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

Anonymous Authors[1]

Abstract

In-context learning (ICL) is a cornerstone of large language model 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.

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 ICL might not be part of the original training dataset. The success of ICL necessitates that the LLM performs certain learning processes during inference. While many previous works aim to demystify ICL from either empirical or theoretical perspectives, the theoretical foundations of ICL remain elusive, especially for complex tasks beyond simple linear regression. This leaves a gap in understanding how full-fledged transformer architectures facilitate ICL of more complex tasks, especially when there exist latent causal structures among the tokens in a sequence.

In this paper, we aim to narrow this gap by studying **how a two-attention-layer transformer is trained to perform ICL of a n-gram Markov chain model**, where each token in the Markov chain statistically depends on 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 (MHA), and a feed-forward network (FFN) layer with normalization. We employ such a transformer model to predict the $(L+1)$-th token of a 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 to the transformer model, which outputs a probability distribution over the vocabulary set for predicting the $(L+1)$-th token.

Under this setting, we aim to answer the following three questions: (i) *Does the gradient flow with respect to 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, where in the first stage, FFN learns the potential parent set; in the second stage, each attention head of the first MHA 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

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

"**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 chi-square mutual information. Finally, the second attention layer is a exponential kernel *classifier* that compares the features at each position with that created for the output position $L+1$, and use 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, which recovers 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.

2. Problem Setup: In-Context Learning of Markov Chains

2.1. In-Context Learning and n-Gram Markov Chains

We study autoregressive transformers trained for in-context learning (ICL). A pretrained transformer is a conditional distribution $f_{\texttt{tf}}(\cdot \,|\, \texttt{prompt})$ over a finite vocabulary $\mathcal{X}$, where $\texttt{prompt}$ is a sequence of tokens in $\mathcal{X}$. We consider unsupervised learning where $f_{\texttt{tf}}$ predicts the $(L+1)$-th token x_{L+1} given the prompt $x_{1:L}$ where the joint distribution of the sequence $x_{1:(L+1)}$ is sampled from a random n-gram Markov chain.

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 $x_{1:r_n}$. 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 $x_l \sim \pi(\cdot \,|\, X_{\texttt{pa}(l)})$ for $l = r_n + 1, \ldots, L+1$. See Figure 1 for an illustration.

Cross-Entropy (CE) 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, we adopt the population CE loss

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

where $\epsilon > 0$ is a small constant introduced for numerical stability. As a remark, we also relax the condition in Nichani et al. (2024) where they need the last token x_L 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}. See §E.3 for further discussion.

2.2. A Two-Layer Transformer Model

We consider a class of two-attention-layer transformer model $\texttt{TF}(M, H, d, D)$ that 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 the RPE window size, H is the number of attention heads, d is the vocabulary size, and D controls the complexity of the FFN. The details of $\texttt{TF}(M, H, d, D)$ are as follows.

Token Embedding, Input and Output. We take $\mathcal{X} = \{e_1, \ldots, e_d\}$ as the vocabulary. Given the input sequence $x_{1:L}$, we denote $X = (x_1, \ldots, x_L)^\top \in \mathbb{R}^{L\times d}$, and append a zero vector $\mathbf{0} \in \mathbb{R}^d$ to the sequence as the place-holder, defining $\widetilde{X} = (x_1, \ldots, x_L, \mathbf{0})^\top \in \mathbb{R}^{(L+1)\times d}$, and fed this extended sequence into the transformer. The output of at the "**0**" position is denoted by $y \in \mathbb{R}^d$.

Relative Positional Embedding. In the first attention layer, we use relative positional embeddings (RPE) to encode the 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 $W_P(i,j)$ to query and key positions (i,j) by

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

In other words, 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*. See Figure 2 for an illustration.

