leq \log L$ it holds by (E.29) that

$$\xi(a) = O(L^{-1/2} \log L) \approx L^{-1/2},$$

where $\approx$ hides logarithmic factors. For f_5, we recall from Proposition E.4 that

$$f_5 = \mathbb{E}_{\pi, X_{-\mathcal{S}^\star} \sim \mu^\pi} \left[\frac{J(X_{-\mathcal{S}^\star}) \cdot e^a}{\left(1 + \mu^\pi(X_{-\mathcal{S}^\star}) \cdot (e^a - 1)\right)^3} \cdot \mu^\pi(X_{-\mathcal{S}^\star}) \right],$$

where for small a we have $f_5 = \Omega(1)$ and for large a we have $f_5 = \Omega(e^{-2a})$. Thus, $e^{-2a} \geq L^{-1/2}$ gives the condition for f_5 to dominate the approximation error, which gives $a = O(\log L)$. In the sequel, we consider the dynamics for $a \leq (\log L)/8$ and give a more rigorous analysis.

We use the notation $x = o(1)$ to denote that a term is much smaller than 1, for example, $(\log \log L)^{-1} = o(1)$. For any x_0 and δ, we write $x = x_0 \pm \delta$ to indicate that x is bounded within $[x_0 - \delta, x_0 + \delta]$. In the following, we assume there exists δ satisfying $\delta \leq \gamma^2/4 \wedge 1/8$ and

$$\delta \cdot \mathbb{E}_{\pi \sim \mathcal{P}} \left[\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}\left(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot)\right) \cdot \left(\mu^\pi(X_{-\mathcal{S}^\star})\right)^2 \right] \geq \xi(\log L),$$

$$\delta \cdot \mathbb{E}_{\pi \sim \mathcal{P}} \left[\sum_{X_{-\mathcal{S}^\star}} \frac{D_{\chi^2}\left(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})\right) \cdot L^{-1/4}}{\mu^\pi(X_{-\mathcal{S}^\star})} \right] \geq \xi(\log L).$$

Note that

$$\xi(\log L) \leq O\left(\frac{\sqrt{M} + d}{L^{1/2} (1 - \lambda)^{1/2} \gamma^{|\mathcal{S}^\star| + 2 + r_n/4}} + \frac{\log L}{L^{1/2}} \right).$$

By additionally noting that $\mu^\pi(X_{-\mathcal{S}^\star}) \geq \gamma^{|\mathcal{S}^\star|}$ thanks to the lower bound of the transition probability, we are able to find such a δ if we have

$$\frac{L}{(\log L)^4} \geq \Omega\left(\frac{1}{\kappa^4 \gamma^{8 + 2|\mathcal{S}^\star|}} \cdot \left(\frac{\sqrt{M} + d}{(1 - \lambda)^{1/2} \gamma^{|\mathcal{S}^\star| + 2 + r_n/4}} \right)^4 \right),$$

where κ is defined as

$$\kappa := \mathbb{E}\left[D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})) \right] \wedge \mathbb{E}\left[D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot)) \right] \wedge 1,$$

and $\Omega(\cdot)$ only hides universal constants. Note that this is already guaranteed by the condition on L in (E.2). In particular, we can just take $\delta = \gamma^2/4 \wedge 1/8$ in the following analysis.

Small a. Consider the case where a is small in the sense that $\mu^\pi(X_{-\mathcal{S}^\star}) \cdot (e^a - 1) \leq \delta$ for any $X_{-\mathcal{S}^\star}$ and $\pi \in \operatorname{supp}(\mathcal{P})$. In fact, one can directly deduce from our previous results that $1 - \delta \leq r_-(a) \leq r^\pi(X_{-\mathcal{S}^\star}; a) < 1$ and

$$1 - 3\delta \leq (r^\pi(X_{-\mathcal{S}^\star}; a))^3 \leq 1.$$

For $J(X_{-\mathcal{S}^\star}; a, \pi)$, we combine the condition that $\mu^\pi(X_{-\mathcal{S}^\star}) \cdot (e^a - 1) \leq \delta$ with (E.32) to obtain that

$$J(X_{-\mathcal{S}^\star}; a, \pi) = \left(1 \pm \gamma^{-2}\delta\right) \cdot D_{\chi^2}\left(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot)\right), \quad \text{where} \quad \gamma^{-2}\delta \leq 1/4.$$

Combining the above two results with Proposition E.4, we have

$$\begin{aligned} f_5 &= \mathbb{E}_{\pi, X_{-\mathcal{S}^\star} \sim \mu^\pi} \left[J(X_{-\mathcal{S}^\star}; a, \pi) \cdot e^a \cdot \left(r^\pi(X_{-\mathcal{S}^\star})\right)^3 \cdot \mu^\pi(X_{-\mathcal{S}^\star}) \right] \\ &= \left(1 \pm (\gamma^{-2} + 3)\delta\right) \cdot \mathbb{E}_{\pi \sim \mathcal{P}} \left[\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}\left(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot)\right) \cdot \mu^\pi(X_{-\mathcal{S}^\star})^2 \right] \cdot e^a. \end{aligned}$$

https://doi.org/10.52202/079017-2127

Also, the noise term $\xi + \psi(a)$ is upper bounded by

$$\begin{aligned}\xi + \psi(\log L) &\le \delta \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\bigg[\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}\big(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot)\big) \cdot \mu^\pi(X_{-\mathcal{S}^\star})^2\bigg] \\ &\le \delta \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\bigg[\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}\big(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot)\big) \cdot \mu^\pi(X_{-\mathcal{S}^\star})^2\bigg] \cdot e^a\end{aligned}$$

by the construction of δ. Combining all the above results, we have the dynamics of a as

$$\partial_t a = \big(1 \pm (\gamma^{-2} + 4)\delta\big) \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\bigg[\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot)) \cdot \mu^\pi(X_{-\mathcal{S}^\star})^2\bigg] \cdot e^a.$$

A simple reformulation gives

$$-\partial_t e^{-a} = \big(1 \pm (\gamma^{-2} + 4)\delta\big) \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\bigg[\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot)) \cdot \mu^\pi(X_{-\mathcal{S}^\star})^2\bigg],$$

which implies that for small a, the growth follows

$$a(t) \le -\log\bigg(e^{-a(0)} - (1 + (\gamma^{-2} + 4)\delta) \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\bigg[\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot)) \cdot \mu^\pi(X_{-\mathcal{S}^\star})^2\bigg] \cdot t\bigg),$$

$$a(t) \ge -\log\bigg(e^{-a(0)} - (1 - (\gamma^{-2} + 4)\delta) \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\bigg[\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot))\mu^\pi(X_{-\mathcal{S}^\star})^2\bigg] \cdot t\bigg).$$

Therefore, in the beginning, $a(t)$ grows super exponentially fast.

Large a. As a grows large such that $\mu^\pi(X_{-\mathcal{S}^\star})(e^a - 1) \ge \delta^{-1}$ for all $X_{-\mathcal{S}^\star}$ and $\pi \in \mathrm{supp}(\mathcal{P})$, we conclude that $0 < r^\pi(X_{-\mathcal{S}^\star}; a) \le r_+(a) \le \delta$ and

$$\frac{r^\pi(X_{-\mathcal{S}^\star}; a)^3}{(\mu^\pi(X_{-\mathcal{S}^\star})e^a)^{-3}} = \frac{(\mu^\pi(X_{-\mathcal{S}^\star})e^a)^3}{(1 + \mu^\pi(X_{-\mathcal{S}^\star})(e^a - 1))^3} = \bigg(1 - \frac{1 - \mu^\pi(X_{-\mathcal{S}^\star})}{1 + \mu^\pi(X_{-\mathcal{S}^\star})(e^a - 1)}\bigg)^3,$$

which imples that

$$1 - 3\delta \le \frac{r^\pi(X_{-\mathcal{S}^\star}; a)^3}{(\mu^\pi(X_{-\mathcal{S}^\star})e^a)^{-3}} \le 1.$$

For $J(X_{-\mathcal{S}^\star}; a, \pi)$, we combine the condition that $\mu^\pi(X_{-\mathcal{S}^\star}) \cdot (e^a - 1) \ge \delta^{-1}$ with (E.33) to obtain that

$$J(X_{-\mathcal{S}^\star}; a, \pi) = (1 \pm \gamma^{-2}\delta) \cdot D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})), \quad \text{where} \quad \gamma^{-2}\delta \le 1/4.$$

Combining the above two results with Proposition E.4, we have

$$\begin{aligned}f_5 &= \mathbb{E}_{\pi, X_{-\mathcal{S}^\star}\sim\mu^\pi}\Big[J(X_{-\mathcal{S}^\star}; a, \pi) \cdot e^a \cdot \big(r^\pi(X_{-\mathcal{S}^\star})\big)^3 \cdot \mu^\pi(X_{-\mathcal{S}^\star})\Big] \\ &= \big(1 \pm (\gamma^{-2} + 3)\delta\big) \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\bigg[\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})) \cdot \frac{e^{-2a}}{\mu^\pi(X_{-\mathcal{S}^\star})}\bigg].\end{aligned}$$

For the noise term $\xi + \psi(a)$, we have

$$\delta \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\bigg[\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})) \cdot \frac{e^{-2a}}{\mu^\pi(X_{-\mathcal{S}^\star})}\bigg] \ge \xi + \psi(a),$$

which can be verified by the condition on δ as well as the fact that we are only considering $a \le (\log L)/8$. We thus have for the gradient that

$$\partial_t a = (1 \pm (\gamma^{-2} + 4)\delta) \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\bigg[\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})) \cdot \frac{e^{-2a}}{\mu^\pi(X_{-\mathcal{S}^\star})}\bigg].$$

By rearranging the terms, we further have

$$\partial_t e^{2a} = (1 \pm (\gamma^{-2} + 4)\delta) \cdot \mathbb{E}_{\pi \sim \mathcal{P}} \bigg[\sum_{X_{-\mathcal{S}^\star}} \frac{D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}))\cdot}{2\mu^\pi(X_{-\mathcal{S}^\star})} \bigg].$$

Suppose this large a regime starts at t_0 with value $a(t_0)$. Thus, for large a, the growth rate is characterized by

$$a(t) = \frac{1}{2} \log \bigg((1 \pm (\gamma^{-2} + 4)\delta) \cdot \mathbb{E}_{\pi \sim \mathcal{P}} \bigg[\sum_{X_{-\mathcal{S}^\star}} \frac{D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}))}{2\mu^\pi(X_{-\mathcal{S}^\star})} \bigg] \cdot (t - t_0) + e^{2a(t_0)} \bigg),$$

which is logarithmically fast. This step ends until a reaches the value $(\log L)/8$. This concludes the proof. □

E.4.1 Additional Proofs for Stage III

We conclude the proof of Stage III by providing the proof of Proposition E.4 and Proposition E.5.

Proof of Proposition E.4. In this paragraph, we aim to gain more insight in f_5. By the definition of f_5 in (E.28), we can rewrite f_5 as

$$\begin{aligned} f_5 &= \mathbb{E}_{\pi, X_{-\mathcal{S}^\star} \sim \mu^\pi} \bigg[\bigg(\sum_{k=1}^{d} \frac{\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star})^2}{\widetilde{\mu}^\pi(z = e_k \,|\, X_{-\mathcal{S}^\star})} - 1 \bigg) \cdot \widetilde{\mu}^\pi(Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star}) \bigg] \\ &= \mathbb{E}_{\pi, X_{-\mathcal{S}^\star} \sim \mu^\pi} \bigg[\sum_{k=1}^{d} \bigg(\frac{\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star})}{\widetilde{\mu}^\pi(z = e_k \,|\, X_{-\mathcal{S}^\star})} - 1 \bigg)^2 \cdot \widetilde{\mu}^\pi(z = e_k \,|\, X_{-\mathcal{S}^\star}) \cdot \widetilde{\mu}^\pi(Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star}) \bigg], \end{aligned}$$

where in the last step, we use the simple fact

$$\sum_x \frac{p(X = x \,|\, Y)^2}{q(X = x \,|\, Y)} - 1 = \sum_x \bigg(\frac{p(X = x \,|\, Y)}{q(X = x \,|\, Y)} - 1 \bigg)^2 \cdot q(X = x \,|\, Y).$$

In the definition of f_5, the key quantity we aim to understand is the reweighted distribution $\widetilde{\mu}^\pi(z, Z \,|\, X_{-\mathcal{S}^\star})$. For the readers' convenience, we copy the definition of the reweighted distribution here:

$$\widetilde{\mu}^\pi(z, Z \,|\, X_{-\mathcal{S}^\star}) = \frac{\mu^\pi(z, Z) \exp\big(a \prod_{h \in \mathcal{S}^\star} \mathbb{1}(z_{-h} = x_{-h})\big)}{\sum_{z', Z'} \mu^\pi(z', Z') \exp\big(a \prod_{h \in \mathcal{S}^\star} \mathbb{1}(z'_{-h} = x_{-h})\big)}, \tag{E.34}$$

A key observation is that the reweighting only depends on the value of $Z_{-\mathcal{S}^\star}$. Let $\bar{\mathcal{S}}^\star = [M] \backslash \mathcal{S}^\star$ and denote by $Z_{-\bar{\mathcal{S}}^\star} = (z_{-h})_{h \in \bar{\mathcal{S}}^\star}$. Following the above observation, we can additionally condition on $Z_{-\mathcal{S}^\star}$ and conclude that

$$\begin{aligned} \widetilde{\mu}^\pi(z, Z_{-\bar{\mathcal{S}}^\star} \,|\, Z_{-\mathcal{S}^\star}, X_{-\mathcal{S}^\star}) &= \frac{\widetilde{\mu}^\pi(z, Z_{-\bar{\mathcal{S}}^\star}, Z_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star})}{\sum_{z', Z'_{-\bar{\mathcal{S}}^\star}} \widetilde{\mu}^\pi(z', Z'_{-\bar{\mathcal{S}}^\star}, Z_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star})} \\ &= \frac{\mu^\pi(z, Z_{-\bar{\mathcal{S}}^\star}, Z_{-\mathcal{S}^\star}) \exp\big(a \prod_{h \in \mathcal{S}^\star} \mathbb{1}(z_{-h} = x_{-h})\big)}{\sum_{z', Z'_{-\bar{\mathcal{S}}^\star}} \mu^\pi(z', Z'_{-\bar{\mathcal{S}}^\star}, Z_{-\mathcal{S}^\star}) \exp\big(a \prod_{h \in \mathcal{S}^\star} \mathbb{1}(z_{-h} = x_{-h})\big)} \\ &= \frac{\mu^\pi(z, Z_{-\bar{\mathcal{S}}^\star}, Z_{-\mathcal{S}^\star})}{\mu^\pi(Z_{-\mathcal{S}^\star})} = \mu^\pi(z, Z_{-\bar{\mathcal{S}}^\star} \,|\, Z_{-\mathcal{S}^\star}), \end{aligned}$$

as when fixing $Z_{-\mathcal{S}^\star}$, the exponential reweighting terms cancel out in the numerator and denominator in the definition (E.34). Using the above identity, we are able to expand $\widetilde{\mu}^\pi(z \,|\, X_{-\mathcal{S}^\star})$ as

$$\begin{aligned} \widetilde{\mu}^\pi(z \,|\, X_{-\mathcal{S}^\star}) &= \sum_{Z_{-\mathcal{S}^\star}} \mu^\pi(z \,|\, Z_{-\mathcal{S}^\star}) \cdot \widetilde{\mu}^\pi(Z_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star}) \\ &= \sum_{Z_{-\mathcal{S}^\star}} \mu^\pi(z \,|\, Z_{-\mathcal{S}^\star}) \cdot \frac{\mu^\pi(Z_{-\mathcal{S}^\star}) + \mu^\pi(X_{-\mathcal{S}^\star})(e^a - 1) \cdot \mathbb{1}(Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star})}{1 + \mu^\pi(X_{-\mathcal{S}^\star})(e^a - 1)} \\ &= \frac{\mu^\pi(z) + \mu^\pi(x = z \,|\, X_{-\mathcal{S}^\star}) \cdot \mu^\pi(X_{-\mathcal{S}^\star}) \cdot (e^a - 1)}{1 + \mu^\pi(X_{-\mathcal{S}^\star}) \cdot (e^a - 1)}. \end{aligned}$$

https://doi.org/10.52202/079017-2127

where the second equality follows from the fact that the reweighing term in $\widetilde{\mu}^\pi$ lifts the likelihood of $Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star}$ by a factor of e^a relative to the base distribution $\mu^\pi(Z_{-\mathcal{S}^\star})$, and the denominator is just the normalization constant. In the sequel, we let $r^\pi(X_{-\mathcal{S}^\star}; a) = (1 + \mu^\pi(X_{-\mathcal{S}^\star}) \cdot (e^a - 1))^{-1}$ be the inverse of the normalization constant. We then have

$$\widetilde{\mu}^\pi(z \,|\, X_{-\mathcal{S}^\star}) = r^\pi(X_{-\mathcal{S}^\star}; a) \cdot \mu^\pi(z) + (1 - r^\pi(X_{-\mathcal{S}^\star}; a)) \cdot \mu^\pi(x = z \,|\, X_{-\mathcal{S}^\star}). \tag{E.35}$$

On the other hand, by definition of $\widetilde{\mu}^\pi$ in (E.34), we directly have

$$\begin{aligned}\widetilde{\mu}^\pi(Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star}) &= \frac{\mu^\pi(X_{-\mathcal{S}^\star})e^a}{\sum_{Z'_{-\mathcal{S}^\star}} \mu^\pi(Z'_{-\mathcal{S}^\star}) \exp\big(a \prod_{h \in \mathcal{S}^\star} \mathbb{1}(z'_{-h} = x_{-h})\big)} \\ &= e^a r^\pi(X_{-\mathcal{S}^\star}; a) \cdot \mu^\pi(X_{-\mathcal{S}^\star}). \end{aligned} \tag{E.36}$$

Combining both (E.35) and (E.36) we have for f_5 that

$$\begin{aligned} f_5 &= \mathbb{E}_{\pi, X_{-\mathcal{S}^\star} \sim \mu^\pi}\Bigg[\sum_{k \in [d]} \left(\frac{\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star})}{r^\pi(X_{-\mathcal{S}^\star}; a) \cdot \mu^\pi(x = e_k) + (1 - r^\pi(X_{-\mathcal{S}^\star}; a)) \cdot \mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star})} - 1\right)^2 \\ &\qquad \cdot \widetilde{\mu}^\pi(z = e_k \,|\, X_{-\mathcal{S}^\star}) \cdot \widetilde{\mu}^\pi(Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star})\Bigg] \\ &= \mathbb{E}_{\pi, X_{-\mathcal{S}^\star} \sim \mu^\pi}\Bigg[\sum_{k \in [d]} \left(\frac{\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) - \mu^\pi(x = e_k)}{r^\pi(X_{-\mathcal{S}^\star}; a) \cdot \mu^\pi(x = e_k) + (1 - r^\pi(X_{-\mathcal{S}^\star}; a)) \cdot \mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star})}\right)^2 \\ &\qquad \cdot \widetilde{\mu}^\pi(z = e_k \,|\, X_{-\mathcal{S}^\star}) \cdot e^a r^\pi(X_{-\mathcal{S}^\star}; a)^3 \cdot \mu^\pi(X_{-\mathcal{S}^\star})\Bigg] \\ &= \mathbb{E}_{\pi, X_{-\mathcal{S}^\star} \sim \mu^\pi}\Bigg[\underbrace{\sum_{k \in [d]} \frac{(\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) - \mu^\pi(x = e_k))^2}{\widetilde{\mu}^\pi(z = e_k \,|\, X_{-\mathcal{S}^\star})}}_{J(X_{-\mathcal{S}^\star}; a, \pi)} \cdot e^a r^\pi(X_{-\mathcal{S}^\star}; a)^3 \cdot \mu^\pi(X_{-\mathcal{S}^\star})\Bigg]. \end{aligned}$$

Here, we note that $J(\cdot; a, \pi)$ is a function depending on both a and π, and can be expanded as

$$J(X_{-\mathcal{S}^\star}; a, \pi) = \sum_{k \in [d]} \frac{(\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) - \mu^\pi(x = e_k))^2}{(1 - r^\pi(X_{-\mathcal{S}^\star}; a))\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) + r^\pi(X_{-\mathcal{S}^\star}; a)\mu^\pi(x = e_k)}.$$

Hence, we complete the proof of Proposition E.4. □

Proof of Proposition E.5. Also, note that $\widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi)$ is convex in r, as by taking the derivative of $\widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi)$ with respect to r, we have

$$\begin{aligned} \frac{\partial \widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi)}{\partial r} &= \sum_{k \in [d]} \frac{(\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) - \mu^\pi(e_k))^3}{((1 - r)\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) + r\mu^\pi(e_k))^2}, \\ \frac{\partial^2 \widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi)}{\partial r^2} &= 2 \cdot \sum_{k \in [d]} \frac{(\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) - \mu^\pi(e_k))^4}{((1 - r)\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) + r\mu^\pi(e_k))^3} \geq 0. \end{aligned}$$

Hence, a naive upper bound for $\widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi)$ is

$$\begin{aligned} \widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi) &\leq \max\{\widetilde{J}(X_{-\mathcal{S}^\star}; 0, \pi), \widetilde{J}(X_{-\mathcal{S}^\star}; 1, \pi)\} \\ &\leq \max\left\{D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})), D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot))\right\}, \end{aligned}$$

where we remind the readers that $D_{\chi^2}(\mu \,\|\, \nu) = \sum_x (\mu(x) - \nu(x))^2/\nu(x)$. Next, we show that $\widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi)$ is Lipschitz continuous in r:

$$
\begin{aligned}
\left|\frac{\partial \widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi)}{\partial r}\right| &= \left|\sum_{k\in[d]} \frac{(\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) - \mu^\pi(e_k))^3}{((1-r)\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) + r\mu^\pi(e_k))^2}\right| \\
&\le \sum_{k\in[d]} \frac{(\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) - \mu^\pi(e_k))^2}{(1-r)\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) + r\mu^\pi(e_k)} \cdot \left|\frac{\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) - \mu^\pi(e_k)}{(1-r)\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) + r\mu^\pi(e_k)}\right| \\
&\le \widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi) \cdot \max\left\{\frac{\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star})}{\mu^\pi(e_k)}, \frac{\mu^\pi(e_k)}{\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star})}\right\} \\
&\le \gamma^{-1} \cdot \max\left\{D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})), D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot))\right\},
\end{aligned}
$$

where we use both the upper bound for $\widetilde{J}(X_{-\mathcal{S}^\star}; r, \pi)$ and the lower bound for the transition kernel that both $\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})$ and $\mu^\pi(\cdot)$ are bounded between γ and 1. □

E.5 Lemma on GIH Approximation Error

Now given the convergence result for the training dynamics, the natural question to ask is how well the learned model implements the GIH mechanism. In the following part of this section, we state the lemma on the approximation error and also present a formal proof of the lemma.