First Attention Layer. The input sequence is first processed by an 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 $L+1$, collected as $V^{(h)} = (v_1^{(h)}, \ldots, v_{L+1}^{(h)})^\top$. For any $l \in [L+1]$, $v_l^{(h)}$ is computed using RPE $W_P^{(h)}$ via

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

Feed-Forward Network with Normalization. After the first attention layer, we concatenate the outputs of the H attention heads and define $V = (V^{(1)}, \ldots, V^{(H)}) \in$

$\mathbb{R}^{(L+1)\times Hd}$. Consequenlty, for the l-the row of V which we denote by $v_l^\top$, we have $v_l^\top = (v_l^{(1)\top}, \ldots, v_l^{(H)\top})$ and for any vector $u \in \mathbb{R}^{Hd}$ in the sequel, we use the notation $u^\top = (u^{(1)\top}, \ldots, u^{(H)\top})$ with block $u^{(h)} \in \mathbb{R}^d$. With embedding dimension d_e, each row 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 $u, v \in \mathbb{R}^{Hd}$, we have

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

Here, the set $[H]_{\le D} = \{\mathcal{S} \subseteq [H] : |\mathcal{S}| \le D\}$ contains all subsets of $[H]$ with cardinality at most D, and $\{c_{\mathcal{S}} : \mathcal{S} \in [H]_{\le D}\}$ are the corresponding trainable parameters of $\phi(\cdot)$. An explicit definition of $\phi(\cdot)$ is available in Lemma E.1.

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 $C_D = \sum_{\mathcal{S} \in [H]_{\le D}} c_{\mathcal{S}}^2$. The normalization scheme is motivated by the popular layer normalization (Ba et al., 2016) in transformer architectures but without trainable parameters. See §B.3 for more discussions.

Second Attention Layer. We define normalized vector sequence as $U = (u_1, \ldots, u_{L+1})^\top$, which together with the original sequence $\widetilde{X}$ are then fed into the second attention layer. This attention layer has a single head and a scalar trainable parameter a. We let $U_{1:L} = (u_1, \ldots, u_L)^\top$ and let $\texttt{Mask}(\cdot)$ denote the mask that sets every entry of the first M rows of a matrix to be $-\infty$. The final output is given by

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

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 is just included here 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, a transformer model in $\texttt{TF}(M, H, d, D)$ consecutively applies the following operations:

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

The trainable parameters of the above transformer model are $\Theta = \big\{a, \{w_{-1}^{(h)}, \ldots, w_{-M}^{(h)}\}_{h\in[H]}, \{c_{\mathcal{S}} : \mathcal{S} \in [H]_{\le D}\}\big\}$. 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. 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.

3. Theoretical Results

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

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 $\mathtt{pa}$ 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 chi-squared mutual information, which is defined as follows: We let (z, Z) denote $(z_{l-M}, \ldots, z_l)$ under the stationary distribution μ^π with $\pi \sim \mathcal{P}$, where $z = z_l$, $Z = (z_{l-M}, \ldots, z_{l-1})$ and $\ell > M$.

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

where the expectation is taken over $\pi \sim \mathcal{P}, (z, Z) \sim \mu^\pi$, $\mu^\pi(z = \cdot \mid Z_{-\mathcal{S}})$ is the conditional distribution of z induced by μ^π given 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 chi-squared mutual information between two variables (Polyanskiy & Wu, 2024) and outputs a reweighted mutual information between $Z_{-\mathcal{S}}$ and z. Define $\mathcal{S}^\star$ as

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

As a remark, with the standard chi-squared mutual information, the optimal $\mathcal{S}^\star$ is the true parent set $\mathtt{pa}$ or a superset of it by the data processing inequality. However, sometimes a true parent can also bear little information about the target and a larger parent set tends to appear less frequently in the context sequence, leading to poor estimation accuracy. To handle this issue, the modification in (3.1) reaches a balance between the *information-richness* and the *model complexity*. See §B.5 for details.

Now we are ready to introduce the Generalized Induction Head (GIH) estimator. For given window size M, parent set degree D, The GIH estimator denoted by $\texttt{GIH}(\cdot; M, D)$ takes the sequence $x_{1:L}$ as input and outputs a vector $y^\star \in \mathbb{R}^d$ as distribution over $\mathcal{X}$ by

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

Here, we define $X_{l-\mathcal{S}^\star}$ as the set $\{x_{l-s} : s \in \mathcal{S}^\star\}$ and $N = \sum_{l>M} \mathbb{1}(X_{l-\mathcal{S}^\star} = X_{L+1-\mathcal{S}^\star})$. In a nutshell, the GIH estimator checks 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 satisfy this condition as the predicted distribution of x_{L+1}. Moreover, the GIH estimator is a generalization of the Induction Head mechanism (Elhage et al., 2021) to the stochastic setting with multiple parents. As we will show in §G.4, there exists a transformer model that implements GIH in its architecture. More importantly, we will show that gradient flow finds such a limiting model.