Lemma E.6. *Suppose Assumption 3.5 holds and consider training a transformer model* $\texttt{TF}(M, H, d, D)$ *with* $H = M$. *Let*

$$
\Delta_1 := 1 - p_{\mathcal{S}^\star}(t_1), \quad \Delta_2 := 1 - \prod_{h\in\mathcal{S}^\star} (\sigma_{-h}^{(h)}(t_2))^2,
$$

where t_1 and t_2 are the ending time for the first two stages of the training, respectively. Suppose the error $\Delta_1, \Delta_2 = O(L^{-1})$ after the first two stages' training, and $a = \Theta(\log L)$ after the last stage's training. Let y be the output of the model in (2.5) after the training and $y^\star$ be the output of the GIH mechanism $\texttt{GIH}(x_{1:L}; M, D)$ *defined in Definition 3.2. Then for any $\pi \in \operatorname{supp}(\mathcal{P})$ and with high probability $1 - O(L^{-1})$, it holds that*

$$
\|y^\star - y\|_1 = O(L^{-a/\log L}).
$$

Proof of Lemma E.6. Let $s_l^\star = \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})$ and $s_l = \langle u_{L+1}, u_l\rangle$. Invoking Lemma F.1, the model misspecification error is bounded by

$$
\max_{M<l\le L} |s_l^\star - s_l| \le (\Delta_1 + \Delta_2) := \Delta. \tag{E.37}
$$

We note that the second layer's attention weight a can be as large as $(\log L)/8$. We are comparing the output of the model with the GIH mechanism $\texttt{GIH}(x_{1:L}; M, D)$. Let $N = \sum_{l>M} \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})$. The output of this GIH mechanism is given by

$$
y^\star := \begin{cases} N^{-1} \cdot \sum_{l=M+1}^{L} x_l \cdot \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h}), & \text{if} \quad N \ge 1, \\ (L-M)^{-1} \cdot \sum_{l=M+1}^{L} x_l, & \text{otherwise.} \end{cases}
$$

We define

$$
\sigma_l^\star = \begin{cases} N^{-1} \cdot \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h}), & \text{if} \quad N \ge 1, \\ (L-M)^{-1}, & \text{otherwise,} \end{cases}
$$

with $\sigma^\star = (\sigma_l^\star)_{l>M}$. Since $\|x_l\|_1 = 1$, the ℓ-1 norm of the difference between $y^\star$ and the model's actual output is given by

$$
\|y^\star - y\|_1 \le \|\sigma^\star - \sigma\|_1 .
$$

Let us define the set $\Gamma = \{L \ge l > M : \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h}) = 1\}$ and $\bar{\Gamma} = \{L \ge l > M : \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h}) = 0\}$. Using (E.37), for $l \in \Gamma$, we have $1 \ge s_l \ge s_l^\star - \Delta = 1 - \Delta$ and for $l \in \bar{\Gamma}$, we have $0 \le s_l \le s_l^\star + \Delta = \Delta$. Consider the normalization factor in the softmax function.

$$
\mathcal{Z} := \sum_{l=M+1}^{L} \exp(a \cdot s_l).
$$

By the split of the set Γ and $\bar{\Gamma}$ and noting that $|\Gamma| = N$, the normalization factor is lower and upper bounded by

$$\mathcal{Z} \geq N \exp(a \cdot (1 - \Delta)) + (L - M - N) \cdot =: \mathcal{Z}_-,$$
$$\mathcal{Z} \leq N \exp(a) + (L - M - N) \cdot \exp(a \cdot \Delta) =: \mathcal{Z}_+.$$

Let us consider the event $N \geq 1$ in the following. We then have for $l \in \Gamma$ that

$$\begin{aligned}
|\sigma_l^\star - \sigma_l| &= \left| \frac{\exp(a \cdot s_l)}{\mathcal{Z}} - \frac{1}{N} \right| \leq \left| \frac{\exp(a)}{\mathcal{Z}_-} - \frac{1}{N} \right| \bigvee \left| \frac{\exp(a \cdot (1 - \Delta))}{\mathcal{Z}_+} - \frac{1}{N} \right| \\
&\leq \left| \frac{1}{N \exp(-a\Delta) + (L - M - N) \cdot \exp(-a)} - \frac{1}{N} \right| \\
&\quad \bigvee \left| \frac{\exp(-2a\Delta)}{N \exp(-a\Delta) + (L - M - N) \exp(-a)} - \frac{1}{N} \right| \\
&\leq \frac{N \cdot (1 - \exp(-a\Delta)) + (L - M - N) \cdot \exp(-a)}{(N \exp(-a\Delta) + (L - M - N) \cdot \exp(-a)) \cdot N} \leq \frac{1 - \exp(-a\Delta)}{N \exp(-a\Delta)} + \frac{L \cdot \exp(-a)}{N^2 \exp(-a\Delta)}.
\end{aligned}$$

Note that $a\Delta = o(1)$ due to the assumption that $\Delta = O(L^{-1})$ and $a = o(L)$. The right hand side is upper bounded by $O(a\Delta/N) + O(L\exp(-a)/N^2)$. For $l \in \bar{\Gamma}$, we have

$$|\sigma_l^\star - \sigma_l| = \sigma_l \leq \frac{\exp(a\Delta)}{\mathcal{Z}_-} \leq \frac{\exp(a \cdot (2\Delta - 1))}{N} = O\left(\frac{\exp(-a)}{N}\right).$$

In summary,

$$\begin{aligned}
\|y^\star - y\|_1 &\leq \|\sigma^\star - \sigma\|_1 \leq \sum_{l \in \Gamma} |\sigma_l^\star - \sigma_l| + \sum_{l \in \bar{\Gamma}} \sigma_l \\
&\leq N \cdot O\left(\frac{a\Delta N + L\exp(-a)}{N^2}\right) + L \cdot O\left(\frac{\exp(-a)}{N}\right) \leq O\left(a\Delta + \frac{L\exp(-a)}{N}\right).
\end{aligned} \tag{E.38}$$

The above inequality holds whenever $N \geq 1$. Now we aim to upper bound the probability that $N = 0$. Note that $N = \sum_{l=M+1}^{L} \mathbb{1}(X_{l-\mathcal{S}^\star} = X_{L+1-\mathcal{S}^\star})$. We consider the following second moment:

$$\begin{aligned}
\mathbb{E}\left[\left((L-M)^{-1} \sum_{l=M+1}^{L} \mathbb{1}(X_{l-\mathcal{S}^\star} = E) - \mu^\pi(E)\right)^2\right] &\leq D_{\chi^2}\left((L-M)^{-1} \sum_{l=M+1}^{L} \mathbb{1}(X_{l-\mathcal{S}^\star} = \cdot) \,\Big\|\, \mu^\pi(\cdot)\right) \\
&\lesssim \frac{M}{L(1-\lambda) \cdot \gamma^{|\mathcal{S}^\star|/2}}, \quad \forall E \in \mathcal{X}^{|\mathcal{S}^\star|},
\end{aligned}$$

where the first inequality holds by noting that $D_{\chi^2}(\mu \,\|\, \nu) = \sum_x (\mu(x) - \nu(x))^2/\nu(x)$ and the last inequality holds by Lemma F.18. Therefore, by the Chebyshev's inequality, we have

$$\mathbb{P}\left(\left| L^{-1} \sum_{l=1}^{L} \mathbb{1}(X_{l-\mathcal{S}^\star} = E) - \mu^\pi(E) \right| \geq t\right) \leq \frac{1}{L(1-\lambda) \cdot \gamma^{|\mathcal{S}^\star|} \cdot t^2}.$$

We can take $t = \min_{E \in \mathcal{X}^{|\mathcal{S}^\star|}} \mu^\pi(E)/2$ and by also taking a union bound over $E \in \mathcal{X}^{|\mathcal{S}^\star|}$ (which gives a $d^{|\mathcal{S}^\star|}$ factor), we conclude that with high probability $\widetilde{O}(1 - L^{-1})$ it holds that $N \geq tL = L \cdot \min_{E \in \mathcal{X}^{|\mathcal{S}^\star|}} \mu^\pi(E)/2$. Thus, it follows from (E.38) that with high probability

$$\|y^\star - y\|_1 \leq O\left(a\Delta + \frac{\exp(-a)}{\min_{E \in \mathcal{X}^{|\mathcal{S}^\star|}} \mu^\pi(E)/2}\right) = O\left(L^{-1}\log L + L^{-a/\log L}\right).$$

Hence, we complete the proof of Lemma E.6. □

https://doi.org/10.52202/079017-2127

F Auxiliary Lemmas and Their Proofs

In this appendix, we present the auxiliary lemmas used to derive the approximation of the gradient flow dynamics in the proof of Theorem 3.6, which is presented in the previous appendix. The proofs of these lemmas are presented right below their statements.

F.1 Useful Inequalities

The following lemma provides a bound on the model misspecification error, which is the difference between the model's output and the ideal output.

Lemma F.1 (Model Misspecification Error)**.** *Let u_{L+1} be the output feature after the FFN & Normalization layer. Then, the model misspecification error defined as*

$$\max_{l\in[L]}\left|\langle u_{L+1}, u_l\rangle - \prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\right|$$

is bounded by $\Delta_1+\Delta_2$, where Δ_1 and Δ_2 are the errors after the training of the first and second stages, respectively, and are defined respectively as

$$\Delta_1 := 1-p_{\mathcal{S}^\star}, \qquad \Delta_2 := 1-\prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)})^2.$$

Proof of Lemma F.1. By definition of the output feature u_l after the FFN & Normalization layer:

$$\langle u_{L+1}, u_l\rangle = \sum_{\mathcal{S}\in[H]_{\le D}} p_{\mathcal{S}}\cdot\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle.$$

As each $v_l^{(h)}$ is a convex combination of $X_{\mathcal{M}(l)}$ where $\mathcal{M}(l)=\{l-M,\dots,l-1\}$, $\|v_l^{(h)}\|_2\le 1$. Thus,

$$\begin{aligned}\left|\langle u_{L+1}, u_l\rangle-\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\right| &= \left|\sum_{\mathcal{S}\in[H]_{\le D}} p_{\mathcal{S}}\cdot\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle-\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\right|\\ &\le\left|-(1-p_{\mathcal{S}^\star})\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle+\sum_{\mathcal{S}\in[H]_{\le D}\backslash\{\mathcal{S}^\star\}} p_{\mathcal{S}}\cdot\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\right|\\ &\le\max\left\{1-p_{\mathcal{S}^\star},\sum_{\mathcal{S}\in[H]_{\le D}\backslash\{\mathcal{S}^\star\}} p_{\mathcal{S}}\right\}=1-p_{\mathcal{S}^\star}=:\Delta_1,\end{aligned}$$

where Δ_1 is the error after the training of the first stage. Since $v_l^{(h)}=\sum_{j\in M}\sigma_{-j}^{(h)}x_{l-j}$, we have

$$\begin{aligned}\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle &= \prod_{h\in\mathcal{S}^\star}\left(\sum_{i,j\in[M]^2}\sigma_{-i}^{(h)}\sigma_{-j}^{(h)}\langle x_{l-i}, x_{L+1-j}\rangle\right)\\ &= \sum_{\{(i_h,j_h)\}_{h\in\mathcal{S}^\star}\in[M]^{2|\mathcal{S}^\star|}}\ \prod_{h\in\mathcal{S}^\star}\sigma_{-i_h}^{(h)}\sigma_{-j_h}^{(h)}\,\mathbb{1}(x_{l-i_h}=x_{L+1-j_h}).\end{aligned}$$

Here in the second equality, we exchange the order of summation and product. The last term of the second equality can be understood as follows. We first pick $|\mathcal{S}^\star|$ index pairs $\{(i_h,j_h)\}_{h\in\mathcal{S}^\star}$ arbitrarily, with each $i_h, j_h\in[H]$. Then we evaluate the product $\prod_{h\in\mathcal{S}^\star}\sigma_{-i_h}^{(h)}\sigma_{-j_h}^{(h)}\,\mathbb{1}(x_{l-i_h}=x_{L+1-j_h})$ given these indices. Then we sum over all possible values that $\{(i_h,j_h)\}_{h\in\mathcal{S}^\star}$ can take.

The above equation implies that

$$\begin{aligned}&\left|\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle-\prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)})^2\,\mathbb{1}(x_{l-h}=x_{L+1-h})\right|\\ &\quad=\left|\sum_{\{(i_h,j_h)\}_{h\in\mathcal{S}^\star}\neq\{(h,h)\}_{h\in\mathcal{S}^\star}}\ \prod_{h\in\mathcal{S}^\star}\sigma_{-i_h}^{(h)}\sigma_{-j_h}^{(h)}\,\mathbb{1}(x_{l-i_h}=x_{L+1-j_h})\right|\\ &\quad\le\sum_{\{(i_h,j_h)\}_{h\in\mathcal{S}^\star}\neq\{(h,h)\}_{h\in\mathcal{S}^\star}}\ \prod_{h\in\mathcal{S}^\star}\sigma_{-i_h}^{(h)}\sigma_{-j_h}^{(h)}\le 1-\prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)})^2=:\Delta_2,\end{aligned}\tag{F.1}$$

where the last inequality follows from the fact that

$$\sum_{(i_h,j_h)_{h\in\mathcal{S}^\star}} \prod_{h\in\mathcal{S}^\star} \sigma^{(h)}_{-i_h}\sigma^{(h)}_{-j_h} = \prod_{h\in\mathcal{S}^\star}\Bigg(\sum_{i,j\in[M]^2} \sigma^{(h)}_{-i}\sigma^{(h)}_{-j}\Bigg) = \prod_{h\in\mathcal{S}^\star}\Bigg(\sum_{i\in[M]} \sigma^{(h)}_{-i}\Bigg)^2 = 1.$$

Here the summation sign in the right-hand side of the second equality indicates that in the last line of (F.1) we sum over all possible values that $\{(i_h, j_h)\}_{h\in\mathcal{S}^\star}$ can take, except for the only case where $(i_h, j_h) = (h, h)$ for all $h \in [H]$.

In summary, by triangle inequality, we have shown that

$$\Bigg|\langle u_{L+1}, u_l\rangle - \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})\Bigg| \le \Delta_1 + \Delta_2.$$

The proof is completed. □

Next, in Lemma F.2, we establish a uniform bound for the quantity involved in the gradient.

Lemma F.2. *Let $y(k) = \sum_{l=M+1}^{L} \sigma_l \,\mathbb{1}(x_l = e_k)$ for each $k \in [d]$ where $\sum_{l=M+1}^{L}\sigma_l = 1$ and $\sigma_l \ge 0$ for all $l \in [L]$. Let ε and C be two positive numbers. For any C-bounded function $f : \mathcal{X}^{L+1} \to [-C, C]$, we have*

$$\Bigg|\sum_{l=M+1}^{L} \sigma_l \cdot \sum_{k=1}^{d}\bigg(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon}\bigg)\cdot f(X)\Bigg| \le 2C.$$

Proof of Lemma F.2. By the triangular inequality, we have

$$\begin{aligned}
&\Bigg|\sum_{l=M+1}^{L} \sigma_l \cdot \sum_{k=1}^{d}\bigg(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon}\bigg)\cdot f(X)\Bigg| \\
&\quad\le C\cdot\Bigg|\sum_{k=1}^{d}\sum_{l=M+1}^{L} \sigma_l\cdot\mathbb{1}(x_l = e_k)\cdot\frac{\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon}\Bigg| + C\cdot\Bigg|\sum_{k=1}^{d}\sum_{l=M+1}^{L}\sigma_l\cdot\frac{y(k)\cdot\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon}\Bigg| \\
&\quad= 2C\cdot\Bigg|\sum_{k=1}^{d}\frac{y(k)\cdot\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon}\Bigg| \le 2C,
\end{aligned}$$

where in the equality, we use the definition $y(k) = \sum_{l=M+1}^{L}\sigma_l\,\mathbb{1}(x_l = e_k)$ and $\sum_{l=M+1}^{L}\sigma_l = 1$. Now we conclude the proof of this lemma. □

F.2 Approximation Errors for Dynamics Analysis

Next, Lemma F.3 addresses the approximation error induced by $\sigma_l \approx 1/L$ in the transformer model. The approximation error will be for $g_{0,\mathcal{S}}$ to $g_{1,\mathcal{S}}$ for Stage I and $g_{h,1}$ to $g_{h,2}$ for Stage II.

Lemma F.3. *For the transformer model defined in (2.5) and any bounded function $f : \mathcal{X}^{L+1} \to \mathbb{R}$ such that $\sup_{x\in\mathcal{X}^L}|f(x)| \le C$ for a constant $C > 0$, define two quantities A and B as*

$$\begin{aligned}
A &:= \sum_{l=M+1}^{L}\mathbb{E}_{X|\pi}\Bigg[\sigma_l(as)\cdot\sum_{k\in[d]}\bigg(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon}\bigg)\cdot f(X)\Bigg], \\
B &:= \frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}_{X|\pi}\Bigg[\bigg(\sum_{k\in[d]}\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\bar{y}(k)+\varepsilon} - \frac{\bar{y}(k)\,\mathbb{1}(x_{L+1} = e_k)}{\bar{y}(k)+\varepsilon}\bigg)\cdot f(X)\Bigg],
\end{aligned}$$

where $s = u_{L+1}^\top U_{1:L}^\top$ and $\bar{y} = (L-M)^{-1}\sum_{l=M+1}^{L} x_l$. Then, for all $a \in [0, 1]$ and $\varepsilon \in (0, 1]$, it holds that

$$|A - B| \le \frac{8Cad}{\varepsilon^2}.$$

Proof of Lemma F.3. By triangular inequality, we have

$$
\begin{aligned}
|A-B| \leq \sum_{l=M+1}^{L} \mathbb{E}\Bigg[\sum_{k\in[d]} \Bigg\{ & \left|\sigma_l(a\cdot s) - \frac{1}{L-M}\right| \cdot \left|\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y(k)+\varepsilon}\right| \\
& + \frac{1}{L-M}\left|\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y(k)+\varepsilon} - \frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\overline{y}(k)+\varepsilon}\right| \\
& + \left|\sigma_l(a\cdot s) - \frac{1}{L-M}\right| \cdot \left|\frac{y(k)\,\mathbb{1}(x_{L+1}=e_k)}{y(k)+\varepsilon}\right|, \\
& + \frac{1}{L-M}\left|\frac{y(k)\,\mathbb{1}(x_{L+1}=e_k)}{y(k)+\varepsilon} - \frac{\overline{y}(k)\,\mathbb{1}(x_{L+1}=e_k)}{\overline{y}(k)+\varepsilon}\right| \Bigg\} \cdot f(X)\Bigg].
\end{aligned}
$$

Note that $0 \leq s_l \leq 1$ for all $l = M+1, \ldots, L$ thanks to the layer normalization. Then, for the softmax operation, we have

$$
\frac{1}{1+(L-M-1)\exp(a)} \leq \sigma_l(a\cdot s) \leq \frac{\exp(a)}{L-M-1+\exp(a)},
$$

which implies that

$$
\begin{aligned}
\left|\sigma_l(a\cdot s) - \frac{1}{L-M}\right| &\leq \max\left\{\frac{1}{L-M} - \frac{1}{1+(L-M-1)\exp(a)}, \frac{\exp(a)}{L-M-1+\exp(a)} - \frac{1}{L-M}\right\} \\
&\leq \frac{\exp(a)-1}{L-M-1}. \qquad \text{(F.2)}
\end{aligned}
$$

Since indicator functions are bounded above by 1, we have

$$
\left|\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y(k)+\varepsilon}\right| \leq \frac{1}{\varepsilon}, \qquad \left|\frac{y(k)\,\mathbb{1}(x_{L+1}=e_k)}{y(k)+\varepsilon}\right| \leq \frac{1}{\varepsilon}, \qquad \text{(F.3)}
$$

For the second term, we have

$$
\begin{aligned}
\left|\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y(k)+\varepsilon} - \frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\overline{y}(k)+\varepsilon}\right| \leq \frac{|\overline{y}(k)-y(k)|}{\varepsilon^2} &\leq \frac{\sum_{l=M+1}^{L}|\sigma_l(a\cdot s^\top) - (L-M)^{-1}|}{\varepsilon^2} \\
&\leq \frac{\exp(a)-1}{\varepsilon^2}, \qquad \text{(F.4)}
\end{aligned}
$$

where the last inequality follows from (F.2). Similarly, the following bound can be derived:

$$
\left|\frac{y(k)\,\mathbb{1}(x_{L+1}=e_k)}{y(k)+\varepsilon} - \frac{\overline{y}(k)\,\mathbb{1}(x_{L+1}=e_k)}{\overline{y}(k)+\varepsilon}\right| \leq \frac{\exp(a)-1}{\varepsilon}. \qquad \text{(F.5)}
$$

Combining (F.2), (F.3), (F.4) and (F.5), it holds that

$$
|A-B| \leq \sum_{l=M+1}^{L} \mathbb{E}\left[4\sum_{k\in[d]} \frac{\exp(a)-1}{\varepsilon^2(L-M)} \cdot f(X)\right] \leq \frac{4Cd(\exp(a)-1)}{\varepsilon^2} \leq \frac{8Cad}{\varepsilon^2},
$$

where the last inequality follows from $\exp(x) - 1 \leq 2x$ for $0 \leq x \leq 1$. This concludes the proof of the lemma. □

Lemma F.4 provides the approximation error introduced by $\mu^\pi(e_k) \approx \overline{y}(k)$ in the transformer model.

Lemma F.4. *For the transformer model defined in* (2.5) *and any bounded function $f : \mathcal{X}^L \to \mathbb{R}$ such that $\sup_{x\in\mathcal{X}^L}|f(x)| \leq C$ for a constant $C > 0$, define two quantities A and B as*

$$
A := \frac{1}{L-M}\sum_{l=M+1}^{L} \mathbb{E}_{X|\pi}\left[\left(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\overline{y}(k)+\varepsilon} - \frac{\overline{y}(k)\,\mathbb{1}(x_{L+1}=e_k)}{\overline{y}(k)+\varepsilon}\right) \cdot f(X)\right],
$$

$$
B := \frac{1}{L-M}\sum_{l=M+1}^{L} \mathbb{E}_{X|\pi}\left[\left(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\mu^\pi(e_k)} - 1\right) \cdot f(X)\right],
$$

where $\bar{y} = (L-M)^{-1}\sum_{l=M+1}^{L} x_l$. Under Assumption 3.5, it holds that

$$|A - B| \le 4C \cdot \frac{(1-\lambda)^{-1/2}(D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1)^{1/4} + 2\sqrt{M}}{L^{1/2}\gamma} + C\gamma^{-1}\varepsilon.$$

where $\mu_0(\cdot)$ is the initial distribution over the first r_n tokens $X_{1:r_n}$. Here we let $D_{\chi^2}(\mu_0 \,\|\, \mu^\pi)$ to denote $D_{\chi^2}(\mu_0(X_{1:r_n} = \cdot) \,\|\, \mu^\pi(X_{1:r_n} = \cdot))$, i.e., the χ^2-divergence between μ_n and the distribution over the first r_n tokens under the stationary distribution μ^π.

Proof of Lemma F.4. Let us use $\bar{y}_X(\cdot)$ to remind the readers that $\bar{y}(\cdot)$ is also a function of X. We simplify the expectation $\mathbb{E}_{X|\pi}$ by $\mathbb{E}$ in this proof. By rearranging the terms, we have

$$\begin{aligned}|A - B| &= \bigg|\frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}\bigg[\bigg(\sum_{k\in[d]}\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\bar{y}_X(k) + \varepsilon} - \sum_{k\in[d]}\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\mu^\pi(e_k)} \\ &\qquad - \sum_{k\in[d]}\frac{\bar{y}_X(k)\cdot \mathbb{1}(x_{L+1} = e_k)}{\bar{y}_X(k) + \varepsilon} + 1\bigg)\cdot f(X)\bigg]\bigg| \\ &= \bigg|\frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}\bigg[\bigg(\sum_{k\in[d]}\Big(\frac{\mu^\pi(e_k) - \bar{y}_X(k)}{(\bar{y}_X(k)+\varepsilon)\cdot\mu^\pi(e_k)} - \frac{\varepsilon}{(\bar{y}_X(k)+\varepsilon)\cdot\mu^\pi(e_k)}\Big)\cdot \mathbb{1}(x_{L+1} = x_l = e_k) \\ &\qquad - \sum_{k\in[d]}\frac{\varepsilon\,\mathbb{1}(x_{L+1} = e_k)}{\bar{y}_X(k) + \varepsilon}\bigg)\cdot f(X)\bigg]\bigg|.\end{aligned}$$

Here, we have three terms to control. For the first error term, we define

$$\begin{aligned}\mathrm{err}_1 :=& \bigg|\frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}\bigg[\sum_{k\in[d]}\frac{\mu^\pi(e_k) - \bar{y}_X(k)}{(\bar{y}_X(k)+\varepsilon)\cdot\mu^\pi(e_k)}\cdot \mathbb{1}(x_{L+1} = x_l = e_k)\cdot f(X)\bigg]\bigg| \\ &\le \frac{C}{L-M}\sum_{l=M+1}^{L}\mathbb{E}\bigg[\sum_{k\in[d]}\frac{|\mu^\pi(e_k) - \bar{y}_X(k)|}{(\bar{y}_X(k)+\varepsilon)\cdot\mu^\pi(e_k)}\cdot \mathbb{1}(x_{L+1} = x_l = e_k)\bigg] \\ &\le C\cdot\mathbb{E}\bigg[\sum_{k\in[d]}\frac{|\mu^\pi(e_k) - \bar{y}_X(k)|}{\mu^\pi(e_k)}\cdot \mathbb{1}(x_{L+1} = e_k)\bigg].\end{aligned}$$

The first inequality above holds by noting that $\sup_X |f(X)| \le C$ and the last inequality holds by noting that $\bar{y}_X(e_k) = (L-M)^{-1}\sum_{l=M+1}^{L}\mathbb{1}(x_l = e_k)$. Using Cauchy-Schwarz inequality, we arrive at

$$\begin{aligned}\mathrm{err}_1 &\le C\cdot\bigg(\mathbb{E}\bigg[\sum_{k\in[d]}\Big(\frac{\mu^\pi(e_k) - \bar{y}_X(k)}{\sqrt{\mu^\pi(e_k)}}\Big)^2\bigg]\cdot\mathbb{E}\bigg[\sum_{k\in[d]}\frac{\mathbb{1}(x_{L+1} = e_k)}{\mu^\pi(e_k)}\bigg]\bigg)^{1/2} \\ &\le C\gamma^{-1/2}\cdot\sqrt{\mathbb{E}\left[D_{\chi^2}\left(\bar{y}_X(\cdot) \,\|\, \mu^\pi(x_{L+1} = \cdot)\right)\right]}.\end{aligned}$$

For the second term, we similarly have

$$\begin{aligned}\mathrm{err}_2 &= \bigg|\frac{1}{L-M}\sum_{l=M+1}^{L}\sum_{k\in[d]}\mathbb{E}\left[\frac{\varepsilon}{(\bar{y}_X(k)+\varepsilon)\mu^\pi(e_k)}\cdot \mathbb{1}(x_{L+1} = x_l = e_k)\cdot f(X)\right]\bigg| \\ &\le C\bigg|\sum_{k\in[d]}\mathbb{E}\bigg[\frac{\varepsilon\cdot\mathbb{1}(x_{L+1} = e_k)}{\mu^\pi(e_k)}\bigg]\bigg| \le C\gamma^{-1}\varepsilon.\end{aligned}$$

Lastly, we have the error term

$$\begin{aligned}\mathrm{err}_3 :=& \frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}\bigg[\sum_{k\in[d]}\frac{\varepsilon\,\mathbb{1}(x_{L+1} = e_k)}{\bar{y}_X(k) + \varepsilon}\cdot f(X)\bigg] \le C\cdot\mathbb{E}\bigg[\sum_{k\in[d]}\frac{\varepsilon\,\mathbb{1}(x_{L+1} = e_k)}{\bar{y}_X(k) + \varepsilon}\bigg] \\ &\le C\cdot\bigg|\mathbb{E}\bigg[\sum_{k\in[d]}\frac{\varepsilon\,\mathbb{1}(x_{L+1} = e_k)}{\mu^\pi(e_k) + \varepsilon}\bigg]\bigg| + C\cdot\bigg|\sum_{k\in[d]}\mathbb{E}\bigg[\frac{\varepsilon(\bar{y}_X(k) - \mu^\pi(e_k))\cdot\mathbb{1}(x_{L+1} = e_k)}{(\mu^\pi(e_k)+\varepsilon)(\bar{y}_X(k)+\varepsilon)}\bigg]\bigg|.\end{aligned}$$

https://doi.org/10.52202/079017-2127

Here, the first term is upper bounded by $C\gamma^{-1}\varepsilon$, and for the second term we have by Cauchy-Schwartz that

$$
\begin{aligned}
& C \cdot \bigg| \sum_{k\in[d]} \mathbb{E}\bigg[\frac{\varepsilon(\bar{y}_X(k) - \mu^\pi(e_k)) \cdot \mathbb{1}(x_{L+1} = e_k)}{(\mu^\pi(e_k) + \varepsilon)(\bar{y}_X(k) + \varepsilon)} \bigg] \bigg| \\
&\quad \le C \cdot \sqrt{\mathbb{E}\bigg[\sum_{k\in[d]} \frac{(\bar{y}_X(k) - \mu^\pi(e_k))^2}{\mu^\pi(e_k)} \bigg] \cdot \mathbb{E}\bigg[\sum_{k\in[d]} \frac{\varepsilon^2\, \mathbb{1}(x_{L+1} = e_k)}{(\bar{y}_X(k) + \varepsilon)^2 \mu^\pi(e_k)} \bigg]} \\
&\quad \le C\gamma^{-1/2} \cdot \sqrt{\mathbb{E}_X D_{\chi^2}(\bar{y}_X(\cdot) \,\|\, \mu^\pi(x_{L+1} = \cdot))},
\end{aligned}
$$

which shares a similar upper bound as err_1. Now we invoke Lemma F.18 to conclude that

$$
\begin{aligned}
|A - B| &\le \mathrm{err}_1 + \mathrm{err}_2 + \mathrm{err}_3 \le 2C\gamma^{-1/2} \cdot \sqrt{\mathbb{E}_X D_{\chi^2}(\bar{y}_X(\cdot) \,\|\, \mu^\pi(x_{L+1} = \cdot))} + C\gamma^{-1}\varepsilon \\
&\le 2C\gamma^{-1/2} \bigg(\frac{4(1-\lambda)^{-1}\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1} + 16M}{L \cdot \min_{x_{L+1}} \mu^\pi(x_{L+1})} \bigg)^{1/2} + C\gamma^{-1} \cdot \varepsilon \\
&\le 2C\gamma^{-1} \cdot \frac{2(1-\lambda)^{-1/2}(D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1)^{1/4} + 4\sqrt{M}}{L^{1/2}} + C\gamma^{-1}\varepsilon.
\end{aligned}
$$

Hence, we complete our proof of Lemma F.4. □

Lemma F.5 covers the approximation error due to the mixing property of the Markov chain.

Lemma F.5. *Let $\mathcal{S} \in [H]_{\le D}$ be a fixed set. For any $h \in \mathcal{S}$, let $\widetilde{\sigma}^{(h)}$ and $\sigma^{(h)}$ be two fixed probability distributions over $[M]$. That is, for any $i, j \in [M]$, we have $\widetilde{\sigma}^{(h)}_{-i}, \sigma^{(h)}_{-j} \in [0,1]$, and $\sum_{i=1}^M \widetilde{\sigma}^{(h)}_{-i} = \sum_{j=1}^M \sigma^{(h)}_{-j} = 1$. Given these distributions over $[M]$, we define*

$$
\widetilde{v}^{(h)}_{L+1} := \sum_{i\in[M]} \widetilde{\sigma}^{(h)}_{-i} \cdot x_{L+1-i}, \qquad \text{and} \qquad v^{(h)}_l := \sum_{j\in[M]} \sigma^{(h)}_{-j} \cdot x_{l-j},
$$

where we let $x_l \in \mathcal{X}$ denote the l-th token in the Markov chain for all $l \in [L+1]$. Moreover, with slight abuse of notation, we let $(z, Z) = (z, z_{-1}, \ldots, z_{-M}) \in \mathcal{X}^{M+1}$ and $(x, X) = (x, x_{-1}, \ldots, x_{-M}) \in \mathcal{X}^{M+1}$ be two independent random variables sampled from the stationary distribution μ^π. We define random variables $\widetilde{v}^{(h)}(Z)$ and $v^{(h)}(X)$ as

$$
\widetilde{v}^{(h)}(Z) := \sum_{i\in[M]} \widetilde{\sigma}^{(h)}_{-i} \cdot z_{-i}, \qquad \text{and} \qquad v^{(h)}(X) := \sum_{j\in[M]} \sigma^{(h)}_{-j} \cdot x_{-j}.
$$

Using $\widetilde{v}^{(h)}_{L+1}$, $v^{(h)}_l$, $\widetilde{v}^{(h)}(Z)$, and $v^{(h)}(X)$, we define two quantities A and B as

$$
\begin{aligned}
A &:= \frac{1}{L-M} \sum_{l=M+1}^{L} \mathbb{E}_{X|\pi}\bigg[\bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\mu^\pi(e_k)} - 1 \bigg) \cdot \prod_{h\in\mathcal{S}} \langle v^{(h)}_l, \widetilde{v}^{(h)}_{L+1} \rangle \bigg], \\
B &:= \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\bigg[\bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(e_k)} - 1 \bigg) \cdot \prod_{h\in\mathcal{S}} \langle \widetilde{v}^{(h)}(Z), v^{(h)}(X) \rangle \bigg],
\end{aligned}
$$

where $\mathbb{E}_{X\,|\,\pi}$ means that the expectation is taken with respect to the randomness of the Markov chain with transition π. Then, under Assumption 3.5, we have

$$
|A - B| \le \frac{8M}{L\gamma} + \frac{16\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda)\gamma^{|\mathcal{S}|/2+1}},
$$

where $\mu_0(\cdot)$ is the initial distribution over the first r_n tokens $X_{1:r_n}$ and $D_{\chi^2}(\mu_0 \,\|\, \mu^\pi)$ is a short-hand notation of $D_{\chi^2}(\mu_0(X_{1:r_n} = \cdot) \,\|\, \mu^\pi(X_{1:r_n} = \cdot))$.

https://doi.org/10.52202/079017-2127

Proof of Lemma F.5. By triangular inequality, we have

$$
\begin{aligned}
|A - B| \leq \Bigg| &\frac{1}{L-M} \sum_{l=M+1}^{L} \mathbb{E}_{X|\pi}\Bigg[\Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\mu^{\pi}(e_k)}\Bigg) \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, \widetilde{v}_{L+1}^{(h)} \rangle \Bigg] \\
&- \mathbb{E}_{(x,X),(z,Z)\sim\mu^{\pi}\times\mu^{\pi}}\Bigg[\Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^{\pi}(e_k)}\Bigg) \cdot \Bigg(\prod_{h\in\mathcal{S}} \langle \widetilde{v}^{(h)}(Z), v^{(h)}(X) \rangle\Bigg)\Bigg]\Bigg| \\
&+ \Bigg|\frac{1}{L-M} \sum_{l=M+1}^{L} \mathbb{E}_{X|\pi}\Bigg[\prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle\Bigg] - \mathbb{E}_{(x,X),(z,Z)\sim\mu^{\pi}\times\mu^{\pi}}\Bigg[\Bigg(\prod_{h\in\mathcal{S}} \langle \widetilde{v}^{(h)}(Z), v^{(h)}(X) \rangle\Bigg)\Bigg]\Bigg|.
\end{aligned}
$$

We will establish the upper bounds for each of the absolute value terms. We first focus on the first absolute value term.

Bounding the First Absolute Value Term. Let $p^{\pi}(X)$ denote the joint distribution of the whole sequence X under kernel π. By the definitions of $\widetilde{v}_{L+1}^{(h)}$ and $v_l^{(h)}$, we have

$$
\begin{aligned}
\langle v_l^{(h)}, v_{L+1}^{(h)} \rangle &= \sum_{i_h, j_h \in [M]} \sigma_{-i_h}^{(h)} \cdot \sigma_{-j_h}^{(h)} \cdot \langle x_{L+1-i_h}, x_{l-j_h} \rangle \\
&= \sum_{i_h, j_h \in [M]} \sum_{k\in[d]} \sigma_{-i_h}^{(h)} \cdot \sigma_{-j_h}^{(h)} \cdot \mathbb{1}(x_{L+1-i_h} = x_{l-j_h} = e_k),
\end{aligned}
$$

where we use (i_h, j_h) as the indices to highlight that they are associated with head h. And we use $k_h \in [d]$ to index all the possible common values for x_{l-i_h} and x_{L+1-j_h}. Then plugging this equality into $\prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle$ and exchanging the order of product and summation, we have

$$
\begin{aligned}
&\Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\mu^{\pi}(e_k)}\Bigg) \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle \qquad \text{(F.6)} \\
&\quad = \sum_{\{(i_h,j_h)\}_{h\in\mathcal{S}}} \sum_{\{k_h\}_{h\in\mathcal{S}}, k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\mu^{\pi}(z = e_k)} \cdot \Bigg(\prod_{h\in\mathcal{S}} \sigma_{-i_h}^{(h)} \cdot \sigma_{-j_h}^{(h)} \cdot \mathbb{1}(x_{L+1-i_h} = x_{l-j_h} = e_{k_h})\Bigg),
\end{aligned}
$$

where the summation means that we sum over all possible values that $\{i_h, j_h, k_h\}_{h\in\mathcal{S}}$ and k can take. Specifically, each i_h and j_h take values in $[M]$, and each k_h and k takes values in $[d]$. Moreover, using the property of indicator functions, we can further simplify (F.6) by gathering all indicators:

$$
\begin{aligned}
&\Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\mu^{\pi}(e_k)}\Bigg) \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle \qquad \text{(F.7)} \\
&\quad = \sum_{\{(i_h,j_h)\}_{h\in\mathcal{S}}} \Bigg(\prod_{h\in\mathcal{S}} \sigma_{-i_h}^{(h)} \cdot \sigma_{-j_h}^{(h)}\Bigg) \cdot \Bigg(\sum_{\{k_h\}_{h\in\mathcal{S}}, k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k, x_{L+1-i_h} = x_{l-j_h} = e_{k_h}, \forall h \in \mathcal{S})}{\mu^{\pi}(z = e_k)}\Bigg).
\end{aligned}
$$

Now we take expectations with respect to the randomness of X on both ends of (F.7) and get

$$
\begin{aligned}
&\frac{1}{L-M} \sum_{l=M+1}^{L} \mathbb{E}\Bigg[\Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\mu^{\pi}(e_k)}\Bigg) \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle\Bigg] \\
&= \sum_{\{(i_h,j_h)\}_{h\in\mathcal{S}}} \Bigg(\prod_{h\in\mathcal{S}} \widetilde{\sigma}_{-i_h}^{(h)} \sigma_{-j_h}^{(h)}\Bigg) \cdot \sum_{\{k_h\}_{h\in\mathcal{S}}, k\in[d]} \frac{\sum_{l=M+1}^{L} p^{\pi}(x_{L+1} = x_l = e_k, x_{L+1-i_h} = x_{l-j_h} = e_{k_h}, \forall h \in \mathcal{S})}{(L-M) \cdot \mu^{\pi}(z = e_k)}.
\end{aligned}
$$

To further simplify the above equality, we define a new probability distribution over $X_{L+1-M:L+1}$ and another subsequence of length $M+1$. Note that $X_{L+1-M:L+1}$ contains is a subsequence with $M+1$ tokens. We let $(z, Z) = (z, z_{-1}, \ldots, z_{-1}, z_{-M})$ denote a random token sequence of size $M+1$ in reverse order. We define a joint distribution $\widehat{p}^{\pi}$ over $X_{L+1-M:L+1}$ and (z, Z) as follows. Let $E = (E_0, E_{-1}, \ldots, E_{-M})$ and $E' = (E'_0, E'_{-1}, \ldots, E'_{-M})$ be two elements in $\mathcal{X}^{M+1}$. That is, each component of E and E' are in $\mathcal{X}$. The probability mass function of $\widehat{p}^{\pi}$ is defined as

$$\widehat{p}^{\pi}\big((x_{L+1}, x_L, \ldots, x_{L+1-M}) = E, (z, Z) = E'\big) \tag{F.8}$$
$$= \frac{1}{L-M} \sum_{l=M+1}^{L} p^{\pi}\big((x_{L+1}, x_L, \ldots, x_{L+1-M}) = E, (x_l, x_{l-1}, \ldots, x_{l=M}) = E'\big).$$

That is, $\widehat{p}^{\pi}$ can be viewed as the joint distribution of $X_{L+1-M:L+1}$ with an averaged distribution of the history. When L is sufficiently large, by the mixing property of the Markov chain, we expect that, under $\widehat{p}^{\pi}$, (z, Z) is approximately independent of $X_{L+1-M:L+1}$, and the marginal distributions of (z, Z) and $X_{L+1-M:L+1}$ are both close to the stationary distribution μ^{π}. We will translate this intuition into a rigorous argument in Lemma F.17, which bounds the total-variation distance between $\widehat{p}^{\pi}$ and the product distribution $\mu^{\pi} \times \mu^{\pi}$.

With $\widehat{p}^{\pi}$ defined in (F.8), we can rewrite the expectation above as

$$\frac{1}{L-M} \sum_{l=M+1}^{L} \mathbb{E}\bigg[\bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\mu^{\pi}(e_k)}\bigg) \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\bigg] \tag{F.9}$$
$$= \sum_{\{(i_h, j_h)\}_{h\in\mathcal{S}}} \bigg(\prod_{h\in\mathcal{S}} \widetilde{\sigma}_{-i_h}^{(h)} \sigma_{-j_h}^{(h)}\bigg) \cdot \sum_{\{k_h\}_{h\in\mathcal{S}}, k\in[d]} \frac{\widehat{p}^{\pi}(x_{L+1} = z = e_k, x_{L+1-i_h} = z_{-j_h} = e_{k_h}, \forall h \in \mathcal{S})}{\mu^{\pi}(z = e_k)}.$$

Similarly, by the definitions of $\widetilde{v}^{(h)}(Z)$ and $v^{(h)}(Z)$, we can write $\langle \widetilde{v}^{(h)}(Z), v^{(h)}(X)\rangle$ as

$$\langle \widetilde{v}^{(h)}(Z), v^{(h)}(X)\rangle = \sum_{i_h, j_h \in [M]} \sum_{k_h\in[d]} \sigma_{-i_h}^{(h)} \cdot \sigma_{-j_h}^{(h)} \cdot \mathbb{1}(z_{-i_h} = x_{j_h} = e_k).$$

Then, multiplying these terms with $h \in \mathcal{S}$, we can write

$$\bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^{\pi}(e_k)}\bigg) \cdot \prod_{h\in\mathcal{S}} \langle \widetilde{v}^{(h)}(Z), v^{(h)}(X)\rangle$$
$$= \sum_{\{(i_h, j_h)\}_{h\in\mathcal{S}}} \bigg(\prod_{h\in\mathcal{S}} \sigma_{-i_h}^{(h)} \cdot \sigma_{-j_h}^{(h)}\bigg) \cdot \bigg(\sum_{\{k_h\}_{h\in\mathcal{S}}, k\in[d]} \frac{\mathbb{1}(z = x = e_k, z_{-i_h} = x_{-j_h} = e_{k_h}, \forall h \in \mathcal{S})}{\mu^{\pi}(z = e_k)}\bigg). \tag{F.10}$$

Recall that here $(z, Z) = (z, z_{-1}, \ldots, z_{-M})$ and $(x, X) = (x, x_{-1}, \ldots, x_{-M})$ are independently sampled from the stationary distribution μ^{π}. Taking the expectation under μ^{π}, we have

$$\mathbb{E}_{(x,X),(z,Z)\sim\mu^{\pi}\times\mu^{\pi}}\bigg[\bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^{\pi}(e_k)}\bigg) \cdot \bigg(\prod_{h\in\mathcal{S}} \langle \widetilde{v}^{(h)}(Z), v^{(h)}(X)\rangle\bigg)\bigg] \tag{F.11}$$
$$= \sum_{\{(i_h, j_h)\}_{h\in\mathcal{S}}} \bigg(\prod_{h\in\mathcal{S}} \widetilde{\sigma}_{-i_h}^{(h)} \sigma_{-j_h}^{(h)}\bigg) \cdot \sum_{\{k_h\}_{h\in\mathcal{S}}, k\in[d]} \frac{\mu^{\pi}(x = e_k, x_{-i_h} = e_{k_h}, \forall h \in \mathcal{S}) \cdot \mu^{\pi}(z = e_k, z_{-j_h} = e_{k_h}, \forall h \in \mathcal{S})}{\mu^{\pi}(z = e_k)}.$$

To bound the first absolute value term in the upper bound on $|A - B|$, we aim to compare (F.9) and (F.11). To this end, let us fix collections of index pairs $(i_h, j_h)_{h\in\mathcal{S}}$. Let $\mathcal{S}_1 = \{i_h : h \in \mathcal{S}\}$ and $\mathcal{S}_2 = \{j_h : h \in \mathcal{S}\}$ be the unique values in $(i_h)_{h\in\mathcal{S}}$ and $(j_h)_{h\in\mathcal{S}}$. Since there might exists two elements h and h' in $\mathcal{S}$ such that $i_h = i_{h'}$ or $j_h = j_{h'}$, $|\mathcal{S}_1|$ and $|\mathcal{S}_2|$ might be strictly less than $|\mathcal{S}|$. As a result, $\widehat{p}^{\pi}(x_{L+1} = z = e_k, x_{L+1-i_h} = z_{-j_h} = e_{k_h}, \forall h \in \mathcal{S})$ only involves random variables x_{L+1}, $X_{L+1-\mathcal{S}_1} = \{x_{L+1-i}\}_{i\in\mathcal{S}_1}$, z, $Z_{-\mathcal{S}_2} = \{z_{-j}\}_{j\in\mathcal{S}_2}$, which are a subset of the random variables defined in (F.8). Similarly,

$$\mu^{\pi}(x = e_k, x_{-i_h} = e_{k_h}, \forall h \in \mathcal{S}) \cdot \mu^{\pi}(z = e_k, z_{-j_h} = e_{k_h}, \forall h \in \mathcal{S})$$

only involves a subset of random variables x, $X_{-\mathcal{S}_1} = \{x_{-i}\}_{i\in\mathcal{S}_1}$, z, and $Z_{-\mathcal{S}_2}$. Let us define $\bar{E} = (E_0, (E_{-i})_{i\in\mathcal{S}_1}) \in \mathcal{X}^{|\mathcal{S}_1|+1}$ and $\bar{E}' = (E'_0, (E'_{-j})_{j\in\mathcal{S}_2}) \in \mathcal{X}^{|\mathcal{S}_2|+1}$. By enumerating $\bar{E}$

in $\mathcal{X}^{|\mathcal{S}_1|+1}$ and $\bar{E}'$ in $\mathcal{X}^{|\mathcal{S}_2|+1}$, we equivalently enumerate all possible values the above random variables can take. Therefore, by comparing (F.9) with (F.11), we have

$$\begin{aligned}
&\sum_{\{k_h\}_{h\in\mathcal{S}},k\in[d]} \Big|\widehat{p}^{\pi}(x_{L+1} = z = e_k, x_{L+1-i_h} = z_{-j_h} = e_{k_h}, \forall h \in \mathcal{S}) \\
&\qquad\qquad - \mu^{\pi}(x = e_k, x_{-i_h} = e_{k_h}, \forall h \in \mathcal{S}) \cdot \mu^{\pi}(z = e_k, z_{-j_h} = e_{k_h}, \forall h \in \mathcal{S})\Big| \\
&\quad = \sum_{\bar{E},\bar{E}'} \Big|\widehat{p}^{\pi}\big((x_{L+1}, X_{L+1-\mathcal{S}_1}) = \bar{E}, (z, Z_{-\mathcal{S}_2}) = \bar{E}'\big) - \mu^{\pi}\big((x_{L+1}, X_{L+1-\mathcal{S}_1}) = \bar{E}\big) \cdot \mu^{\pi}\big((z, Z_{-\mathcal{S}_2}) = E'\big)\Big| \\
&\qquad\qquad \cdot \mathbb{1}(E_0 = E'_0, E_{-i_h} = E'_{-j_h}, \forall h \in \mathcal{S}) \\
&\quad \le 2\|\widehat{p}^{\pi}(Y = \cdot, Y' = \cdot) - \mu^{\pi}(Y = \cdot) \times \mu^{\pi}(Y' = \cdot)\|_{\mathrm{TV}},
\end{aligned} \tag{F.12}$$

where in the last line, we use Y and Y' as placeholders for the random variables $(x_{L+1}, X_{L+1-\mathcal{S}_1})$ and $(z, Z_{-\mathcal{S}_2})$ respectively. In the first equality, we sum over $\bar{E} \in \mathcal{X}^{|\mathcal{S}_1|+1}$ and $\bar{E}' \in \mathcal{X}^{|\mathcal{S}_2|+1}$, and the last inequality follows from the definition of total variation distance and dropping the indicator. By Lemma F.17, this total variation distance is bounded by

$$\begin{aligned}
&2\|\widehat{p}^{\pi}(Y = \cdot, Y' = \cdot) - \mu^{\pi}(Y = \cdot) \times \mu^{\pi}(Y' = \cdot)\|_{\mathrm{TV}} \\
&\quad \le \frac{4M}{L} + \frac{8\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^{\pi}) + 1}}{L(1-\lambda) \cdot \sqrt{\min_{x_{L+1}, X_{L+1-\mathcal{S}_1}} \mu^{\pi}(x_{L+1}, X_{L+1-\mathcal{S}_1})}} \\
&\quad \le \frac{4M}{L} + \frac{8\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^{\pi}) + 1}}{L(1-\lambda) \cdot \gamma^{(|\mathcal{S}|+1)/2}},
\end{aligned} \tag{F.13}$$

where the last inequality holds by Corollary F.15 and the fact that $|\mathcal{S}_1| \le |\mathcal{S}|$. Specifically, Corollary F.15 implies that the density function of the joint distribution of x_{L+1} and $X_{L+1-\mathcal{S}_1}$ is lower bounded by $\gamma^{|\mathcal{S}_1|+1} \ge \gamma^{|\mathcal{S}|+1}$. Thus, combining (F.12) and (F.13), we have

$$\begin{aligned}
&\Bigg|\sum_{\{k_h\}_{h\in\mathcal{S}},k\in[d]} \frac{\widehat{p}^{\pi}(x_{L+1} = z = e_k, x_{L+1-i_h} = z_{-j_h} = e_{k_h}, \forall h \in \mathcal{S})}{\mu^{\pi}(z = e_k)} \\
&\qquad - \sum_{\{k_h\}_{h\in\mathcal{S}},k\in[d]} \frac{\mu^{\pi}(x = e_k, x_{-i_h} = e_{k_h}, \forall h \in \mathcal{S}) \cdot \mu^{\pi}(z = e_k, z_{-j_h} = e_{k_h}, \forall h \in \mathcal{S})}{\mu^{\pi}(z = e_k)}\Bigg| \\
&\quad \le \frac{1}{\gamma} \cdot \left(\frac{4M}{L} + \frac{8\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^{\pi}) + 1}}{L(1-\lambda) \cdot \gamma^{(|\mathcal{S}|+1)/2}}\right).
\end{aligned} \tag{F.14}$$