3.2. Convergence Guarantee of Gradient Flow

In the following, we present the convergence guarantee for gradient flow. To simplify our discussion, we consider the case where $H = M$. That is, there are enough heads to implement the GIH mechanism by letting each head copy a unique parent token from the window of size M. In the following, when we discuss the correspondence between "head" and "parent", we always refer to the mapping from head h to parent x_{l-h} for any $h \in [H]$ and $l > M$, which is without loss of generality. Let us first introduce the paradigm of gradient-flow training.

Training Paradigm. Now we train a transformer $\texttt{TF}(M, H, d, D)$ in (2.5) 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) with $f_{\texttt{tf}}$ replaced by the transformer model in (2.5) with parameter Θ. We train parameter Θ using gradient descent, under the ideal setting with infinite training data and infinitesimal step size. That is, we study the dynamics of gradient flow with respect to the loss $\mathcal{L}(\Theta)$:

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

To simplify the analysis, we consider a three-stage training paradigm where in each stage only one part of the weights gets trained. See §B.1 for a detailed table.

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

Theorem 3.1 (Convergence of Gradient Flow)**.** *Suppose Assumption B.1 and Assumption B.3 hold. Then the following holds for the three-stage training of gradient flow when L is sufficiently large.*

Stage I: Parent Selection by FFN. *Let $C_D(t) = \sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}(t)^2$ and $p_{\mathcal{S}^\star}(t) = c_{\mathcal{S}^\star}^2(t)/C_D(t)$. Then in the first stage of length $t_1 \asymp C_D(0)\log(L\log L)/(a(0)\Delta\widetilde{I}_{\chi^2})$, 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,*

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

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 of length $t_2 \asymp (L\log L)/(a(0)\Delta\widetilde{I}_{\chi^2})$, it holds for all $t \in [t_1, t_1 + t_2)$ that*

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

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 $a(t) \le \log(c_1/\delta)$, it holds that $\partial a(t) \asymp e^{a(t)}$; (ii) After $a(t)$ has grown such that $a(t) \ge \log(c_2/\delta)$, then $\partial_t a(t) \asymp 1/a(t)$ until it reaches the value $(1-\delta)\log L/4$.*

See §F for a proof sketch and §G for the detailed proof. An experimental demonstration for the three stages' s dynamics is in Figure 4. From Theorem 3.1, we can interpret that:

- The first stage's training on FFN is learning a ***selector*** that selects an informative set $\mathcal{S}^\star$ by realizing the corresponding feature embedding through the polynomial kernel.
- The second stage's training on the RPE turns the first attention layer into a ***copier*** by establishing the correspondence between the attention heads and the parents in the selected $\mathcal{S}^\star$.
- Given that the previous two stages have prepared the feature mapping ϕ such that $\langle\phi(v_l), \phi(v_{L+1})\rangle \approx \mathbb{1}(X_{l-\mathcal{S}^\star} = X_{L+1-\mathcal{S}^\star})$, the last stage enforces the GIH mechanism by increasing the scalar weight a in the second attention layer, which serves as an exponential kernel ***classifier***. The two sub-stages with distinct growth rates can be clearly seen from Figure 4(c), where $\partial a(t)$ is initially large and gradually decays.

In fact, we theoretically show that the limiting model upon convergence implements the GIH mechanism with τ going to infinity up to an $O(L^{-(1-\delta)/4})$ error. We defer the formal statement and proof to §G.4. 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***.

Another takeaway from Theorem 3.1 is that the FFN layer evolves exponentially faster than the RPE in the first attention layer, suggesting that we can actually train them together without splitting the first two stages. Indeed, this is validated by experiments in §D.

References

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

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

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

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

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

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

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

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

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

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

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

Chen, S., Yang, D., Li, J., Wang, S., Yang, Z., and Wang, Z. Adaptive model design for markov decision process. In *International Conference on Machine Learning*, pp. 3679–3700. PMLR, 2022.