Therefore, to bound the first absolute value term, we combine (F.9), (F.11), and (F.14) and use triangle inequality to get

$$\begin{aligned}
&\Bigg|\frac{1}{L-M} \sum_{l=M+1}^{L} \mathbb{E}_{X|\pi}\Bigg[\Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\mu^{\pi}(e_k)}\Bigg) \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, \widetilde{v}_{L+1}^{(h)} \rangle\Bigg] \\
&\qquad - \mathbb{E}_{(x,X),(z,Z)\sim\mu^{\pi}\times\mu^{\pi}}\Bigg[\Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^{\pi}(e_k)}\Bigg) \cdot \Bigg(\prod_{h\in\mathcal{S}} \langle \widetilde{v}^{(h)}(Z), v^{(h)}(X) \rangle\Bigg)\Bigg]\Bigg| \\
&\quad \le \frac{1}{\gamma} \cdot \left(\frac{4M}{L} + \frac{8\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^{\pi}) + 1}}{L(1-\lambda) \cdot \gamma^{(|\mathcal{S}|+1)/2}}\right) \cdot \sum_{\{(i_h,j_h)\}_{h\in\mathcal{S}}} \left(\prod_{h\in\mathcal{S}} \widetilde{\sigma}_{-i_h}^{(h)} \sigma_{-j_h}^{(h)}\right).
\end{aligned} \tag{F.15}$$

Furthermore, recall that $\widetilde{\sigma}^{(h)}$ and $\sigma^{(h)}$ are probability distributions over $[M]$ for all $h \in \mathcal{S}$. By going over all possible values that $\{(i_h, j_h)\}_{h\in\mathcal{S}}$ can take, we have

$$\sum_{\{(i_h,j_h)\}_{h\in\mathcal{S}}} \left(\prod_{h\in\mathcal{S}} \widetilde{\sigma}_{-i_h}^{(h)} \sigma_{-j_h}^{(h)}\right) = \prod_{h\in\mathcal{S}} \left(\sum_{k\in[M]} \widetilde{\sigma}_{-k}^{(h)}\right) \cdot \left(\sum_{k\in[M]} \sigma_{-k}^{(h)}\right) = 1. \tag{F.16}$$

Plugging this equality into (F.15), we show that the upper bound in (F.15) can be reduced to the right-hand side of (F.14).

Bounding the Second Absolute Value Term. For the second absolute value term, an analogous argument can be applied. In fact, the proof is simpler because we only need to handle $\langle v_l^{(h)}, v_{L+1}^{(h)}$ and $\langle \widetilde{v}^{(h)}(Z), v^{(h)}(Z)\rangle$ and do not have indicators $\mathbb{1}(x_{L+1} = x_l)$ and $\mathbb{1}(x = z)$.

Similar to the derivation in (F.9) and (F.11),

$$\begin{aligned}
&\bigg|\frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}\bigg[\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, \widetilde{v}_{L+1}^{(h)}\rangle\bigg] - \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\bigg[\bigg(\prod_{h\in\mathcal{S}}\langle \widetilde{v}^{(h)}(Z), v^{(h)}(X)\rangle\bigg)\bigg]\bigg| \\
&\qquad = \bigg|\sum_{\{(i_h,j_h)\}_{h\in\mathcal{S}}}\bigg(\prod_{h\in\mathcal{S}}\widetilde{\sigma}_{-i_h}^{(h)}\sigma_{-j_h}^{(h)}\bigg)\cdot\sum_{\{k_h\}_{h\in\mathcal{S}}}\Big(\widehat{p}^\pi(x_{L+1-i_h} = z_{-j_h} = e_{k_h}, \forall h\in\mathcal{S}) \qquad \text{(F.17)}\\
&\qquad\qquad\qquad - \mu^\pi(x_{-i_h} = e_{k_h}, \forall h\in\mathcal{S})\cdot\mu^\pi(z_{-j_h} = e_{k_h}, \forall h\in\mathcal{S})\Big)\bigg|.
\end{aligned}$$

Similar to (F.12), for any fixed collection of index pairs $(i_h, j_h)_{h\in\mathcal{S}}$, we let $\mathcal{S}_1 = \{i_h : h\in\mathcal{S}\}$ and $\mathcal{S}_2 = \{j_h : h\in\mathcal{S}\}$ denote the unique values in $(i_h)_{h\in\mathcal{S}}$ and $(j_h)_{h\in\mathcal{S}}$. By Lemma F.17, we have

$$\begin{aligned}
&\bigg|\sum_{\{k_h\}_{h\in\mathcal{S}}}\Big(\widehat{p}^\pi(x_{L+1-i_h} = z_{-j_h} = e_{k_h}, \forall h\in\mathcal{S}) - \mu^\pi(x_{-i_h} = e_{k_h}, \forall h\in\mathcal{S})\cdot\mu^\pi(z_{-j_h} = e_{k_h}, \forall h\in\mathcal{S})\Big)\bigg| \\
&\qquad \le 2\big\|\widehat{p}^\pi(\widetilde{Y} = \cdot, \widetilde{Y}' = \cdot) - \mu^\pi(\widetilde{Y} = \cdot)\times\mu^\pi(\widetilde{Y}' = \cdot)\big\|_{\mathrm{TV}} \le \frac{4M}{L} + \frac{8\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}}{L(1-\lambda)\cdot\gamma^{|\mathcal{S}|/2}}. \qquad \text{(F.18)}
\end{aligned}$$

Here we use $\widetilde{Y}$ and $\widetilde{Y}'$ as placeholders for random variables $X_{L+1-\mathcal{S}_1}$ and $Z_{-\mathcal{S}_2}$. We note that Lemma F.17 can be applied to any subsets of $X_{L+1-M:L+1}$ and (z, Z). Therefore, combining (F.16), (F.17), and (F.18), we conclude that

$$\begin{aligned}
&\bigg|\frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{E}\bigg[\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, \widetilde{v}_{L+1}^{(h)}\rangle\bigg] - \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\bigg[\bigg(\prod_{h\in\mathcal{S}}\langle \widetilde{v}^{(h)}(Z), v^{(h)}(X)\rangle\bigg)\bigg]\bigg| \\
&\quad \le \frac{4M}{L} + \frac{8\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}}{L(1-\lambda)\gamma^{|\mathcal{S}|/2}}.
\end{aligned}$$

Note that the second upper bound is dominated by the previous one. This completes the proof of Lemma F.5. □

Lemma F.6 provides an approximation result using the definition of the modified χ^2-mutual information.

Lemma F.6. *Consider a fixed set $\mathcal{S}\in[H]_{\le D}$. For any $h\in\mathcal{S}$, let $\widetilde{\sigma}^{(h)}$ and $\sigma^{(h)}$ be two probability distributions over $[M]$. That is, for any $i, j\in[M]$, we have $\widetilde{\sigma}_{-i}^{(h)}, \sigma_{-j}^{(h)}\in[0,1]$, and $\sum_{i=1}^M\widetilde{\sigma}_{-i}^{(h)} = \sum_{j=1}^M\sigma_{-j}^{(h)} = 1$. Moreover, we let $(z, Z) = (z, z_{-M}, \ldots, z_{-1})\in\mathcal{X}^{M+1}$ and $(x, X) = (x, x_{-M}, \ldots, x_{-1})\in\mathcal{X}^{M+1}$ be two independent random variables sampled from the stationary distribution μ^π. We define random variables $\widetilde{v}^{(h)}(Z)$ and $v^{(h)}(X)$ as*

$$\widetilde{v}^{(h)}(Z) := \sum_{i\in[M]}\widetilde{\sigma}_{-i}^{(h)}\cdot z_{-i}, \qquad \text{and} \qquad v^{(h)}(X) := \sum_{j\in[M]}\sigma_{-j}^{(h)}\cdot x_{-j}.$$

Let $(i_h^\star, j_h^\star)_{h\in\mathcal{S}}$ be any fixed collection of index pairs, where $i_h^\star\in[M]$ and $j_h^\star\in[M]$ for all $h\in\mathcal{S}$. We define quantities A and B as

$$A := \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\bigg[\bigg(\sum_{k\in[d]}\frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(e_k)} - 1\bigg)\cdot\prod_{h\in\mathcal{S}}\langle\widetilde{v}^{(h)}(Z), v^{(h)}(X)\rangle,\bigg],$$

$$B := \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\bigg[\prod_{h\in\mathcal{S}}\mathbb{1}(x_{-i_h^\star} = z_{-j_h^\star})\cdot\bigg(\sum_{k=1}^{d}\frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(e_k)} - 1\bigg)\bigg].$$

Under Assumption 3.5, it holds that

$$\bigg|\mathbb{E}_\pi[A] - \prod_{h\in\mathcal{S}}\sigma_{-i_h^\star}^{(h)}\widetilde{\sigma}_{-j_h^\star}^{(h)}\cdot B\bigg| \le \bigg(1 - \prod_{h\in\mathcal{S}}\sigma_{-i_h^\star}^{(h)}\widetilde{\sigma}_{-j_h^\star}^{(h)}\bigg)\cdot I_{\chi^2}(\mathcal{S}^\star).$$

Proof of Lemma F.6. To simplify the notation, we define a signal set $\Gamma(\mathcal{S})$ and an error set $\bar{\Gamma}(\mathcal{S})$ as

$$\Gamma(\mathcal{S}) := \{(i_h^\star, j_h^\star)_{h\in\mathcal{S}}\}, \qquad \bar{\Gamma}(\mathcal{S}) := \left\{(i_h, j_h)_{h\in\mathcal{S}} \in ([M]\times[M])^{|\mathcal{S}|}\right\} \backslash \Gamma(\mathcal{S}).$$

Similar to (F.10), we can write $\mathbb{E}_\pi[A]$ as

$$\mathbb{E}_\pi[A] = \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\Bigg[\sum_{\{(i_h,j_h)\}_{h\in\mathcal{S}}} \prod_{h\in\mathcal{S}} \sigma^{(h)}_{-i_h}\widetilde{\sigma}^{(h)}_{-j_h} \cdot \mathbb{1}(x_{-i_h} = z_{-j_h}) \cdot \Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(e_k)} - 1\Bigg)\Bigg],$$

where we exchange the order of product and summation. Using the notation $\bar{\Gamma}(\mathcal{S})$, we can split the summation into two parts:

$$\begin{aligned}
\mathbb{E}_\pi[A] &= \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\Bigg[\prod_{h\in\mathcal{S}} \sigma^{(h)}_{-i_h^\star}\widetilde{\sigma}^{(h)}_{-j_h^\star} \cdot \mathbb{1}(x_{-i_h^\star} = z_{-j_h^\star}) \cdot \Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(e_k)} - 1\Bigg)\Bigg] \\
&\quad + \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\Bigg[\sum_{\{(i_h,j_h)\}_{h\in\mathcal{S}}\in\bar{\Gamma}(\mathcal{S})} \prod_{h\in\mathcal{S}} \sigma^{(h)}_{-i_h}\widetilde{\sigma}^{(h)}_{-j_h} \cdot \mathbb{1}(x_{-i_h} = z_{-j_h})\Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(e_k)} - 1\Bigg)\Bigg] \\
&= \prod_{h\in\mathcal{S}} \sigma^{(h)}_{-i_h^\star}\widetilde{\sigma}^{(h)}_{-j_h^\star} \cdot B \\
&\quad + \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\Bigg[\sum_{(i_h,j_h)_{h\in\mathcal{S}}\in\bar{\Gamma}(\mathcal{S})} \prod_{h\in\mathcal{S}} \sigma^{(h)}_{-i_h}\widetilde{\sigma}^{(h)}_{-j_h} \cdot \mathbb{1}(x_{-i_h} = z_{-j_h})\Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(e_k)} - 1\Bigg)\Bigg].
\end{aligned}$$

Here last equality holds by the definition of the B and the fact that $\widetilde{\sigma}^{(h)}$ and $\sigma^{(h)}$ are fixed vectors.

Therefore, to prove this lemma, it suffices to upper bound the second term above. To this end, we apply Lemma F.7 stated below for any fixed set of indices $(i_h, j_h)_{h\in\mathcal{S}} \in \bar{\Gamma}(\mathcal{S})$. Specifically, let $\mathcal{S}_1 = \{i_h \colon h \in \mathcal{S}\}$ and $\mathcal{S}_2 = \{j_h \colon h \in \mathcal{S}\}$ denote the unique values of $(i_h)_{h\in\mathcal{S}}$ and $(j_h)_{h\in\mathcal{S}}$. Lemma F.7 implies that

$$\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\Bigg[\prod_{h\in\mathcal{S}} \mathbb{1}(x_{-i_h} = z_{-j_h}) \cdot \Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(z = e_k)} - 1\Bigg)\Bigg] \le I_{\chi^2}(\mathcal{S}^\star). \tag{F.19}$$

Combining (F.19) with the fact that

$$\sum_{(i_h,j_h)_{h\in\mathcal{S}}\in\bar{\Gamma}(\mathcal{S})} \prod_{h\in\mathcal{S}} \sigma^{(h)}_{-i_h}\widetilde{\sigma}^{(h)}_{-j_h} = 1 - \prod_{h\in\mathcal{S}} \sigma^{(h)}_{-i_h^\star}\widetilde{\sigma}^{(h)}_{-j_h^\star},$$

the desired term is bounded above by $(1-\prod_{h\in\mathcal{S}} \sigma^{(h)}_{-i_h^\star}\widetilde{\sigma}^{(h)}_{-j_h^\star})\cdot I_{\chi^2}(\mathcal{S}^\star)$, which concludes the proof. □

Lemma F.7. *Let $\mathcal{S} \in [H]_{\le D}$ be a fixed subset and let $\{(i_h, j_h)\}_{h\in\mathcal{S}}$ be a fixed collection of index pairs, where $i_h, j_h \in [M]$ for all $h \in \mathcal{S}$. Let $\mathcal{S}_1 = \{i_h \colon h \in \mathcal{S}\}$ and $\mathcal{S}_2 = \{j_h \colon h \in \mathcal{S}\}$ denote the unique values of $(i_h)_{h\in\mathcal{S}}$ and $(j_h)_{h\in\mathcal{S}}$. We let $(z, Z) = (z, z_{-M}, \ldots, z_{-1}) \in \mathcal{X}^{M+1}$ and $(x, X) = (x, x_{-M}, \ldots, x_{-1}) \in \mathcal{X}^{M+1}$ be two independent random variables sampled from the stationary distribution μ^π, where π is the transition kernel of the Markov chain and is sampled from prior $\mathcal{P}$. If Assumption 3.5 holds, it follows that*

$$\begin{aligned}
&\mathbb{E}_{\pi\sim\mathcal{P},(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\Bigg[\prod_{h\in\mathcal{S}} \mathbb{1}(x_{-i_h} = z_{-j_h}) \cdot \Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(z = e_k)} - 1\Bigg)\Bigg] \\
&\qquad \le \frac{1}{2}\left(\widetilde{I}_{\chi^2}(\mathcal{S}_1) + \widetilde{I}_{\chi^2}(\mathcal{S}_2)\right) \le \widetilde{I}_{\chi^2}(\mathcal{S}^\star),
\end{aligned}$$

where $\widetilde{I}_{\chi^2}(\mathcal{S})$ is the modified χ^2-mutual information defined in Definition 3.1 and $\mathcal{S}^\star = \operatorname{argmax}_{\mathcal{S}\in[H]_{\le D}} \widetilde{I}_{\chi^2}(\mathcal{S})$.

Proof of Lemma F.7. We first note that it is allowed $|\mathcal{S}_1| \ne |\mathcal{S}_2|$ as there could be duplicate values in both $(i_h)_{h\in\mathcal{S}}$ and $(j_h)_{h\in\mathcal{S}}$, while $\mathcal{S}_1$ and $\mathcal{S}_2$ are the unique values. In the sequel, we let $X_{-\mathcal{S}_1}$ denote $\{x_{-i_h}\}_{h\in\mathcal{S}}$ and let $Z_{-\mathcal{S}_2}$ denote $\{z_{-j_h}\}_{h\in\mathcal{S}}$, where repeated elements are removed. Moreover, we let

$\{X_{-\mathcal{S}_1} = Z_{-\mathcal{S}_2}\}$ be the event that $x_{-i_h} = z_{-j_h}$ for all $h \in \mathcal{S}$. Notice that $\prod_{h\in\mathcal{S}} \mathbb{1}(x_{-i_h} = z_{-j_h}) = \mathbb{1}(X_{-\mathcal{S}_1} = Z_{-\mathcal{S}_2})$. Then, we have

$$
\begin{aligned}
&\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\left[\prod_{h\in\mathcal{S}} \mathbb{1}(x_{-i_h} = z_{-j_h}) \cdot \left(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(z = e_k)} - 1\right)\right] \qquad \text{(F.20)}\\
&\quad = \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\left[\mathbb{1}(X_{-\mathcal{S}_1} = Z_{-\mathcal{S}_2}) \cdot \left(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(z = e_k)} - 1\right)\right]\\
&\quad = \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\left[\left(\sum_{k\in[d]} \frac{\mu^\pi(x = e_k|X_{-\mathcal{S}_1}) \cdot \mu^\pi(z = e_k|Z_{-\mathcal{S}_2})}{\mu^\pi(z = e_k)} - 1\right) \cdot \mathbb{1}(X_{-\mathcal{S}_1} = Z_{-\mathcal{S}_2})\right]\\
&\quad = \mathbb{E}_{\pi,(X,Z)\sim\mu^\pi\times\mu^\pi}\left[\sum_{k\in[d]} \left(\frac{\mu^\pi(x = e_k|X_{-\mathcal{S}_1})}{\mu^\pi(x = e_k)} - 1\right) \cdot \left(\frac{\mu^\pi(z = e_k|Z_{-\mathcal{S}_2})}{\mu^\pi(z = e_k)} - 1\right) \cdot \mu^\pi(z = e_k) \cdot \mathbb{1}(X_{-\mathcal{S}_1} = Z_{-\mathcal{S}_2})\right].
\end{aligned}
$$

Here in the second equality, we take a conditional expectation given $X_{-\mathcal{S}_1}$ and $Z_{-\mathcal{S}_2}$. The last equality can be verified by direct computation. To simplify the expectation above, we aim to transform the indicator of $\mathbb{1}(X_{-\mathcal{S}_1} = Z_{-\mathcal{S}_2})$ into probabilities involving $X_{-\mathcal{S}_1}$ and $Z_{-\mathcal{S}_2}$. To this end, we need to explicitly enumerate all possible values that $X_{-\mathcal{S}_1}$ and $Z_{-\mathcal{S}_2}$ can take. This is challenging, as there may be duplicated values in both i_h and j_h, and thus $X_{-\mathcal{S}_1}$ and $Z_{-\mathcal{S}_2}$ can have different sizes. However, since $\mathcal{S}_1$ is a "reduction" of $\{i_h\}_{h\in\mathcal{S}}$, we can revert to the original space and consider $E = (E_h)_{h\in\mathcal{S}} \in \mathcal{X}^{|\mathcal{S}|}$ that *respects* the reduction from $\{i_h\}_{h\in\mathcal{S}}$ to $\mathcal{S}_1$. Here each E_h is the value x_{i_h} takes. In other words, with $(i_h)_{h\in\mathcal{S}}$ that might have duplicated values, we consider the values taken by $(x_{i_h})_{h\in\mathcal{S}}$, with duplicates allowed. And E has the same duplication structure as $(x_{i_h})_{h\in\mathcal{S}}$. In the following, we describe these values by introducing the notion of compatibility.

Definition F.8 (Compatible Value Set). *We say that $E \in \mathcal{X}^{|\mathcal{S}|}$ is* compatible *with $(i_h)_{h\in\mathcal{S}}$ if, for any $h \neq h'$ such that $i_h = i_{h'}$, we have $E_h = E_{h'}$. In other words, the* unique *values in E can be indexed by $\{i_h\}_{h\in\mathcal{S}} = \mathcal{S}_1$ if E is compatible with $(i_h)_{h\in\mathcal{S}}$.*

By this definition, E is compatible with $(i_h)_{h\in\mathcal{S}}$ if it respect duplication pattern of $(x_{i_h})_{h\in\mathcal{S}}$. If $i_h = i_{h'}$, then we know that x_{i_h} and $x_{i_{h'}}$ is the same token. Since x_{i_h} and $x_{i_{h'}}$ take values E_h and $E_{h'}$, we must have $E_h = E_{h'}$. As a concrete example, suppose $\mathcal{S} = \{1, 2, 3\}$, and the values of $(i_h)_{h\in\mathcal{S}}$ are given by $(i_1, i_2, i_3) = (1, 2, 1)$. Therefore, we have $\mathcal{S}_1 = \{1, 2\}$, which contains the unique values of (i_1, i_2, i_3). Now, let $E = (E_1, E_2, E_3)$. For E to be *compatible* with $(i_h)_{h\in\mathcal{S}}$, we must have $E_1 = E_3$ since $i_1 = i_3$. There is no restriction on E_2. So, a compatible value set for this example could be $E = (a, b, a)$, where a and b are elements of $\mathcal{X}$.

In the sequel, we define $\mathcal{E}$ as the set of vectors in $\mathcal{X}^{|\mathcal{S}|}$ that are compatible with both $\{i_h\}_{h\in\mathcal{S}}$ and $\{j_h\}_{h\in\mathcal{S}}$, i.e.,

$$
\mathcal{E} = \left\{E \in \mathcal{X}^{|\mathcal{S}|} \mid E \text{ is compatible with both } (i_h)_{h\in\mathcal{S}} \text{ and } (j_h)_{h\in\mathcal{S}}\right\}.
$$

The compatibility condition allows us to assign x_{-i_h}, z_{-j_h} the value E_h for all $h \in \mathcal{S}$ when $E \in \mathcal{E}$. Under this assignment, the constraint $X_{-\mathcal{S}_1} = Z_{-\mathcal{S}_2}$ is automatically satisfied. We use the notation $\{X_{-\mathcal{S}_1} = E\}$ to denote the event that $x_{-i_h} = E_h$ for all $h \in \mathcal{S}$, and similarly for $Z_{-\mathcal{S}_2} = E$. In particular, we are able to rewrite the indicator $\mathbb{1}(X_{-\mathcal{S}_1} = Z_{-\mathcal{S}_2})$ as $\sum_{E\in\mathcal{E}} \mathbb{1}(X_{-\mathcal{S}_1} = E, Z_{-\mathcal{S}_2} =$

E). Then we can rewrite (F.20) by separating $X_{-\mathcal{S}_1}$ and $Z_{-\mathcal{S}_2}$ as

$$\begin{aligned}
&\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\left[\prod_{h\in\mathcal{S}}\mathbb{1}(x_{-i_h}=z_{-j_h})\cdot\left(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(z=e_k)}-1\right)\right]\\
&\quad=\mathbb{E}_\pi\Bigg[\sum_{E\in\mathcal{E}}\sum_{k\in[d]}\left(\frac{\mu^\pi(x=e_k|X_{-\mathcal{S}_1}=E)}{\mu^\pi(x=e_k)}-1\right)\cdot\left(\frac{\mu^\pi(z=e_k|Z_{-\mathcal{S}_2}=E)}{\mu^\pi(z=e_k)}-1\right)\\
&\qquad\qquad\qquad\qquad\qquad\qquad\cdot\mu^\pi(X_{-\mathcal{S}_1}=E)\cdot\mu^\pi(Z_{-\mathcal{S}_2}=E)\cdot\mu^\pi(z=e_k)\Bigg]\\
&\quad\le\frac{1}{2}\mathbb{E}_\pi\left[\sum_{E\in\mathcal{E}}\sum_{k\in[d]}\left(\frac{\mu^\pi(x=e_k|X_{-\mathcal{S}_1})}{\mu^\pi(x=e_k)}-1\right)^2\cdot\mu^\pi(z=e_k)\cdot\big(\mu^\pi(X_{-\mathcal{S}_1}=E)\big)^2\right]\\
&\qquad+\frac{1}{2}\mathbb{E}_\pi\left[\sum_{E\in\mathcal{E}}\sum_{k\in[d]}\left(\frac{\mu^\pi(z=e_k|Z_{-\mathcal{S}})}{\mu^\pi(z=e_k)}-1\right)^2\cdot\mu^\pi(z=e_k)\cdot\big(\mu^\pi(Z_{-\mathcal{S}_2}=E)\big)^2\right].
\end{aligned}\tag{F.21}$$

where in the last inequality, we apply $ab\le a^2+b^2/2$.

Next, for each $E\in\mathcal{E}$, consider $E'=(E'_i)_{i\in\mathcal{S}_1}$ such that

$$E'_{i_h}=E_h,\quad\forall h\in\mathcal{S}.\tag{F.22}$$

Note that for each $E\in\mathcal{E}$, E' must exist and is unique. The existence follows from the compatibility definition, which allows us to index all the unique values in E by restricting the indices to the set $\mathcal{S}_1$. The uniqueness is due to the fact that (F.22) completely determines all the values in E' because enumerating over i_h for $h\in\mathcal{S}$ is just the same as enumerating over i for $i\in\mathcal{S}_1$. In fact, E' contains all the unique values of E. In the above example, we have $\mathcal{S}_1=\{1,2\}$ and thus $E'=(a,b)$ when $E=(a,b,a)$.

Since E' is uniquely defined based on E, we are able to define an operator $\mathcal{J}_1$ that maps $E\in\mathcal{E}$ to $E'\in\mathcal{X}^{|\mathcal{S}_1|}$ according to the mapping given in (F.22). Let $\mathcal{J}_1(\mathcal{E})$ be the image of $\mathcal{E}$ under $\mathcal{J}_1$. It is important to note that for each $E'\in\mathcal{J}_1(\mathcal{E})$, there is also a unique pre-image $E\in\mathcal{E}$ such that $\mathcal{J}_1(E)=E'$ according to the rule (F.22). ***Therefore, $\mathcal{J}_1$ is an one-to-one mapping from $\mathcal{E}$ to $\mathcal{J}_1(\mathcal{E})$.*** In the following, for any $E'\in\mathcal{J}_1(\mathcal{E})$, we denote by $\{X_{-\mathcal{S}_1}=E'\}$ the event where $x_{-i}=E'_i$ for all $i\in\mathcal{S}_1$. Equivalently, we have $x_{-i_h}=E'_{i_h}=E_h$ for all $h\in\mathcal{S}$. Thus, the event $\{X_{-\mathcal{S}_1}=E'\}$ is exactly the same as $\{X_{-\mathcal{S}_1}=E\}$ introduced above. Therefore, the first term on the right hand side of (F.21) can be reformulated as

$$\begin{aligned}
&\frac{1}{2}\cdot\mathbb{E}_\pi\left[\sum_{E\in\mathcal{E}}\sum_{k\in[d]}\left(\frac{\mu^\pi(x=e_k|X_{-\mathcal{S}_1})}{\mu^\pi(x=e_k)}-1\right)^2\cdot\mu^\pi(z=e_k)\cdot\mu^\pi(X_{-\mathcal{S}_1}=E)^2\right]\\
&\quad=\frac{1}{2}\cdot\mathbb{E}_\pi\left[\sum_{E'\in\mathcal{J}_1(\mathcal{E})}\sum_{k\in[d]}\left(\frac{\mu^\pi(x=e_k|X_{-\mathcal{S}_1})}{\mu^\pi(x=e_k)}-1\right)^2\cdot\mu^\pi(z=e_k)\cdot\big(\mu^\pi(X_{-\mathcal{S}_1}=E')\big)^2\right]\\
&\quad\le\frac{1}{2}\cdot\mathbb{E}_\pi\left[\sum_{E'\in\mathcal{X}^{|\mathcal{S}_1|}}\sum_{k\in[d]}\left(\frac{\mu^\pi(x=e_k|X_{-\mathcal{S}_1})}{\mu^\pi(x=e_k)}-1\right)^2\cdot\mu^\pi(z=e_k)\cdot\big(\mu^\pi(X_{-\mathcal{S}_1}=E')\big)^2\right]=\frac{1}{2}\widetilde{I}_{\chi^2}(\mathcal{S}_1),
\end{aligned}$$

where the equality follows from the bijection between $\mathcal{E}$ and $\mathcal{J}_1(\mathcal{E})$, and the last inequality holds by noting that $\mathcal{J}_1(\mathcal{E})\subseteq\mathcal{X}^{|\mathcal{S}_1|}$. The last equality follows from the definition of the modified mutual information. The argument for the second term on the right hand side of (F.21) is similar, and we hence conclude that

$$\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\times\mu^\pi}\left[\prod_{h\in\mathcal{S}}\mathbb{1}(x_{-i_h}=z_{-j_h})\cdot\left(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(z=e_k)}-1\right)\right]\le\frac{1}{2}\widetilde{I}_{\chi^2}(\mathcal{S}_1)+\frac{1}{2}\widetilde{I}_{\chi^2}(\mathcal{S}_2).$$

Lastly, note that $\widetilde{I}_{\chi^2}(\mathcal{S})\le\widetilde{I}_{\chi^2}(\mathcal{S}^\star)$ for any $\mathcal{S}\in[H]_{\le D}$ by the optimality of $\mathcal{S}^\star$. Hence, we complete the proof of Lemma F.7. □

Lemma F.9 quantifies the approximation error from $\sigma_l\approx\sigma_l^\star$ and $y(k)\approx y^\star(k)$ for Stage III.

Lemma F.9. *For the transformer model defined in (2.5), define two quantities f_1 and f_2 as*

$$f_1 := \mathbb{E}\Bigg[\sum_{l=M+1}^{L} \sigma_l \sum_{k=1}^{d}\bigg(\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y(k)+\varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1}=e_k)}{y(k)+\varepsilon}\bigg)\cdot \prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Bigg],$$

$$f_2 := \mathbb{E}\Bigg[\sum_{l=M+1}^{L} \sigma_l^\star \sum_{k=1}^{d}\bigg(\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y^\star(k)+\varepsilon} - \frac{y^\star(k)\,\mathbb{1}(x_{L+1}=e_k)}{y^\star(k)+\varepsilon}\bigg)\cdot \prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\Bigg],$$

where the expectation is taken over all the randomness in the data, and

$$\sigma_l^\star := \frac{\exp\big(a\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\big)}{\sum_{l'=1}^{L}\exp\big(a\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l'-h}=x_{L+1-h})\big)},\quad y^\star(k) := \sum_{l=M+1}^{L}\sigma_l^\star\,\mathbb{1}(x_l=e_k),$$

with $\mathcal{S}^\star$ is the optimal information set. Under Assumption 3.5, it holds that

$$|f_1 - f_2| \le 12\cdot(1+a\varepsilon^{-1})\cdot(\Delta_1+\Delta_2),$$

where $\Delta_1 := 1 - p_{\mathcal{S}^\star}$ and $\Delta_2 := 1 - \prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)})^2$.

Proof. We separate the approximation error into three parts $|f_1 - f_2| \le \mathrm{err}_1 + \mathrm{err}_2 + \mathrm{err}_3$, which are explained in detail as follows.

The First Error Term. Here, the first error err_1 captures the error of replacing $\prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h} = x_{L+1-h})$ with $\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle$ in f_2:

$$\mathrm{err}_1 := \Bigg|\mathbb{E}\Bigg[\sum_{l=M+1}^{L}\sigma_l\cdot\sum_{k\in[d]}\bigg(\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y^\star(k)+\varepsilon} - \frac{y^\star(k)\,\mathbb{1}(x_{L+1}=e_k)}{y^\star(k)+\varepsilon}\bigg)\cdot \\ \cdot\bigg(\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle - \prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\bigg)\Bigg]\Bigg|,$$

Using Lemma F.2, we have

$$\Bigg|\sum_{l=M+1}^{L}\sigma_l\cdot\sum_{k\in[d]}\bigg(\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y^\star(k)+\varepsilon} - \frac{y^\star(k)\,\mathbb{1}(x_{L+1}=e_k)}{y^\star(k)+\varepsilon}\bigg)\Bigg| \le 2.$$

Then using Lemma F.1, we conclude that

$$\mathrm{err}_1 \le 2\sup_{l\in[L]}\Bigg|\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle - \prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\Bigg| \le 2(\Delta_1+\Delta_2).$$

The Second Error Term. The second error term characterizes the difference in σ_l and $\sigma_l^\star$:

$$\mathrm{err}_2 := \Bigg|\mathbb{E}\Bigg[\sum_{l=M+1}^{L}(\sigma_l^\star - \sigma_l)\cdot\sum_{k\in[d]}\bigg(\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y^\star(k)+\varepsilon} - \frac{y^\star(k)\,\mathbb{1}(x_{L+1}=e_k)}{y^\star(k)+\varepsilon}\bigg)\cdot\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Bigg]\Bigg|.$$

To characterize such an error, we invoke equation (53) in Lemma 5.1 of Chen et al. (2022). This lemma states that for σ and $\sigma^\star$ being the output of the softmax function with scaling parameters a for s and $s^\star$ respectively, i.e.,

$$\sigma = \frac{\exp(as)}{\sum_{l=M+1}^{L}\exp(as_l)},\qquad \text{and}\qquad \sigma^\star = \frac{\exp(as^\star)}{\sum_{l=M+1}^{L}\exp(as_l^\star)},$$

it holds that $\|\sigma - \sigma^\star\|_1 \le 4a\cdot\|s - s^\star\|_\infty$. Consequently, we have $\|\sigma - \sigma^\star\|_1 \le 4a\cdot(\Delta_1+\Delta_2)$ by Lemma F.1. We notice that

$$\Bigg|\sum_{k\in[d]}\bigg(\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y^\star(k)+\varepsilon} - \frac{y^\star(k)\,\mathbb{1}(x_{L+1}=e_k)}{y^\star(k)+\varepsilon}\bigg)\cdot\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Bigg| \le \max\{\varepsilon^{-1}, 1\} = \varepsilon^{-1}.$$

Thus, $\mathrm{err}_2 \le 4a\varepsilon^{-1}\cdot(\Delta_1+\Delta_2)$.

https://doi.org/10.52202/079017-2127

The Third Error Term. The last error term characterizes the difference between $y^\star$ and y:

$$\begin{aligned}\mathrm{err}_3 := &\bigg|\mathbb{E}\bigg[\sum_{l=M+1}^{L}\sigma_l\cdot\sum_{k\in[d]}\Big(\frac{1}{y^\star(k)+\varepsilon}-\frac{1}{y(k)+\varepsilon}\Big)\cdot\mathbb{1}(x_{L+1}=x_l=e_k)\cdot\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)},v_{L+1}^{(h)}\rangle\bigg]\bigg| \\ &+\bigg|\mathbb{E}\bigg[\sum_{l=M+1}^{L}\sigma_l\cdot\sum_{k\in[d]}\Big(\frac{y^\star(k)}{y^\star(k)+\varepsilon}-\frac{y(k)}{y(k)+\varepsilon}\Big)\cdot\mathbb{1}(x_{L+1}=e_k)\cdot\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)},v_{L+1}^{(h)}\rangle\bigg]\bigg|.\end{aligned}$$

By noting that $y^\star=\sum_{l=M+1}^{L}\sigma_l^\star x_l$ and $y=\sum_{l=M+1}^{L}\sigma_l x_l$, we have

$$\|y^\star-y\|_1=\bigg\|\sum_{l=M+1}^{L}(\sigma_l^\star-\sigma_l)x_l\bigg\|_1\le\sum_{l=M+1}^{L}|\sigma_l^\star-\sigma_l|_1\cdot\|x_l\|_1\le\|\sigma-\sigma^\star\|_1\le 4a\cdot(\Delta_1+\Delta_2).$$

The first term of err_3 can be bounded by

$$\begin{aligned}&\bigg|\mathbb{E}\bigg[\sum_{l=M+1}^{L}\sigma_l\cdot\sum_{k\in[d]}\Big(\frac{1}{y^\star(k)+\varepsilon}-\frac{1}{y(k)+\varepsilon}\Big)\cdot\mathbb{1}(x_{L+1}=x_l=e_k)\cdot\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)},v_{L+1}^{(h)}\rangle\bigg]\bigg| \\ &\quad\le\sum_{k\in[d]}\frac{|y(k)-y^\star(k)|}{(y^\star(k)+\varepsilon)(y(k)+\varepsilon)}\cdot y(k)\,\mathbb{1}(x_{L+1}=e_k)\le\|y-y^\star\|_1\cdot\varepsilon^{-1}\le 4a\varepsilon^{-1}(\Delta_1+\Delta_2).\end{aligned}$$

Moreover, for the second term of err_3, we have

$$\begin{aligned}&\bigg|\mathbb{E}\bigg[\sum_{l=M+1}^{L}\sigma_l\cdot\sum_{k\in[d]}\Big(\frac{y^\star(k)}{y^\star(k)+\varepsilon}-\frac{y(k)}{y(k)+\varepsilon}\Big)\cdot\mathbb{1}(x_{L+1}=e_k)\cdot\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)},v_{L+1}^{(h)}\rangle\bigg]\bigg| \\ &\quad\le\sum_{k\in[d]}\frac{|y(k)-y^\star(k)|}{(y^\star(k)+\varepsilon)(y(k)+\varepsilon)}\cdot\varepsilon\cdot\mathbb{1}(x_{L+1}=e_k)\le 4a\varepsilon^{-1}(\Delta_1+\Delta_2).\end{aligned}$$

It then holds that

$$\begin{aligned}|f_1-f_2|&\le\mathrm{err}_1+\mathrm{err}_2+\mathrm{err}_3\le 2(\Delta_1+\Delta_2)+4a\varepsilon^{-1}(\Delta_1+\Delta_2)+8a\varepsilon^{-1}(\Delta_1+\Delta_2)\\&=12\cdot(1+a\varepsilon^{-1})\cdot(\Delta_1+\Delta_2).\end{aligned}$$

Therefore, we complete the proof of Lemma F.9. □

Lemma F.10. *Let us define for brevity,*

$$\widetilde{\mu}_X^\pi(z,Z)=\widetilde{\mu}^\pi(z,Z\,|\,X_{L+1-\mathcal{S}^\star})=\frac{\mu^\pi(z,Z)\exp\big(a\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(z_{-h}=x_{L+1-h})\big)}{\sum_{z',Z'}\mu^\pi(z',Z')\exp\big(a\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(z'_{-h}=x_{L+1-h})\big)},$$

where $Z=(z_{-M},\ldots,z_{-1})$ and μ^π is the stationary distribution of the Markov chain over a window of size $M+1$. We denote by $\widetilde{\mu}_X^\pi(e_k)=\widetilde{\mu}_X^\pi(z=e_k)$ where $\widetilde{\mu}_X^\pi(z)$ is the marginal distribution for z and serves as the population counterpart for $y^\star=\sum_{l=M+1}^{L}\sigma_l^\star x_l$. We define quantity A and B as

$$A:=\mathbb{E}\bigg[\sum_{l=M+1}^{L}\sigma_l^\star\sum_{k=1}^{d}\Big(\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y^\star(k)+\varepsilon}-\frac{y^\star(k)\,\mathbb{1}(x_{L+1}=e_k)}{y^\star(k)+\varepsilon}\Big)\prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\bigg].$$

$$B:=\mathbb{E}\bigg[\sum_{l=M+1}^{L}\sigma_l^\star\sum_{k=1}^{d}\Big(\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\widetilde{\mu}_X^\pi(e_k)}-\mathbb{1}(x_{L+1}=e_k)\Big)\prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\bigg].$$

Under Assumption 3.5, we have

$$|A-B|\le\frac{8(1-\lambda)^{-1/2}(D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1)^{1/4}+8\sqrt{M}}{L^{1/2}\cdot\gamma^{|\mathcal{S}^\star|+1}}+\frac{2d\varepsilon}{\gamma}.$$

Proof of Lemma F.10. The proof follows the same arguments as Lemma F.4. We remind the readers that $y^\star(k)$ is also a function of the whole chain X. We note that

$$
\begin{aligned}
|A-B| &= \Bigg|\mathbb{E}\Bigg[\sum_{l=M+1}^{L}\sigma_l^\star\cdot\Bigg(\sum_{k\in[d]}\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y^\star(k)+\varepsilon}-\sum_{k\in[d]}\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\widetilde{\mu}_X^\pi(e_k)} \\
&\qquad -\sum_{k\in[d]}\frac{y^\star(k)\,\mathbb{1}(x_{L+1}=e_k)}{y^\star(k)+\varepsilon}+1\Bigg)\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\Bigg]\Bigg| \\
&= \Bigg|\mathbb{E}\Bigg[\sum_{l=M+1}^{L}\sigma_l^\star\cdot\Bigg(\sum_{k\in[d]}\Bigg(\frac{\widetilde{\mu}_X^\pi(e_k)-y^\star(k)}{(y^\star(k)+\varepsilon)\cdot\widetilde{\mu}_X^\pi(e_k)}-\frac{\varepsilon}{(y^\star(k)+\varepsilon)\cdot\widetilde{\mu}_X^\pi(e_k)}\Bigg)\cdot\mathbb{1}(x_{L+1}=x_l=e_k) \\
&\qquad -\sum_{k\in[d]}\frac{\varepsilon\,\mathbb{1}(x_{L+1}=e_k)}{y^\star(k)+\varepsilon}\Bigg)\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\Bigg]\Bigg|.
\end{aligned}
$$

To handle this error, we define three error terms as

$$
\begin{aligned}
\mathrm{err}_1 &:= \Bigg|\mathbb{E}\Bigg[\sum_{k\in[d]}\frac{\widetilde{\mu}_X^\pi(e_k)-y^\star(k)}{(y^\star(k)+\varepsilon)\cdot\widetilde{\mu}_X^\pi(e_k)}\cdot\sum_{l=M+1}^{L}\sigma_l^\star\cdot\mathbb{1}(x_{L+1}=x_l=e_k)\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\Bigg]\Bigg|, \\
\mathrm{err}_2 &:= \Bigg|\mathbb{E}\Bigg[\sum_{k\in[d]}\frac{\varepsilon}{(y^\star(k)+\varepsilon)\cdot\widetilde{\mu}_X^\pi(e_k)}\cdot\sum_{l=M+1}^{L}\sigma_l^\star\cdot\mathbb{1}(x_{L+1}=x_l=e_k)\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\Bigg]\Bigg|, \\
\mathrm{err}_3 &:= \Bigg|\mathbb{E}\Bigg[\sum_{k\in[d]}\frac{\varepsilon}{y^\star(k)+\varepsilon}\cdot\mathbb{1}(x_{L+1}=e_k)\cdot\sum_{l=M+1}^{L}\sigma_l^\star\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\Bigg]\Bigg|.
\end{aligned}
$$

For the first error term, we have that

$$
\begin{aligned}
\mathrm{err}_1 &\le \mathbb{E}\Bigg[\sum_{k\in[d]}\frac{|\widetilde{\mu}_X^\pi(e_k)-y^\star(k)|}{(y^\star(k)+\varepsilon)}\cdot\sum_{l=M+1}^{L}\frac{\sigma_l^\star\,\mathbb{1}(x_l=e_k)}{\widetilde{\mu}_X^\pi(e_k)}\Bigg] \\
&= \mathbb{E}\Bigg[\sum_{k\in[d]}\frac{|\widetilde{\mu}_X^\pi(e_k)-y^\star(k)|}{(y^\star(k)+\varepsilon)}\cdot\frac{y^\star(k)}{\widetilde{\mu}_X^\pi(e_k)}\Bigg]\le\gamma^{-1}\cdot\mathbb{E}\Bigg[\sum_{k\in[d]}|\widetilde{\mu}_X^\pi(e_k)-y^\star(k)|\Bigg],
\end{aligned}
$$

where we recall that by assumption, γ provides a lower bound for $\pi(\cdot\,|\,X_{\mathtt{pa}})$, hence also a lower bound for $\widetilde{\mu}_X^\pi(e_k)$. Next, we invoke Proposition F.19 which provides an upper bound for the difference between the empirical and population distributions in terms of the ℓ_1-norm:

$$
\begin{aligned}
\mathbb{E}\Big[\Big\|\widetilde{\mu}_X^\pi(z=\cdot)-y^\star(\cdot)\Big\|_1\Big] &\le \frac{4\big((1-\lambda)^{-1}\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}+4M\big)^{1/2}}{L^{1/2}\cdot\min_{\pi,x_{L+1},X_{L+1-\mathcal{S}^\star}}\mu^\pi(x_{L+1},X_{L+1-\mathcal{S}^\star})} \\
&\le \frac{4(1-\lambda)^{-1/2}(D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1)^{1/4}+8\sqrt{M}}{L^{1/2}\cdot\gamma^{|\mathcal{S}^\star|+1}}.
\end{aligned}
\tag{F.23}
$$

Hence, we control the first error term.

For the second error term, we follow the same procedure and obtain an upper bound as

$$
\mathrm{err}_2\le\mathbb{E}\Bigg[\sum_{k\in[d]}\frac{\varepsilon}{\widetilde{\mu}_X^\pi(e_k)}\cdot\sum_{l=M+1}^{L}\frac{\sigma_l^\star\,\mathbb{1}(x_l=e_k)}{(y^\star(k)+\varepsilon)}\Bigg]\le\mathbb{E}\Bigg[\sum_{k\in[d]}\frac{\varepsilon}{\widetilde{\mu}_X^\pi(e_k)}\Bigg]\le\gamma^{-1}d\varepsilon.
$$

https://doi.org/10.52202/079017-2127

For the last error term, it holds that

$$
\begin{aligned}
\mathrm{err}_3 &\le \mathbb{E}\Big[\sum_{k\in[d]} \frac{\varepsilon}{y^\star(k)+\varepsilon}\cdot \mathbb{1}(x_{L+1}=e_k)\Big] \\
&\le \Big|\mathbb{E}\Big[\sum_{k\in[d]} \frac{\varepsilon\,\mathbb{1}(x_{L+1}=e_k)}{\widetilde{\mu}^\pi_X(e_k)+\varepsilon}\Big]\Big| + \Big|\sum_{k\in[d]}\mathbb{E}\Big[\frac{\varepsilon(y^\star(k)-\widetilde{\mu}^\pi_X(e_k))\cdot\mathbb{1}(x_{L+1}=e_k)}{(\widetilde{\mu}^\pi_X(e_k)+\varepsilon)(y^\star(k)+\varepsilon)}\Big]\Big| \\
&\le \frac{\varepsilon}{\gamma} + \mathbb{E}\Big[\sum_{k\in[d]}\frac{|y^\star(k)-\widetilde{\mu}^\pi_X(e_k)|}{\gamma}\Big] \le \frac{\varepsilon}{\gamma} + \frac{4(1-\lambda)^{-1/2}(D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1)^{1/4}+8\sqrt{M}}{L^{1/2}\cdot\gamma^{|\mathcal{S}^\star|+1}}.
\end{aligned}
$$

where the last inequality follows directly from (F.23).