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

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

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

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

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

Edelman, B. L., Goel, S., Kakade, S., and Zhang, C. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pp. 5793–5831. PMLR, 2022.

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

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

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

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

Giannou, A., Rajput, S., Sohn, J.-Y., Lee, K., Lee, J. D., and Papailiopoulos, D. Looped transformers as programmable computers. In Krause, A., Brunskill, E., Cho, K., Engelhardt, B., Sabato, S., and Scarlett, J. (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 11398–11442. PMLR, 23–29 Jul 2023.

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

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

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

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

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

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

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

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

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

Li, Y., Huang, Y., Ildiz, M. E., Rawat, A. S., and Oymak, S. Mechanics of next token prediction with self-attention. In *International Conference on Artificial Intelligence and Statistics*, pp. 685–693. PMLR, 2024.

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

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

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

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

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

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

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

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

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

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

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

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

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

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

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

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

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

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

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

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

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

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

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

Von Oswald, J., Niklasson, E., Randazzo, E., Sacramento, J., Mordvintsev, A., Zhmoginov, A., and Vladymyrov, M. Transformers learn in-context by gradient descent. In *International Conference on Machine Learning*, pp. 35151–35174. PMLR, 2023.

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

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

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

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

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

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

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

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

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

A. Organization of The Appendix

The appendices are organized as follows:

- In §B, we provide details omitted from the main text due to space constraints.
- In §C, we present an in-depth discussion on the related works.
- In §D, we discuss the experimental details.
- In §E, we provide explicit expressions for the FFN realizing a low-degree polynomial kernel, and review basics related to concepts mentioned in the main text.
- In §F, we provide a high-level overview of the proof of our main results.
- In §G, we present the proof for Theorem 3.1.
- In §H, we collect auxiliary results used in the proof of Theorem 3.1.

B. Additional Details for The Main Text

B.1. Table for Training Stages

Stage	Block to Train	Weights to Train	Duration
Stage I	FFN, layer 1	$\{c_{\mathcal{S}}\}_{\mathcal{S}\in[H]_{\le D}}$	$t_1 \asymp (C_D(0)\log L)/(a(0)\Delta\widetilde{I}_{\chi^2})$
Stage II	Attention RPE, layer 1	$\{w^{(h)}\}_{h\in[H]}$	$t_2 \asymp (L\log L)/(a(0)\Delta\widetilde{I}_{\chi^2})$
Stage III	Attention weight, layer 2	a	-

The three-stage training paradigm is presented in the above table. Specifically, we train the FFN layer in the first stage, then the first attention layer in the second stage, and finally the second attention layer in the last stage. In each stage, the parameters of other components of the model are frozen.

B.2. Figures for Illustration and Experiment Results

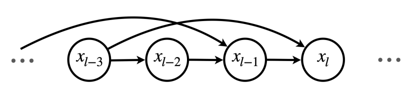

Figure 1. A 2-gram Markov chain with parent set $\mathtt{pa} = \{-1, -3\}$.

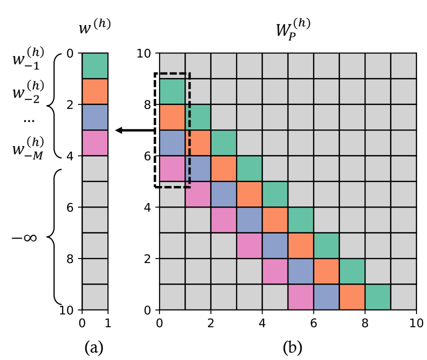

Figure 2. Illustration of the relationship between RPE vector $w^{(h)}$ and corresponding matrix $W_P^{(h)}$.