In summary, the difference between f_2 and f_3 is bounded by

$$
|f_2-f_3| \le \mathrm{err}_1+\mathrm{err}_2+\mathrm{err}_3 \le \frac{8(1-\lambda)^{-1/2}(D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1)^{1/4}+8\sqrt{M}}{L^{1/2}\cdot\gamma^{|\mathcal{S}^\star|+1}} + \frac{2d\varepsilon}{\gamma},
$$

which completes our proof of Lemma F.10. □

The following lemmas are for analyzing the error $|f_3-f_4|$ for Stage III.

Lemma F.11. *We define*

$$
\begin{aligned}
A &:= \mathbb{E}\Big[\sum_{l=M+1}^{L}\sigma_l^\star\cdot\sum_{k=1}^{d}\Big(\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\widetilde{\mu}^\pi_X(e_k)} - \mathbb{1}(x_{L+1}=e_k)\Big)\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\Big], \\
B &:= \mathbb{E}_{X,(z,Z)\sim\widetilde{\mu}^\pi_X}\Big[\sum_{k=1}^{d}\Big(\frac{\mathbb{1}(x_{L+1}=z=e_k)}{\widetilde{\mu}^\pi_X(e_k)} - \mathbb{1}(x_{L+1}=e_k)\Big)\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(z_{l-h}=x_{L+1-h})\Big],
\end{aligned}
$$

where

$$
\begin{aligned}
\sigma_l^\star &:= \frac{\exp\big(a\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\big)}{\sum_{l'=1}^{L}\exp\big(a\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l'-h}=x_{L+1-h})\big)}, \\
\widetilde{\mu}^\pi_X(z,Z) &:= \widetilde{\mu}^\pi(z,Z\,|\,X_{L+1-\mathcal{S}^\star}) = \frac{\mu^\pi(z,Z)\exp\big(a\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(z_{-h}=x_{L+1-h})\big)}{\sum_{z',Z'}\mu^\pi(z',Z')\exp\big(a\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(z'_{-h}=x_{L+1-h})\big)}.
\end{aligned}
$$

Under Assumption 3.5, we have

$$
|A-B| \le \frac{8\gamma^{-1}(1-\lambda)^{-1/2}(D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1)^{1/4}+16\gamma^{-1}\sqrt{M}}{L^{1/2}\cdot\gamma^{|\mathcal{S}^\star|+1}}.
$$

Proof of Lemma F.11. For $Z=(z_{-M},\ldots,z_{-1})$ and $Z'=(z'_{-M},\ldots,z'_{-1})$, we let $Z_{-\mathcal{S}^\star}=(z_{-h})_{h\in\mathcal{S}^\star}$, we define

$$
\begin{aligned}
\widehat{\mu}^\pi_X(z,Z) &= \frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{1}(x_l=z, X_{l-M:l-1}=Z), \\
R(Z,X_{L+1-\mathcal{S}^\star}) &= \exp\Big(a\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(z_{-h}=x_{L+1-h})\Big).
\end{aligned}
$$

Using these notations, we can rewrite the normalizing factor in $\widetilde{\mu}^\pi_X$ and $\sigma_l^\star$ respectively as

$$
\Phi = \sum_{z,Z}\mu^\pi(z,Z)\cdot R(Z,X_{L+1-\mathcal{S}^\star}), \quad \widehat{\Phi} = \sum_{z,Z}\widehat{\mu}^\pi_X(z,Z)\cdot R(Z,X_{L+1-\mathcal{S}^\star}).
$$

We also define

$$
\phi(z,Z_{-\mathcal{S}^\star}) = \mu^\pi(z,Z_{-\mathcal{S}^\star})\cdot R(Z_{-\mathcal{S}^\star},X_{-\mathcal{S}^\star}), \quad \widehat{\phi}(z,Z_{-\mathcal{S}^\star}) = \widehat{\mu}^\pi_X(z,Z_{-\mathcal{S}^\star})\cdot R(Z_{-\mathcal{S}^\star},X_{L+1-\mathcal{S}^\star}).
$$

https://doi.org/10.52202/079017-2127

If we further define $\widehat{\nu}_X^\pi(z, Z_{-\mathcal{S}^\star}) = \sum_{l=M+1}^{L} \mathbb{1}(x_l = z, X_{l-\mathcal{S}^\star} = Z_{-\mathcal{S}^\star})$, then we have

$$\widehat{\nu}_X^\pi(z, Z_{-\mathcal{S}^\star}) = \frac{\widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star}) \cdot R(Z_{-\mathcal{S}^\star}, X_{L+1-\mathcal{S}^\star})}{\widehat{\Phi}} = \frac{\widehat{\phi}(z, Z_{-\mathcal{S}^\star})}{\widehat{\Phi}},$$
$$\widetilde{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star}) = \frac{\mu^\pi(z, Z_{-\mathcal{S}^\star}) \cdot R(Z_{-\mathcal{S}^\star}, X_{L+1-\mathcal{S}^\star})}{\Phi} = \frac{\phi(z, Z_{-\mathcal{S}^\star})}{\Phi}.$$

Using the above definitions and relationship, A and B can be rewritten as

$$A = \mathbb{E}\bigg[\sum_{k=1}^{d} \frac{\widehat{\phi}(e_k, X_{L+1-\mathcal{S}^\star})}{\widehat{\Phi} \cdot \widetilde{\mu}_X^\pi(e_k)} - \frac{\widehat{\phi}(X_{L+1-\mathcal{S}^\star})}{\widehat{\Phi}}\bigg], \quad B = \mathbb{E}\bigg[\sum_{k=1}^{d} \frac{\phi(e_k, X_{L+1-\mathcal{S}^\star})}{\Phi \cdot \widetilde{\mu}_X^\pi(e_k)} - \frac{\phi(X_{L+1-\mathcal{S}^\star})}{\Phi}\bigg].$$

Therefore, the difference between A and B is given by

$$|A - B| \le \frac{2}{\gamma} \cdot \mathbb{E}\bigg[\sum_{z, Z_{-\mathcal{S}^\star}} \bigg|\frac{\phi(z, Z_{-\mathcal{S}^\star})}{\Phi} - \frac{\widehat{\phi}(z, Z_{-\mathcal{S}^\star})}{\widehat{\Phi}}\bigg|\bigg] \le \frac{2}{\gamma} \cdot \mathbb{E}\bigg[\sum_{z, Z_{-\mathcal{S}^\star}} \Big|\widetilde{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star}) - \widehat{\nu}_X^\pi(z, Z_{-\mathcal{S}^\star})\Big|\bigg]$$
$$\le \frac{8\gamma^{-1} \cdot \big((1-\lambda)^{-1}\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1} + 4M\big)^{1/2}}{L^{1/2} \cdot \min_{x_{L+1}, X_{L+1-\mathcal{S}^\star}} \mu^\pi(x_{L+1}, X_{L+1-\mathcal{S}^\star})}.$$

where the last inequality follows from the result in Proposition F.19. Invoking the lower bound $\mu^\pi(x_{L+1}, X_{L+1-\mathcal{S}^\star}) \ge \gamma^{|\mathcal{S}^\star|+1}$, we complete the proof of Lemma F.11. □

F.3 Lemmas on Concentration of Markov Chain

Recall that we previously define $X = (x_1, \ldots, x_L)$ as the observed sequence and x_{L+1} as the value at time $L+1$ to be predicted. For generality, we will use $X = (x_1, \ldots, x_{L+1})$ to denote the whole sequence in the following proof. We denote by $p^\pi(\cdot)$ the joint distribution for the sequence X with kernel π. Recall that we have the parent set $\mathtt{pa} = \{-r_1, \ldots, -r_n\}$, and as the start of a chain, we sample the first r_n tokens by $(x_1, \ldots, x_{r_n}) \sim \mu_0$.

In the sequel, we will study concentration properties of the Markov chain X for a window of tokens with window size at most M, where $M > r_n$. To proceed, let us consider a fixed set $\mathcal{S} \subseteq [M]$. For any $l \in [M+1, L+1]$, we define $Y_l = (x_l, X_{l-\mathcal{S}})$ as a new vector containing the token at position l and also the tokens in the past $\mathcal{S}$ positions prior to x_l. Here, we follow the convention that $X_{l-\mathcal{S}} = (X_{l-i})_{i \in \mathcal{S}}$. We also consider another fixed subset $\mathcal{S}' \subseteq [M]$ and similarly define $Y_l' = (x_l, X_{l-\mathcal{S}'})$.

The concentration properties of the Markov chain are rooted in the fact that when conditioning on all the parents, the current token is independent of all the past tokens. Given the parent set structure $\mathtt{pa} = \{-r_1, \ldots, -r_n\}$, we aim to make Y_{L+1} approximately independent of Y_l by conditioning on some intermediate parent sets. To this end, we define $A = (x_{L+1-M}, \ldots, x_{L-M+r_n}) \in \mathcal{X}^{r_n}$ and $B_l = (x_{l-r_n+1}, \ldots, x_l) \in \mathcal{X}^{r_n}$ as these intermediate parent sets. By the Markov property and the parent set structure, we have the following conditional independence relations:

$$Y_{L+1} \perp\!\!\!\perp (B_l, Y_l) \,|\, A, \quad (Y_{L+1}, A) \perp\!\!\!\perp Y_l \,|\, B_l, \quad \forall l = M+1, \ldots, L-M+r_n.$$

To illustrate, let us consider the first condition $Y_{L+1} \perp\!\!\!\perp (B_l, Y_l) \,|\, A$. When $l \le L - M + r_n$, the B_l and Y_l are both contained in the history $\{x_k \colon k \le L - M + r_n\} = A \cup \{x_k \colon k \le L - M\}$. When conditioning on A, the randomness of (B_l, Y_l) is measurable by the σ-algebra generated by the "past" $\{x_k \colon k \le L - M\}$. Moreover, the randomness of Y_{L+1} is measurable by the σ-algebra generated by the "future" $\{x_k \colon k \in [L+1-M+r_n, L+1]\}$ when conditioning on A. Notice that the parent to the any element in the future $\{x_k \colon k \in [L+1-M+r_n, L+1]\}$ is either contained in A, or can be generated conditioned on A without touching further history $\{x_k \colon k \le L - M\}$. Thus, by the Markov property, conditioning on A, Y_{L+1} is independent of the past $\{x_k \colon k \le L - M\}$, and in particular, (B_l, Y_l). Similarly, since B contains the parent of x_{l+1}, conditioning on B, Y_l is independent of x_{l+1} and later tokens. Moreover, given B, the randomness of Y_l comes from the randomness of $x_{l-M}, \ldots, x_{l-r_n}$. Since $l \le L - M + r_n$, we have $L + 1 - M \ge l + 1 - r_n$. As a result, conditioning on B, the randomness of (Y_{L+1}, A) comes from tokens generated no earlier than x_{l+1}. Therefore, (Y_{L+1}, A) and Y_l are conditionally independent given B_l. We visualize the definition of Y_{L+1}, A, B_l, and Y_l in Figure 10

https://doi.org/10.52202/079017-2127

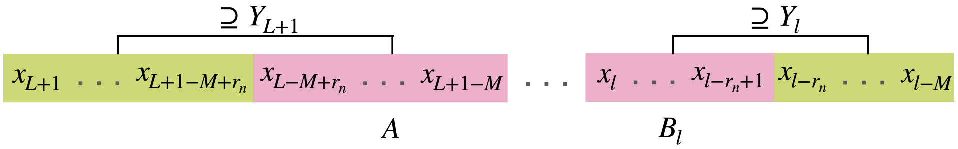

$\supseteq Y_{L+1}$
$\supseteq Y_l$
x_{L+1} . . . $x_{L+1-M+r_n}$
x_{L-M+r_n} . . . x_{L+1-M}
. . .
x_l . . . x_{l-r_n+1}
x_{l-r_n} . . . x_{l-M}
A
B_l

Figure 10: Illustration of the definition of Y_{L+1}, A, B_l, and Y_l. When conditioned on A, Y_{L+1} is independent of (B_l, Y_l). When conditioned on B_l, Y_l is independent of (A, Y_{L+1}).

Similarly, for $Y_l' = (x_l, X_{l-\mathcal{S}'})$ defined using the subset $\mathcal{S}'$, we also parallel conditional independence relations:

$$Y_{L+1}' \perp\!\!\!\perp (B_l, Y_l') \,|\, A, \quad (Y_{L+1}', A) \perp\!\!\!\perp Y_l' \,|\, B_l, \quad \forall l = M+1, \ldots, L-M+r_n.$$

In particular, we also have

$$Y_{L+1} \perp\!\!\!\perp (B_l, Y_l') \,|\, A, \quad (Y_{L+1}, A) \perp\!\!\!\perp Y_l' \,|\, B_l, \quad \forall l = M+1, \ldots, L-M+r_n. \tag{F.24}$$

Using $\{Y_l, Y_l'\}$, we define a joint distribution $\widehat{p}^\pi$ over $2 + |\mathcal{S}| + |\mathcal{S}'|$ tokens as follows. For any $E \in \mathcal{X}^{|\mathcal{S}|+1}$ and $E' \in \mathcal{X}^{|\mathcal{S}'|+1}$, the probability mass function of $\widehat{p}^\pi$ is defined as

$$\begin{aligned} &\widehat{p}^\pi(Y_{L+1} = E, Y' = E') \\ &:= \frac{1}{L-M} \sum_{l=M+1}^{L} p^\pi(Y_{L+1} = E, Y_l' = E') \\ &= \frac{1}{L-M} \sum_{l=M+1}^{L} \sum_{A, B_l} \mu^\pi(Y_{L+1} = E \,|\, A) \cdot P_\pi^{L-M+r_n-l}(A \,|\, B_l) \cdot p^\pi(Y_l' = E' \,|\, B_l) \cdot p^\pi(B_l). \end{aligned} \tag{F.25}$$

Here, Y' is just a placeholder for Y_l' as $\widehat{p}$ takes an average over l and does not depend on any specific position index. The summation $\sum_{A,B_l}$ means we sum over all possible values that A and B_l can take. In the last line of (F.25), we decompose the joint distribution $p^\pi(Y_{L+1} = E, Y_l' = E')$ into the product of the conditional distributions by the Markov property in (F.24). That is,

$$\begin{aligned} p^\pi(Y_{L+1} = E, Y_l' = E') &= \sum_{A,B_l} p^\pi(Y_{L+1} = E, Y_l' = E', A, B_l) \\ &= \sum_{A,B_l} p^\pi(B_l) \cdot p^\pi(Y_{L+1} = E, A \,|\, B_l) \cdot p^\pi(Y_l = E' \,|\, B_l) \\ &= \sum_{A,B_l} p^\pi(Y_{L+1} = E \,|\, A) \cdot p^\pi(A \,|\, B_l) \cdot p^\pi(Y_l = E' \,|\, B_l) \cdot p^\pi(B_l). \end{aligned}$$

Here the second equality follows from the fact that $(Y_{L+1}, A) \perp\!\!\!\perp Y_l' \,|\, B_l$ and the last equality follows from the fact that $Y_{L+1} \perp\!\!\!\perp (B_l, Y_l') \,|\, A$, which implies $p^\pi(Y_{L+1} = E \,|\, A, B_l) = p^\pi(Y_{L+1} = E \,|\, A)$. Moreover, we denote by P_π^i the i-step transition kernel of the chain, which corresponds to the i-th power of the transition matrix P_π. Here, we are following the convention in the main text that

$$P_\pi(Z', Z) = \pi(z_l' \mid Z_{\mathsf{pa}(l)}) \cdot \mathbf{1}(Z_{l-r_n+1:-1}' = Z_{l-r_n+1:-1}). \tag{F.26}$$

In the following, we always consider a fixed transition kernel π and omit the superscript/subscript π in the matrix notation. We denote the transition matrix by P_π and the stationary distribution by μ^π for a window of length r_n. For the transition matrix, we index each row by the next r_n-window Z' and each column by the current r_n-window Z. Under this notation, since both A and B_l have lengths r_n, we have

$$p^\pi(A \,|\, B_l) = P_\pi^{L-M+r_n-l}(A, B_l). \tag{F.27}$$

Here $P_\pi^{L-M+r_n-l}(A \,|\, B_l)$ corresponds to the (A, B_l)-entry of the matrix $(P_\pi)^{L-M+r_n-l}$. Combining (F.24) and (F.27), we obtain the last equality in (F.25).

In the sequel, to simplify the notation, we write P_π and μ^π as P and μ respectively. Let us consider the reweighted transition kernel

$$K := \mathrm{diag}\big(\sqrt{\mu}\big)^{-1} \cdot P \cdot \mathrm{diag}\left(\sqrt{\mu}\right),$$

where $\sqrt{\mu}$ is the element-wise square root of μ. Since the transition matrix is *primitive* by assumption and having only one eigenvalue with value one on its spectral circle, we also have for K that the leading eigenvalue is one with eigenvector $\sqrt{\mu}$, i.e. $\sqrt{\mu} = K\sqrt{\mu}$ and $\sqrt{\mu}^\top = \sqrt{\mu}^\top K$. However, the projection in the leading eigenspace (or the Perron projection) is not of our interest. The following property of K will be useful in the subsequent proof.

Proposition F.12. *For the reweighted transition matrix K, we have for any integer $i \geq 0$*

$$P^i - \mu\mathbf{1}^\top = \mathrm{diag}\left(\sqrt{\mu}\right) \cdot \left(K - \sqrt{\mu}\sqrt{\mu}^\top\right)^i \cdot \mathrm{diag}\big(\sqrt{\mu}^{-1}\big)$$

Proof of Proposition F.12.

$$\begin{aligned} P^i - \mu\mathbf{1}^\top &= \Big(\mathrm{diag}\big(\sqrt{\mu}\big) \cdot K \cdot \mathrm{diag}\big(\sqrt{\mu}\big)^{-1}\Big)^i - \mu\mathbf{1}^\top \\ &= \mathrm{diag}\big(\sqrt{\mu}\big) \cdot \Big(K^i - \sqrt{\mu}\sqrt{\mu}^\top\Big) \cdot \mathrm{diag}\big(\sqrt{\mu}\big)^{-1} \\ &= \mathrm{diag}\left(\sqrt{\mu}\right) \cdot \left(K - \sqrt{\mu}\sqrt{\mu}^\top\right)^i \cdot \mathrm{diag}\big(\sqrt{\mu}^{-1}\big), \end{aligned}$$

where the last equality holds by noting that $K - \sqrt{\mu}\sqrt{\mu}^\top$ project $\sqrt{\mu}$ to the zero vector, and for any $v \perp \sqrt{\mu}$, we have $(K - \sqrt{\mu}\sqrt{\mu}^\top)v = Kv$. Thus for any test vector x:

$$\begin{aligned} (K - \sqrt{\mu}\sqrt{\mu}^\top)^i x &= (K - \sqrt{\mu}\sqrt{\mu}^\top)^i (x - \langle\sqrt{\mu}, x\rangle \cdot \sqrt{\mu}) \\ &= K^i(x - \langle\sqrt{\mu}, x\rangle \cdot \sqrt{\mu}) = (K^i - \sqrt{\mu}\sqrt{\mu}^\top)x. \end{aligned}$$

This completes the proof of Proposition F.12. □

Indeed, the second largest eigenvalue of K (in magnitude) determines the mixing rate of the chain. Let λ denote the eigenvalue of K with the second largest magnitude.

Furthermore, if the transition kernel π admits a lower bound $\gamma > 0$, then we can guarantee that both p^π and μ^π admit a uniform lower bound.

Proposition F.13 (Uniform Lower Bound)**.** *Suppose $\pi(\cdot \mid X_{\texttt{pa}}) \geq \gamma$ uniformly for some $\gamma > 0$ and $\texttt{pa} = \{-r_1, \ldots, -r_n\}$. Suppose $X_{1:r_n} \sim \mu_0(\cdot)$ where $\mu_0 \in \Delta(\mathcal{X}^{r_n})$. Then for any S tokens $x_{l_1}, x_{l_2}, \ldots, x_{l_S}$ such that $l_s \geq r_n$ for any $s \in [S]$, we have*

$$p^\pi(x_{l_1}, \ldots, x_{l_S}) \geq \gamma^S.$$

Using Proposition F.13, we show that the transition matrix P_π is primitive.

Corollary F.14 (Uniform Lower Bound Implies Primitive Transition)**.** *Under the condition of Proposition F.13, with $\pi(\cdot \mid X_{\texttt{pa}}) \geq \gamma > 0$, the transition matrix defined in (F.26) is primitive.*

Proof of Corollary F.14. If the initial distribution is set to be any one-hot vector in $\Delta(\mathcal{X}^{r_n})$, and taking $x_{l_1}, \ldots, x_{l_S}$ in Proposition F.13 to be $x_{r_n+1}, \ldots, x_{2r_n}$, we conclude that $p^\pi(X_{r_n+1:2r_n} \mid X_{1:r_n}) > 0$ holds for any $X_{r_n+1:2r_n}, X_{1:r_n} \in \mathcal{X}^{r_n}$. Recall from the definition that for a primitive matrix P, we can find some positive integer k such that P^k has all positive entries. For our case, we can set $k = r_n$ and everything follows by noting that $p^\pi(X_{r_n+1:2r_n} \mid X_{1:r_n}) = P_\pi^{r_n}(X_{r_n+1:2r_n}, X_{1:r_n})$. □

Another corollary of Proposition F.13 is that, if we take $\mu_0 = \mu^\pi$, which is the stationary distribution, we can replace p^π in Proposition F.13 by μ^π.

Corollary F.15. *Suppose $\pi(\cdot \mid X_{\texttt{pa}}) \geq \gamma$ uniformly for some $\gamma > 0$ and $\texttt{pa} = \{-r_1, \ldots, -r_n\}$. For the stationary distribution μ^π and S tokens $x_{l_1}, x_{l_2}, \ldots, x_{l_S}$ such that $l_s \geq r_n$ for any $s \in [S]$, we have $\mu^\pi(x_{l_1}, \ldots, x_{l_S}) \geq \gamma^S$.*

We prove Proposition F.13 as follows.

https://doi.org/10.52202/079017-2127

Proof of Proposition F.13. Without loss of generality, suppose that $M \le l_1 < l_2 < \ldots < l_S$. We will prove the statement by induction on the number of tokens S. If $S = 1$, we can rewrite

$$p^\pi(x_{l_1}) = \sum_{X_{\texttt{pa}(l_1)}} \pi(x_{l_1} \,|\, X_{\texttt{pa}(l_1)}) p^\pi(X_{\texttt{pa}(l_1)}) \ge \sum_{X_{\texttt{pa}(l_1)}} \gamma \cdot p^\pi(X_{\texttt{pa}(l_1)}) \ge \gamma.$$

Now, suppose the statement holds for $1, 2, \ldots, S-1$. Let $Y = x_{l_1}, \ldots, x_{l_{s-1}}$. Then, we have

$$\begin{aligned} p^\pi(x_{l_1}, \ldots, x_{l_S}) &= \sum_{X_{\texttt{pa}(l_S)} \setminus Y} \pi(x_{l_S} \,|\, X_{\texttt{pa}(l_S)}) \cdot p^\pi(Y) \cdot p^\pi(X_{\texttt{pa}(l_S)} \setminus Y) \\ &\ge \sum_{X_{\texttt{pa}(l_S)} \setminus Y} \gamma \cdot p^\pi(Y) \cdot p^\pi(X_{\texttt{pa}(l_S)} \setminus Y) = \gamma \cdot p^\pi(Y) \ge \gamma^S, \end{aligned}$$

where the last inequality holds by the induction condition. Hence, we finish the proof. □

Before analyzing $\widehat{p}^\pi$, we first study a simpler convergence result: quantifying the closeness between $\sum_{l=M+1}^{L} \eta^{L-l} p^\pi(B_l = b) / \sum_{l=M+1}^{L} \eta^{L-l}$ and $\mu^\pi(b)$ for certain values of $\eta \in (0, 1]$.

Lemma F.16. *Following the notations introduced above, for the Markov chain with parent set* $\texttt{pa} = \{-r_1, \ldots, -r_n\}$*, let* $D_{\chi^2}(\mu_0 \,\|\, \mu^\pi)$ *be the* χ^2*-divergence between the initial distribution* μ_0 *and the stationary distribution* μ^π *over the first* r_n *tokens. Take any* $\mathcal{S} \subseteq [M]$ *and let* $Y_l = (x_l, X_{l-\mathcal{S}})$ *for* $l = M+1, \ldots, L+1$*. Suppose* $L/2 \ge M \ge r_n$*. We have*

$$\left\| \frac{\sum_{l=M+1}^{L} p^\pi(Y_l = \cdot)}{L - M} - \mu^\pi(Y_{L+1} = \cdot) \right\|_{\mathrm{TV}} \le \frac{2\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda)}, \tag{F.28}$$

$$\| p^\pi(Y_{L+1} = \cdot) - \mu^\pi(Y_{L+1} = \cdot) \|_{\mathrm{TV}} \le \lambda^{L-M} \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}. \tag{F.29}$$

Proof of Lemma F.16. Let $c_l = \eta^{L-l} / \sum_{l=r_n}^{L-M+r_n} \eta^{L-l}$, where $\eta \in [0, 1]$ is a constant to be determined. Denote by μ_0, a vector of length $|\mathcal{X}|^{r_n}$, the initial distribution of the chain. We begin by quantifying the total variation (TV) distance:

$$\left\| \frac{\sum_{l=r_n}^{L-M+r_n} \lambda^{L-l} p^\pi(B_l = \cdot)}{\sum_{l=r_n}^{L-M+r_n} \lambda^{L-l}} - \mu^\pi(\cdot) \right\|_{\mathrm{TV}} = \left\| \sum_{l=r_n}^{L-M+r_n} c_l \cdot \left(p^\pi(B_l = \cdot) - \mu^\pi(\cdot) \right) \right\|_{\mathrm{TV}}.$$

Let $b \in \mathcal{X}^{r_n}$, representing the value for a length-r_n window. Using matrix notation, we have:

$$\begin{aligned} \sum_{l=r_n}^{L-M+r_n} c_l \left(p^\pi(B_l = b) - \mu^\pi(b) \right) &= \sum_{l=r_n}^{L-M+r_n} c_l \cdot \mathbf{1}_b^\top P^{l-r_n} (\mu_0 - \mu) = \sum_{l=r_n}^{L-M+r_n} c_l \cdot \mathbf{1}_b^\top (P^{l-r_n} - \mu \mathbf{1}^\top) \mu_0 \\ &= \sum_{l=r_n}^{L-M+r_n} c_l \cdot \mathbf{1}_B^\top \mathrm{diag}\left(\sqrt{\mu}\right) \left(K - \sqrt{\mu}\sqrt{\mu}^\top \right)^{l-r_n} \mathrm{diag}\left(\sqrt{\mu}\right)^{-1} \mu_0, \end{aligned}$$

where $\mathbf{1}_b$ is the indicator vector corresponding to b. The last equality follows from Proposition F.12. For any test vector $u \in \{0, 1\}^{|\mathcal{X}|^{r_n}}$, using the variational representation of TV distance:

$$\begin{aligned} \left\| \sum_{l=r_n}^{L-M+r_n} c_l \left(p^\pi(B_l = \cdot) - \mu^\pi(\cdot) \right) \right\|_{\mathrm{TV}} &= \max_{u \in \{0,1\}^{|\mathcal{X}|^{r_n}}} u^\top \sum_{l=r_n}^{L-M+r_n} c_l \left(p^\pi(B_l = \cdot) - \mu^\pi(\cdot) \right) \\ &= \max_{u \in \{0,1\}^{|\mathcal{X}|^{r_n}}} \sum_{l=r_n}^{L-M+r_n} c_l \cdot u^\top \mathrm{diag}\left(\sqrt{\mu}\right) \cdot \left(K - \sqrt{\mu}\sqrt{\mu}^\top \right)^{l-r_n} \cdot \mathrm{diag}\left(\sqrt{\mu}\right)^{-1} \cdot \mu_0 \\ &\le \sum_{l=r_n}^{L-M+r_n} c_l \cdot \lambda^{l-r_n} \cdot \left\| \mathrm{diag}\left(\sqrt{\mu}\right)^{-1} \cdot \mu_0 \right\|_2 = \sum_{l=r_n}^{L-M+r_n} c_l \cdot \lambda^{l-r_n} \cdot \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}, \end{aligned} \tag{F.30}$$

https://doi.org/10.52202/079017-2127

where the inequality holds by $\|u^\top \operatorname{diag}\left(\sqrt{\mu}\right)\|_2 \le \|\sqrt{\mu}\|_2 = 1$ and $K - \sqrt{\mu}\sqrt{\mu}^\top$ has leading eigenvalue with magnitude λ. The last identity follows directly from the definition of the χ^2-divergence that $D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1 = \sum_b \mu_0(b)^2/\mu^\pi(b)$.

Substituting the definition of c_l, we have

$$\left\| \frac{\sum_{l=r_n}^{L-M+r_n} \eta^{L-l} p^\pi(B_l = b)}{\sum_{l=r_n}^{L-M+r_n} \eta^{L-l}} - \mu^\pi(A = b) \right\|_{\mathrm{TV}} \le \frac{\sum_{l=r_n}^{L-M+r_n} \eta^{L-l} \cdot \lambda^{l-r_n} \cdot \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{\sum_{l=r_n}^{L-M+r_n} \eta^{L-l}}.$$

We consider two special cases. In the first case, we set $\eta = \lambda$, which gives us

$$\begin{aligned} \left\| \frac{\sum_{l=r_n}^{L-M+r_n} \lambda^{L-l} p^\pi(B_l = b)}{\sum_{l=r_n}^{L-M+r_n} \lambda^{L-l}} - \mu^\pi(A = b) \right\|_{\mathrm{TV}} &\le \frac{\sum_{l=r_n}^{L-M+r_n} \lambda^{L-r_n} \cdot \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{(1 - \lambda^{L-M})/(1-\lambda)} \\ &\le \frac{L \cdot \lambda^{L-r_n} \cdot (1-\lambda)}{1 - \lambda^{L-M}} \cdot \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}. \end{aligned}$$

In the second case, we set $\eta = 1$, which gives us

$$\left\| \frac{\sum_{l=r_n}^{L-M+r_n} p^\pi(B_l = \cdot)}{L - M} - \mu^\pi(A = \cdot) \right\|_{\mathrm{TV}} \le \frac{\sum_{l=r_n}^{L-M+r_n} \lambda^{l-r_n} \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L - M} \le \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{(L-M)(1-\lambda)}.$$

Note that the TV distance is an f-divergence. Thus, we can use the data processing inequality to obtain the desired result for Y_l from the above inequality. To do so, note that $\sum_{l=M+1}^{L} p^\pi(Y_l = \cdot)/(L-M)$ and $\mu^\pi(Y_{L+1} = \cdot)$ can be transformed from $\sum_{l=r_n}^{L-M+r_n-1} p^\pi(B_l = \cdot)/(L-M)$ and $\mu^\pi(A = \cdot)$ by the same emission kernel

$$p^\pi(Y_{L+1} = \cdot \,|\, A = \cdot) = p^\pi(Y_l = \cdot \,|\, B_{l-M+r_n} = \cdot) = \mu^\pi(Y_{L+1} = \cdot \,|\, A = \cdot) = \mu^\pi(Y_l = \cdot \,|\, B_{l-M+r_n} = \cdot).$$

Therefore, by the data processing inequality, it holds that

$$\left\| \frac{\sum_{l=M+1}^{L} p^\pi(Y_l = \cdot)}{L - M} - \mu^\pi(Y_{L+1} = \cdot) \right\|_{\mathrm{TV}} \le \left\| \frac{\sum_{l=r_n}^{L-M+r_n} p^\pi(B_l = \cdot)}{L - M} - \mu^\pi(A = \cdot) \right\|_{\mathrm{TV}} \le \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{(L-M)(1-\lambda)}.$$

Similarly for $p^\pi(Y_{L+1} = \cdot)$ and $\mu^\pi(\cdot)$, we have

$$\begin{aligned} &\|p^\pi(Y_{L+1} = \cdot) - \mu^\pi(Y_{L+1} = \cdot)\|_{\mathrm{TV}} \le \|p^\pi(A = \cdot) - \mu^\pi(A = \cdot)\|_{\mathrm{TV}} \\ &\quad \le \max_{u \in \{0,1\}^{|\mathcal{X}|^{r_n}}} u^\top \cdot \operatorname{diag}\left(\sqrt{\mu}\right) \cdot \left(K - \sqrt{\mu}\sqrt{\mu}^\top\right)^{L-M} \cdot \operatorname{diag}\left(\sqrt{\mu}\right)^{-1} \cdot \mu_0 \le \lambda^{L-M} \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}, \end{aligned}$$

where the latter two inequality follows from the same arguments as in (F.30). Hence, the proof is completed. □

We have established that the average $\sum_{l=M+1}^{L} p^\pi(Y_l = \cdot)/(L-M)$ converges to $\mu^\pi(A = \cdot)$ in total variation distance. This represents a "first-order" convergence since it involves the average of the marginal distribution of Y_l. However, the quantity of interest in (F.25) is the average of the joint distribution of Y_{L+1} and Y_l, which concerns "second-order" convergence. This is studied in the following lemma.

Lemma F.17. *Following the notations introduced above, for the Markov chain with parent set* $\mathtt{pa} = \{-r_1, \ldots, -r_n\}$*, let* $D_{\chi^2}(\mu_0 \,\|\, \mu^\pi)$ *be the* χ^2*-divergence between the initial distribution* μ_0 *and the stationary distribution* μ^π *over the first* r_n *tokens. Take any* $\mathcal{S}, \mathcal{S}' \subseteq [M]$ *and let* $Y_l = (x_l, X_{l-\mathcal{S}})$ *and* $Y_l' = (x_l, X_{l-\mathcal{S}'})$ *for* $l = M+1, \ldots, L+1$*. Suppose* $L/2 \ge M \ge r_n$*. For* $\widehat{p}^\pi$ *defined in* (F.25)*, we have*

$$\|\widehat{p}^\pi(Y_{L+1} = \cdot, Y' = \cdot) - \mu^\pi(Y_{L+1} = \cdot) \times \mu^\pi(Y' = \cdot)\|_{\mathrm{TV}} \le \frac{2M}{L} + \frac{4\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda) \cdot \sqrt{\min_E \mu^\pi(Y_{L+1} = E)}}.$$

In particular, we have

$$\begin{aligned} &\left\| \widehat{p}^\pi(Y_{L+1} = \cdot, Y' = \cdot) - \mu^\pi(Y_{L+1} = \cdot) \times \left(\frac{1}{L-M} \sum_{l=M+1}^{L} p^\pi(Y_l' = \cdot) \right) \right\|_{\mathrm{TV}} \\ &\quad \le \frac{2M}{L} + \frac{2\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda) \cdot \sqrt{\min_E \mu^\pi(Y_{L+1} = E)}}. \end{aligned} \tag{F.31}$$

Proof of Lemma F.17. Let us take $\mu^\pi(E)\cdot(L-M)^{-1}\sum_{l=M+1}^{L} p^\pi(Y_l = E')$ as the intermediate distribution, and we have by (F.25) that

$$
\begin{aligned}
&\widehat{p}^\pi(Y_{L+1}=E, Y'=E') - \mu^\pi(Y_{L+1}=E)\cdot\Bigg(\frac{1}{L-M}\sum_{l=M+1}^{L} p^\pi(Y_l=E')\Bigg)\\
&= \underbrace{\frac{1}{L-M}\sum_{l=M+1}^{L-M+r_n}\sum_{A,B_l}\mu^\pi(Y_{L+1}=E\,|\,A)\cdot\big(P^{L-l-(M-r_n)}(A\,|\,B_l)-\mu^\pi(A)\big)\cdot p^\pi(Y_l'=E'\,|\,B_l)\cdot p^\pi(B_l)}_{\text{(I)}}\\
&\quad+\underbrace{\frac{1}{L-M}\sum_{l=L-M+r_n+1}^{L}\big(p^\pi(Y_{L+1}=E,Y_l'=E')-\mu^\pi(Y_{L+1}=E)p^\pi(Y_l'=E')\big)}_{\text{(II)}}.
\end{aligned}
\tag{F.32}
$$

where we use the fact that $\sum_A \mu^\pi(Y_{L+1}=E\,|\,A)\mu^\pi(A)=\mu^\pi(Y=E)$ for the first line. The second term on the right hand side can be easily controlled as we already have an L^{-1} factor. We let TV_0 be the total variation distance of the second term. It is easy to see that

$$
\mathrm{TV}_0 := \frac{1}{2}\sum_{E,E'}|\text{(II)}| \le \frac{M-r_n}{L-M} \le \frac{M}{L-M},
$$

where we remark that (II) is a function of both E and E', and the total variation distance is just taking the sum of the absolute values of the differences. Here, we also use the fact that $L\ge 2M$. Using Proposition F.12, we can also rewrite the first term on the right hand side of (F.32) in the matrix form as

$$
\begin{aligned}
\text{(I)} = \frac{1}{L-M}\sum_{l=M+1}^{L-M+r_n}&\mu^\pi(Y_{L+1}=\cdot\,|\,A=\cdot)\cdot\mathrm{diag}\left(\sqrt{\mu}\right)\cdot\left(K-\sqrt{\mu}\sqrt{\mu}^\top\right)^{L-l-(M-r_n)}\cdot\mathrm{diag}\left(\sqrt{\mu}\right)^{-1}\\
&\cdot\mathrm{diag}(p^\pi(B_l=\cdot))\cdot p^\pi(Y_l'=\cdot\,|\,B_l=\cdot)^\top.
\end{aligned}
$$

When considering the ℓ_1-norm of the above term, we introduce a test matrix U of shape $|\mathcal{X}|^{|Y_{L+1}|}\times|\mathcal{X}|^{|Y_{L+1}|}$ with each element of U chosen from $\{0,1\}$. Let TV_1 be the total variation distance of the first term (I). Then, we have

$$
\begin{aligned}
\mathrm{TV}_1 \le \max_U \mathrm{Tr}\Bigg[&\frac{1}{L-M}\sum_{l=M+1}^{L-M+r_n}\mu^\pi(Y_{L+1}=\cdot\,|\,A=\cdot)\cdot\mathrm{diag}\left(\sqrt{\mu}\right)\cdot\left(K-\sqrt{\mu}\sqrt{\mu}^\top\right)^{L-l-(M-r_n)}\\
&\cdot\mathrm{diag}\left(\sqrt{\mu}\right)^{-1}\cdot\mathrm{diag}(p^\pi(B_l=\cdot))\cdot p^\pi(Y_l'=\cdot\,|\,B_l=\cdot)^\top\cdot U(\cdot,\cdot)^\top\Bigg].
\end{aligned}
$$

To upper bound this quantity, we consider each row of U as $U(E,\cdot)=u(\cdot\,|\,E)^\top$. Note that $u(\cdot\,|\,E)$ is also a $\{0,1\}$-valued vector. By expanding the trace, we have

$$
\begin{aligned}
\mathrm{TV}_1 \le \sum_E \max_{u(\cdot\,|\,E)}\frac{1}{L-M}\sum_{l=M+1}^{L-M+r_n}&\mu^\pi(Y_{L+1}=E\,|\,A=\cdot)\cdot\mathrm{diag}\left(\sqrt{\mu}\right)\\
&\cdot\left(K-\sqrt{\mu}\sqrt{\mu}^\top\right)^{L-l-(M-r_n)}\cdot\mathrm{diag}\left(\sqrt{\mu}\right)^{-1}\cdot\mathrm{diag}(p^\pi(B_l=\cdot))\cdot p^\pi(Y_l'=\cdot\,|\,B_l=\cdot)^\top\cdot u(\cdot\,|\,E).
\end{aligned}
$$

Note that the ℓ_2-norm of the vector in the last line can be upper bounded by

$$
\begin{aligned}
&\left\|\left(K-\sqrt{\mu}\sqrt{\mu}^\top\right)^{L-l-(M-r_n)}\cdot\mathrm{diag}\left(\sqrt{\mu}\right)^{-1}\cdot\mathrm{diag}(p^\pi(B_l=\cdot))\cdot p^\pi(Y_l'=\cdot\,|\,B_l=\cdot)^\top\cdot u(\cdot\,|\,E)\right\|_2\\
&\quad\le\left\|\lambda^{L-l-(M-r_n)}\cdot\mathrm{diag}\left(\sqrt{\mu}\right)^{-1}\cdot\mathrm{diag}(p^\pi(B_l=\cdot))\cdot\mathbf{1}\right\|_2=\lambda^{L-l-(M-r_n)}\left\|\mathrm{diag}\left(\sqrt{\mu}\right)^{-1}\cdot p^\pi(B_l=\cdot)\right\|_2\\
&\quad=\lambda^{L-l-(M-r_n)}\sqrt{D_{\chi^2}(p^\pi(B_l=\cdot)\,\|\,\mu^\pi(B_l=\cdot))+1}\le\lambda^{L-l-(M-r_n)}\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1},
\end{aligned}
\tag{F.33}
$$

where the first inequality holds by noting that $p^\pi(Y_l' = \cdot \,|\, B_l = \cdot)^\top \cdot u(\cdot \,|\, E)$ is a vector with element within $[0, 1]$, and also invoking the operator norm of the matrix $\widetilde{K} - \sqrt{\mu}\sqrt{\mu}^\top$. The second identity follows from the definition of the χ^2-divergence that $D_{\chi^2}(p^\pi(B_l = \cdot) \,\|\, \mu^\pi(\cdot)) + 1 = \sum_b p^\pi(B_l = b)^2/\mu^\pi(b)$. The last inequality is the data processing inequality as $p^\pi(B_l = \cdot)$ can be transformed from $\mu_0(B_{r_n})$ and $\mu^\pi(B_l)$ can be transformed from $\mu^\pi(B_{r_n})$ by the same emission kernel $\mu^\pi(B_l = \cdot \,|\, B_{r_n} = \cdot)$. Consequently, we have for the TV distance that

$$
\begin{aligned}
\mathrm{TV}_1 \le \frac{1}{L-M} \sum_{l=M+1}^{L-M+r_n} \lambda^{L-l-(M-r_n)} \cdot \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1} \\
\cdot \max_{\{v(\cdot \,|\, E)\}_E \colon \|v(\cdot \,|\, E)\|_2 \le 1} \sum_{E,A} \mu^\pi(Y_{L+1} = E \,|\, A) \cdot \sqrt{\mu^\pi(A)} \cdot v(A \,|\, E) \\
\le \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{(L-M)(1-\lambda)} \cdot \max_{\{v(\cdot \,|\, E)\}_E \colon \|v(\cdot \,|\, E)\|_2 \le 1} \sum_{A,E} \frac{\mu^\pi(A \,|\, Y_{L+1} = E)}{\sqrt{\mu^\pi(A)}} \cdot v(A \,|\, E) \cdot \mu^\pi(Y_{L+1} = E) \\
\le \max_{\{v(\cdot \,|\, E)\}_E \colon \|v(\cdot \,|\, E)\|_2 \le 1} \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{(L-M)(1-\lambda)} \cdot \sqrt{I_{\chi^2}(A; Y_{L+1}) + 1} \cdot \sqrt{\sum_{A,E} v(A \,|\, E)^2 \cdot \mu^\pi(Y_{L+1} = E)}.
\end{aligned}
$$

where in the first equality, we use the variational form of the ℓ_2-norm for vector $\mu^\pi(Y_{L+1} = E \,|\, A = \cdot) \cdot \mathrm{diag}(\sqrt{\mu})$. In the second inequality, we apply (F.33) and use the Bayes rule. The last inequality follows from the Cauchy-Schwarz inequality. Here, the mutual information $I_{\chi^2}(A; Y_{L+1}) + 1$ can be upper bounded by

$$
I_{\chi^2}(A; Y_{L+1}) + 1 = \sum_{A,E} \frac{\mu^\pi(Y_{L+1} = E \,|\, A)}{\mu^\pi(Y_{L+1} = E)} \cdot \mu^\pi(Y_{L+1} = E, A) \le \frac{1}{\min_E \mu^\pi(Y_{L+1} = E)},
$$

and the last term involving $v(A \,|\, E)$ can be upper bounded by 1 thanks to the constraint on $v(\cdot \,|\, E)$. In conclusion,

$$
\mathrm{TV}_1 \le \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{(L-M)(1-\lambda) \cdot \sqrt{\min_E \mu^\pi(Y_{L+1} = E)}}.
$$

Lastly, let us relate the intermediate distribution to the final distribution $\mu^\pi(Y = \cdot) \times \mu^\pi(Y' = \cdot)$, where we define the total variation distance TV_2 as

$$
\mathrm{TV}_2 := \left\| \mu^\pi(\cdot) \cdot \left(\frac{1}{L-M} \sum_{l=M+1}^{L} p^\pi(Y_l' = \cdot) \right) - \mu^\pi(\cdot) \cdot \mu^\pi(\cdot) \right\|_{\mathrm{TV}} = \left\| \left(\frac{1}{L-M} \sum_{l=M+1}^{L} p^\pi(Y_l' = \cdot) \right) - \mu^\pi(\cdot) \right\|_{\mathrm{TV}}.
$$

Invoking (F.28) of Lemma F.16, we have this quantity upper bounded by

$$
\mathrm{TV}_2 \le \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{(L-M)(1-\lambda)}.
$$

Using the triangular inequality for the total variation distance, we have

$$
\begin{aligned}
&\|\widehat{p}^\pi(Y_{L+1} = \cdot, Y' = \cdot) - \mu^\pi(Y_{L+1} = \cdot) \times \mu^\pi(Y' = \cdot)\|_{\mathrm{TV}} \\
&\quad \le \mathrm{TV}_0 + \mathrm{TV}_1 + \mathrm{TV}_2 \\
&\quad \le \frac{M}{L-M} + \frac{2\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{(L-M)(1-\lambda) \cdot \sqrt{\min_E \mu^\pi(Y_{L+1} = E)}} \\
&\quad \le \frac{2M}{L} + \frac{4\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda) \cdot \sqrt{\min_E \mu^\pi(Y_{L+1} = E)}},
\end{aligned}
$$

and the upper bound for (F.31) follows by the same arguments. Hence, the proof is completed. □

In the following, we use a similar technique as in Lemma F.17 to derive a bound for the chi-square divergence.

Lemma F.18. *For the χ^2-divergence between the empirical distribution $(L-M)^{-1}\sum_{l=M+1}^{L} \mathbb{1}(Y_l = \cdot)$ and the stationary distribution $\mu^\pi(\cdot)$, we have*

$$\mathbb{E}\biggl[D_{\chi^2}\biggl(\frac{1}{L-M}\sum_{l=M+1}^{L}\mathbb{1}(Y_l=\cdot)\,\bigg\|\,\mu^\pi(Y_{L+1}=\cdot)\biggr)\biggr] \le \frac{4(1-\lambda)^{-1}\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}+16M}{L\cdot\min_E \mu^\pi(Y_{L+1}=E)},$$

where the expectation is with respect to $X \sim p^\pi$.

Proof of Lemma F.18. By definition of the χ^2-divergence, what we aim to bound is just

$$\begin{aligned}
&\mathbb{E}\biggl[\sum_E \biggl((L-M)^{-1}\sum_{l=M+1}^{L}\mathbb{1}(Y_l=E)-\mu^\pi(E)\biggr)^2\Big/\mu^\pi(E)\biggr]\\
&\quad=\mathbb{E}\biggl[\sum_E\frac{(L-M)^{-2}\sum_{l,l'=M+1}^{L}\mathbb{1}(Y_l=Y_{l'}=E)-\mu^\pi(E)^2}{\mu^\pi(E)}\biggr]\\
&\quad=\mathbb{E}\biggl[\sum_E\sum_{l,l'=M+1}^{L}\frac{\mathbb{1}(Y_l=Y_{l'}=E)}{(L-M)^2\mu^\pi(E)}-1\biggr]=\sum_E\sum_{l,l'=M+1}^{L}\frac{p^\pi(Y_l=Y_{l'}=E)}{(L-M)^2\mu^\pi(E)}-1.
\end{aligned}$$

To study the above quantity, for $l \ge 2M - r_n + 2$, we define

$$J_1(l) := \sum_E \sum_{l'=M+1}^{l-M+r_n-1}\frac{p^\pi(Y_l=Y_{l'}=E)}{(L-M)^2\mu^\pi(E)}-\frac{l-2M+r_n}{(L-M)^2}.$$

Following our convention, we let $A_l = X_{l-M:l-M+r_n-1}$ and $B_{l'} = X_{l'-r_n+1:l'}$ be two length-r_n window and by the Markov property, we have

$$Y_{l+1} \perp\!\!\!\perp (B_{l'}, Y_{l'})\,|\,A_l, \quad (Y_{l+1}, B_l) \perp\!\!\!\perp Y_{l'}\,|\,B_{l'}.$$

Let us fix an index $l \ge 2M - r_n + 2$ and take a summation over $M + 1 \le l' \le l - M + r_n - 1$. Expanding the joint distribution, we have

$$\begin{aligned}
J_1(l) &:= \frac{1}{(L-M)^2}\sum_{l'=M+1}^{l-M+r_n-1}\sum_{E,A_l,B_{l'}}\mu^\pi(Y_l=E\,|\,A_l)\cdot\bigl(P^{l-l'-M+r_n-1}(A_l\,|\,B_{l'})-\mu^\pi(A_l)\bigr)\\
&\qquad\cdot p^\pi(Y_{l'}=E\,|\,B_{l'})\cdot p^\pi(B_{l'})\cdot\mu^\pi(Y_{l'}=E)^{-1}\\
&=\frac{1}{(L-M)^2}\sum_{l'=M+1}^{l-M+r_n-1}\mathrm{Tr}\Bigl[\mu^\pi(Y_l=\cdot\,|\,A_l=\cdot)\cdot\mathrm{diag}\,(\sqrt{\mu})\cdot\bigl(K-\sqrt{\mu}\sqrt{\mu}^\top\bigr)^{l-l'-M+r_n-1}\\
&\qquad\cdot\mathrm{diag}\bigl(\sqrt{\mu}\bigr)^{-1}\cdot\mathrm{diag}(p^\pi(B_{l'}=\cdot))\cdot p^\pi(Y_{l'}=\cdot\,|\,B_{l'}=\cdot)^\top\cdot\mathrm{diag}(\mu^\pi(Y_{l'}=\cdot)^{-1})\Bigr]\\
&=\frac{1}{(L-M)^2}\sum_{l'=M+1}^{l-M+r_n-1}\mathrm{Tr}\Bigl[\mathrm{diag}(\mu^\pi(Y_{l'}=\cdot)^{-1/2})\cdot\mu^\pi(Y_l=\cdot\,|\,A_l=\cdot)\cdot\mathrm{diag}\,(\sqrt{\mu})\\
&\qquad\cdot\bigl(K-\sqrt{\mu}\sqrt{\mu}^\top\bigr)^{l-l'-M+r_n-1}\cdot\mathrm{diag}\bigl(\sqrt{\mu}\bigr)^{-1}\cdot\mathrm{diag}(p^\pi(B_{l'}=\cdot))\\
&\qquad\cdot p^\pi(Y_{l'}=\cdot\,|\,B_{l'}=\cdot)^\top\cdot\mathrm{diag}(\mu^\pi(Y_{l'}=\cdot)^{-1/2})\Bigr],
\end{aligned}$$

where the first identity follows from the fact that

$$\begin{aligned}
&\sum_{E,A_l,B_l'}\mu^\pi(Y_l=E\,|\,A_l)\cdot\mu^\pi(A_l)\cdot p^\pi(Y_{l'}=E\,|\,B_{l'})\cdot p^\pi(B_{l'})\cdot\mu^\pi(Y_{l'}=E)^{-1}\\
&\quad=\sum_E p^\pi(Y_{l'}=E)\cdot\mu^\pi(Y_{l'}=E)\cdot\mu^\pi(Y_{l'}=E)^{-1}=1,
\end{aligned}$$

and the second identity follows from Proposition F.12. We next invoke the Cauchy-Schwarz inequality for trace, i.e., $\mathrm{Tr}(W^\top V)^2 \le \mathrm{Tr}(W^\top W)\,\mathrm{Tr}(V^\top V)$, where we take

$$\begin{aligned}
W^\top &= \mathrm{diag}(\mu^\pi(Y_l=\cdot)^{-1/2})\cdot\mu^\pi(Y_l=\cdot\,|\,A_l=\cdot)\cdot\mathrm{diag}(\sqrt{\mu})\cdot\bigl(K-\sqrt{\mu}\sqrt{\mu}^\top\bigr)^{l-l'-M+r_n-1},\\
V &= \mathrm{diag}\bigl(\sqrt{\mu}\bigr)^{-1}\cdot\mathrm{diag}(p^\pi(B_{l'}=\cdot))\cdot p^\pi(Y_{l'}=\cdot\,|\,B_{l'}=\cdot)^\top\cdot\mathrm{diag}(\mu^\pi(Y_{l'}=\cdot)^{-1/2})\\
&= \mathrm{diag}\bigl(\sqrt{\mu}\bigr)\cdot p^\pi(Y_{l'}=\cdot\,|\,B_{l'}=\cdot)^\top\cdot\mathrm{diag}(\mu^\pi(Y_{l'}=\cdot)^{-1/2})
\end{aligned}$$

Note that

$$\sqrt{\mathrm{Tr}(W^\top W)} \le \lambda^{l-l'-M+r_n-1} \cdot \sqrt{\mathrm{Tr}\left(\mathrm{diag}(\mu^\pi(Y_l=\cdot)^{-1})\mu^\pi(Y_l=\cdot \,|\, A=\cdot)\mathrm{diag}\left(\mu\right)\mu^\pi(Y_l=\cdot \,|\, A=\cdot)^\top\right)}$$
$$= \lambda^{l-l'-M+r_n-1} \cdot \sqrt{\sum_{A_l,Y_l} \frac{\mu^\pi(Y_l,A_l)^2}{\mu^\pi(Y_l)\cdot\mu^\pi(A_l)}}.$$

Following the same calculation, we have

$$\sqrt{\mathrm{Tr}(V^\top V)} = \sqrt{\sum_{Y_{l'},B_{l'}} \frac{p^\pi(Y_{l'},B_{l'})^2}{\mu^\pi(Y_{l'})\mu^\pi(B_{l'})}}.$$

Therefore,

$$J_1(l) \le \frac{1}{(L-M)^2} \sum_{l'=M+1}^{l-M+r_n-1} \lambda^{l-l'-M+r_n-1} \cdot \sqrt{\sum_{A_l,Y_l} \frac{\mu^\pi(Y_l,A_l)^2}{\mu^\pi(Y_l)\cdot\mu^\pi(A_l)} \cdot \sum_{Y_{l'},B_{l'}} \frac{p^\pi(Y_{l'},B_{l'})^2}{\mu^\pi(Y_{l'})\mu^\pi(B_{l'})}}.$$

We further have

$$\sum_{Y_{l'},B_{l'}} \frac{p^\pi(Y_{l'},B_{l'})^2}{\mu^\pi(Y_{l'})\mu^\pi(B_{l'})} \le \max_{Y_{l'},B_{l'}} \left\{\frac{p^\pi(Y_{l'} \,|\, B_{l'})}{\mu^\pi(Y_{l'})}\right\} \cdot \sum_{B_{l'}} \frac{p^\pi(B_{l'})^2}{\mu^\pi(B_{l'})}$$
$$\le \frac{D_{\chi^2}(p^\pi(B_{l'}=\cdot)\,\|\,\mu^\pi(B_{l'}=\cdot))+1}{\min_E \mu^\pi(Y_{L+1}=E)} \le \frac{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}{\min_E \mu^\pi(Y_{L+1}=E)},$$

where the last inequality holds by the data processing inequality. Similarly, we have

$$\sum_{A_l,Y_l} \frac{\mu^\pi(Y_l,A_l)^2}{\mu^\pi(Y_l)\cdot\mu^\pi(A_l)} \le \max_{Y_l,A_l} \left\{\frac{\mu^\pi(Y_l \,|\, A_l)}{\mu^\pi(Y_l)}\right\} \le \frac{1}{\min_E \mu^\pi(Y_{L+1}=E)}.$$

Therefore, we conclude that

$$J_1(l) \le \frac{\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}}{(L-M)^2(1-\lambda)\cdot\min_E \mu^\pi(Y_{L+1}=E)},$$

and

$$2\sum_{l=2M-r_n+2}^{L} J_1(l) \le \frac{2\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}}{(L-M)(1-\lambda)\cdot\min_E \mu^\pi(Y_{L+1}=E)},$$

where we double the value as $l > l'$ only contributes to half of the terms in the double summation. Note that in the above summation for $l > l'$, we only include terms satisfying $l - l' \ge M - r_n + 1$ and $l - (M+1) \ge M - r_n + 1$. For the remaining (l, l') not included above, each term is bounded above by

$$\left|\frac{1}{(L-M)^2}\left(\sum_E \frac{p^\pi(Y_l = Y_{l'} = E)}{\mu^\pi(E)} - 1\right)\right| \le \frac{1}{(L-M)^2 \min_E \mu^\pi(Y_{L+1}=E)},$$

and we have no more than $4L(M - r_n + 1)$ of these terms in total. As a result, we conclude with $L/2 \ge M \ge r_n$ that

$$J_1 \le \frac{4(1-\lambda)^{-1}\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}+16M}{L\cdot\min_E \mu^\pi(Y_{L+1}=E)}.$$

Hence, we complete the proof of Lemma F.18. □

Proposition F.19. *Let us define*

$$\widetilde{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star}) = \frac{\mu^\pi(z, Z_{-\mathcal{S}^\star})\exp\left(a\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(z_{-h}=x_{L+1-h})\right)}{\sum_{z',Z'_{-\mathcal{S}^\star}}\mu^\pi(z', Z'_{-\mathcal{S}^\star})\exp\left(a\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(z'_{-h}=x_{L+1-h})\right)},$$

https://doi.org/10.52202/079017-2127

where $Z_{-\mathcal{S}^\star} = (z_{-h})_{h\in\mathcal{S}^\star}$ and μ^π is the stationary distribution of the Markov chain over a window of size $M+1$. We also treat $\widetilde{\mu}_X^\pi(\cdot)$ as a length $|\mathcal{X}|$ vector where $\mathcal{X}$ is the state space of the Markov chain. Let $\widehat{\nu}_X^\pi(z, Z_{-\mathcal{S}^\star}) = \sum_{l=M+1}^{L} \sigma_l^\star \, \mathbb{1}(x_l = z, X_{l-\mathcal{S}^\star} = Z_{-\mathcal{S}^\star})$ where

$$\sigma_l^\star = \frac{\exp(a \cdot \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h}))}{\sum_{l'=M+1}^{L} \exp(a \cdot \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l'-h} = x_{L+1-h}))}.$$

Then, we have

$$\mathbb{E}_X\left[\|\widetilde{\mu}_X^\pi(z=\cdot, Z_{-\mathcal{S}^\star}=\cdot) - \widehat{\nu}_X^\pi(z=\cdot, Z_{-\mathcal{S}^\star}=\cdot)\|_1\right] \le \frac{4\big((1-\lambda)^{-1}\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi)+1} + 4M\big)^{1/2}}{L^{1/2} \cdot \min_{x_{L+1}, X_{L+1-\mathcal{S}^\star}} \mu^\pi(x_{L+1}, X_{L+1-\mathcal{S}^\star})}.$$

Proof of Proposition F.19. To unify the notations, we let $Z = (z_{-M}, \ldots, z_{-1})$ and define

$$\widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star}) = \frac{1}{L-M} \sum_{l=M+1}^{L} \mathbb{1}(x_l = z, X_{l-\mathcal{S}^\star} = Z_{-\mathcal{S}^\star}),$$
$$R(Z_{-\mathcal{S}^\star}, X_{L+1-\mathcal{S}^\star}) = \exp\left(a \cdot \mathbb{1}(Z_{-\mathcal{S}^\star} = x_{L+1-\mathcal{S}^\star})\right).$$

Using these notations, we can define the normalizing factor in $\widetilde{\mu}_X^\pi$ and $y_X^\star$ respectively as

$$\Phi = \sum_{z, Z_{-\mathcal{S}^\star}} \mu^\pi(z, Z_{-\mathcal{S}^\star}) \cdot R(Z_{-\mathcal{S}^\star}, X_{L+1-\mathcal{S}^\star}), \quad \widehat{\Phi} = \sum_{z, Z_{-\mathcal{S}^\star}} \widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star}) \cdot R(Z_{-\mathcal{S}^\star}, X_{L+1-\mathcal{S}^\star}).$$

We also define

$$\phi(z, Z_{-\mathcal{S}^\star}) = \mu^\pi(z, Z_{-\mathcal{S}^\star}) \cdot R(Z_{-\mathcal{S}^\star}, X_{L+1-\mathcal{S}^\star}), \quad \widehat{\phi}(z, Z_{-\mathcal{S}^\star}) = \widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star}) \cdot R(Z_{-\mathcal{S}^\star}, X_{L+1-\mathcal{S}^\star}).$$

We can then rewrite the objective as

$$\begin{aligned}
&\|\widetilde{\mu}_X^\pi(z=\cdot, Z_{-\mathcal{S}^\star}=\cdot) - \widehat{\nu}_X^\pi(z=\cdot, Z_{-\mathcal{S}^\star}=\cdot)\|_1 \\
&\quad = \sum_{z, Z_{-\mathcal{S}^\star}} \left| \frac{\phi(z, Z_{-\mathcal{S}^\star})}{\Phi} - \frac{\widehat{\phi}(z, Z_{-\mathcal{S}^\star})}{\widehat{\Phi}} \right| \le \sum_{z, Z_{-\mathcal{S}^\star}} \frac{\widehat{\phi}(z, Z_{-\mathcal{S}^\star}) \cdot |\widehat{\Phi} - \Phi| + |\phi(z, Z_{-\mathcal{S}^\star}) - \widehat{\phi}(z, Z_{-\mathcal{S}^\star})| \cdot \widehat{\Phi}}{\Phi \cdot \widehat{\Phi}} \\
&\quad = \frac{|\widehat{\Phi} - \Phi| + \sum_{z, Z_{-\mathcal{S}^\star}} |\phi(z, Z_{-\mathcal{S}^\star}) - \widehat{\phi}(z, Z_{-\mathcal{S}^\star})|}{\Phi} \le \frac{2\sum_{z, Z_{-\mathcal{S}^\star}} |\phi(z, Z_{-\mathcal{S}^\star}) - \widehat{\phi}(z, Z_{-\mathcal{S}^\star})|}{\Phi}.
\end{aligned}$$

Furthermore, notice that

$$\begin{aligned}
&\frac{\sum_{z, Z_{-\mathcal{S}^\star}} |\phi(z, Z_{-\mathcal{S}^\star}) - \widehat{\phi}(z, Z_{-\mathcal{S}^\star})|}{\Phi} = \frac{\sum_{z, Z_{-\mathcal{S}^\star}} |(\mu^\pi(z, Z_{-\mathcal{S}^\star}) - \widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star})) \cdot R(Z_{-\mathcal{S}^\star}, X_{L+1-\mathcal{S}^\star})|}{\sum_{z, Z_{-\mathcal{S}^\star}} \mu^\pi(z, Z_{-\mathcal{S}^\star}) \cdot R(Z_{-\mathcal{S}^\star}, X_{L+1-\mathcal{S}^\star})} \\
&\quad \le \frac{\sum_{z, Z_{-\mathcal{S}^\star}} |(\mu^\pi(z, Z_{-\mathcal{S}^\star}) - \widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star}))| + (e^a - 1)\sum_{z, Z_{-\mathcal{S}^\star} \in \Gamma_X} |\mu^\pi(z, Z_{-\mathcal{S}^\star}) - \widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star})|}{1 + (e^a - 1) \cdot \sum_{z, Z_{-\mathcal{S}^\star} \in \Gamma_X} \mu^\pi(z, Z_{-\mathcal{S}^\star})} \\
&\quad \le \sum_{z, Z_{-\mathcal{S}^\star}} |\mu^\pi(z, Z_{-\mathcal{S}^\star}) - \widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star})| + \frac{\sum_{z, Z_{-\mathcal{S}^\star} \in \Gamma_X} |(\mu^\pi(z, Z_{-\mathcal{S}^\star}) - \widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star}))|}{\sum_{z, Z_{-\mathcal{S}^\star} \in \Gamma_X} \mu^\pi(z, Z_{-\mathcal{S}^\star})}.
\end{aligned} \tag{F.34}$$

where we define $\Gamma_X = \{Z_{-\mathcal{S}^\star} : Z_{-\mathcal{S}^\star} = X_{L+1-\mathcal{S}^\star}\}$. Note that when $Z_{-\mathcal{S}^\star} \in \Gamma_X$, we have $R(Z_{-\mathcal{S}^\star}, X_{L+1-\mathcal{S}^\star}) = e^a$ and when $Z_{-\mathcal{S}^\star} \notin \Gamma_X$, we have $R(Z_{-\mathcal{S}^\star}, X_{L+1-\mathcal{S}^\star}) = 1$. For the first term on the right-hand side of (F.34), we have by Cauchy-Schwarz that

$$\begin{aligned}
\mathbb{E}_X\left[\sum_{z, Z_{-\mathcal{S}^\star}} |\mu^\pi(z, Z_{-\mathcal{S}^\star}) - \widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star})|\right] &\le \left(\mathbb{E}_X\left[\sum_{z, Z_{-\mathcal{S}^\star}} \frac{(\mu^\pi(z, Z_{-\mathcal{S}^\star}) - \widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star}))^2}{\mu^\pi(z, Z_{-\mathcal{S}^\star})}\right]\right)^{1/2} \\
&\le \left(\frac{4(1-\lambda)^{-1}\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi)+1} + 16M}{L \cdot \min_{x_{L+1}, X_{L+1-\mathcal{S}^\star}} \mu^\pi(x_{L+1}, X_{L+1-\mathcal{S}^\star})}\right)^{1/2},
\end{aligned}$$

https://doi.org/10.52202/079017-2127

where in the last inequality, we invoke Lemma F.18 where we take $Y_l = x_l$ in the lemma. For the second term on the right hand of (F.34), we note that

$$\begin{aligned}
&\mathbb{E}_X\left[\frac{\sum_{z,Z_{-\mathcal{S}^\star}\in\Gamma_X}|(\mu^\pi(z,Z_{-\mathcal{S}^\star})-\widehat{\mu}^\pi_X(z,Z_{-\mathcal{S}^\star}))|}{\sum_{z,Z_{-\mathcal{S}^\star}\in\Gamma_X}\mu^\pi(z,Z_{-\mathcal{S}^\star})}\right]\\
&\quad\le\sum_{E,z}\mathbb{E}_X\left[\frac{|\mu^\pi(z,Z_{-\mathcal{S}^\star}=E)-\widehat{\mu}^\pi_X(z,Z_{-\mathcal{S}^\star}=E)|}{\mu^\pi(Z_{-\mathcal{S}^\star}=E)}\cdot\mathbb{1}(X_{L+1-\mathcal{S}^\star}=E)\right]\\
&\quad\le\sum_{E,z}\left(\mathbb{E}_X\left[\Big(\frac{\mu^\pi(z,Z_{-\mathcal{S}^\star}=E)-\widehat{\mu}^\pi_X(z,Z_{-\mathcal{S}^\star}=E)}{\sqrt{\mu^\pi(Z_{-\mathcal{S}^\star}=E)}}\Big)^2\right]\cdot\frac{p^\pi(X_{L+1-\mathcal{S}^\star}=E)}{\mu^\pi(Z_{-\mathcal{S}^\star}=E)}\right)^{1/2}\\
&\quad\le\left(\mathbb{E}_X\left[\sum_{E,z}\frac{(\mu^\pi(z,Z_{-\mathcal{S}^\star}=E)-\widehat{\mu}^\pi_X(z,Z_{-\mathcal{S}^\star}=E))^2}{\mu^\pi(Z_{-\mathcal{S}^\star}=E)}\right]\cdot\sum_{E,z}\frac{p^\pi(X_{L+1-\mathcal{S}^\star}=E)}{\mu^\pi(Z_{-\mathcal{S}^\star}=E)}\right)^{1/2},
\end{aligned}\tag{F.35}$$

where the last two inequalities follow from the Cauchy-Schwarz inequality. We have an upper bound for the second term on the right-hand side of (F.35) that

$$\left(\sum_{E,z}\frac{p^\pi(X_{L+1-\mathcal{S}^\star}=E)}{\mu^\pi(Z_{-\mathcal{S}^\star}=E)}\right)^{1/2}\le\sqrt{\frac{1}{\min_E\mu^\pi(Z_{-\mathcal{S}^\star}=E)}}.$$

We can also apply Lemma F.18 to the first term with $Y_{L+1}=(x_{L+1},X_{L+1-\mathcal{S}^\star})$ and conclude that

$$\begin{aligned}
&\left(\mathbb{E}_X\left[\sum_{E,z}\frac{(\mu^\pi(z,Z_{-\mathcal{S}^\star}=E)-\widehat{\mu}^\pi_X(z,Z_{-\mathcal{S}^\star}=E))^2}{\mu^\pi(Z_{-\mathcal{S}^\star}=E)}\right]\right)^{1/2}\\
&\qquad\le\left(\frac{4(1-\lambda)^{-1}\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}+16M}{L\cdot\min_{x_{L+1},X_{L+1-\mathcal{S}^\star}}\mu^\pi(x_{L+1},X_{L+1-\mathcal{S}^\star})}\right)^{1/2}.
\end{aligned}$$

In summary, we have

$$\begin{aligned}
&\mathbb{E}_X\left[\|\widetilde{\mu}^\pi_X(e_k)-y^\star(k)\|_1\right]\\
&\quad\le\frac{2}{\min_{x_{L+1},X_{L+1-\mathcal{S}^\star}}\mu^\pi(x_{L+1},X_{L+1-\mathcal{S}^\star})}\cdot\left(\frac{(1-\lambda)^{-1}\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}+4M}{L}\right)^{1/2}\\
&\qquad+2\left(\frac{(1-\lambda)^{-1}\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}+4M}{L\cdot\min_{x_{L+1},X_{L+1-\mathcal{S}^\star}}\mu^\pi(x_{L+1},X_{L+1-\mathcal{S}^\star}))}\right)^{1/2}.
\end{aligned}$$

Note that the second term is dominated by the first term. Thus, we conclude the proof of Proposition F.19. □

https://doi.org/10.52202/079017-2127

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The convergence results are established in Theorem 3.6 and we relate the learned transformer model to the generalized induction head in the following discussions.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discussed in §B.2

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

https://doi.org/10.52202/079017-2127

Justification: We provide the full set of assumptions in Assumption 3.3 and Assumption 3.5. We provide a complete and correct proof in §E and §F.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provide the complete details for our numerical experiment in §B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

https://doi.org/10.52202/079017-2127

Answer: [No]

Justification: We do not release the data and code, but the details provided in §B are sufficient for reproducing the synthetic data and the experiment results.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://nips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental Setting/Details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: We provide all the training and test details in §B.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment Statistical Significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [No]

Justification: We do not report error bars because the training behavior is consistent across different runs, and the goal of the experiments is to corroborate our main theoretical results, rather than achieving better performance on benchmarks.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).

https://doi.org/10.52202/079017-2127

- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We provide information about compute resources in §B.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: We have reviewed the NeurIPS Code of Ethics and we confirm that our submission adheres to the guidelines therein.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: In the current paper, we focus on developing theoretical understanding of transformers, and the goal is to analyze existing architectures instead of proposing new models for better performance. Therefore, we do not see immediate societal impact of our paper.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

https://doi.org/10.52202/079017-2127

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release data or models that have a high risk for misuse.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: We use common and standard Python libraries and write our own code for the experiments. Also, the data used in experiments is synthetic.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: We do not introduce new assets in the current paper.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: We perform theoretical analysis and numerical simulations on synthetic data, and the process is not related to human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: As clarified in the answer above, our study is not related to human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

https://doi.org/10.52202/079017-2127