Illustration of Figure 4 on the three training stages:

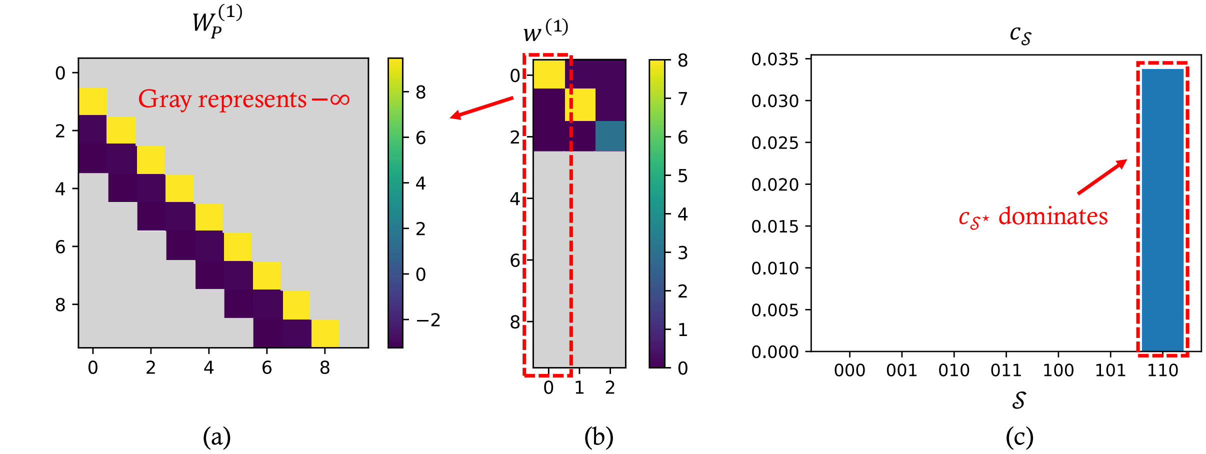

Figure 3. Limiting model of $\texttt{TF}(M = 3, H = 3, d = 3, D = 2)$ trained using gradient descent with $L = 100$, $\texttt{pa} = \{-1, -2\}$: (a) The top left 10 by 10 block of $W_P^{(1)}$ that attends to the -1 parent. (b) The RPE weight heatmap for all 3 heads. (c) One $c^\star_{\mathcal{S}}$ dominates. Here, $\mathcal{S}^\star$ represented by "110" means that $\mathcal{S}^\star = \{1, 2\}$, which is the exact parent set.

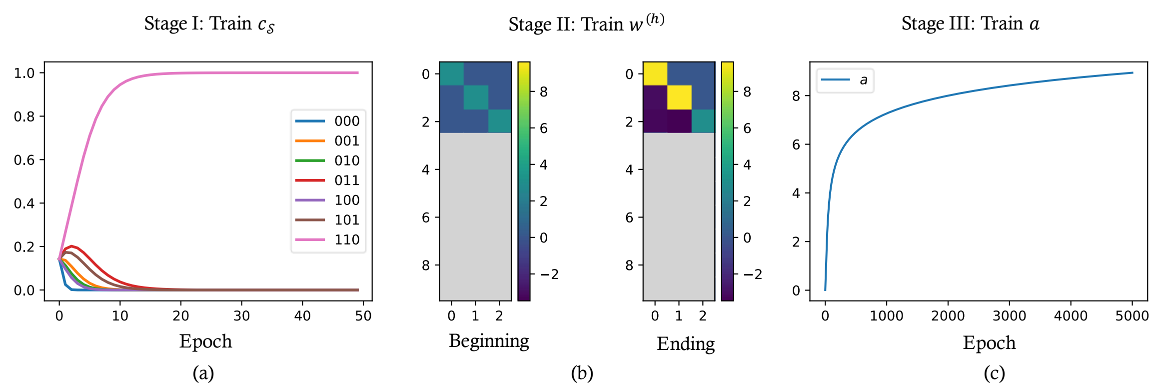

Figure 4. Training courses of 3 stages for $\texttt{TF}(M = 3, H = 3, d = 3, D = 2)$ trained with $L = 100$, $\texttt{pa} = \{-1, -2\}$. (a) In Stage I, a dominating $c_{\mathcal{S}^\star}$ was learned with $\mathcal{S}^\star = \{1, 2\}$ being the exact parent set. (b) In Stage II, the first two heads were trained to attend to parents -1 and -2, respectively. (c) In Stage III, the value of a increased monotonically.

- The first stage's training on FFN is learning a ***selector*** that selects an informative set $\mathcal{S}^\star$ by realizing the corresponding feature embedding through the polynomial kernel. In Figure 4(a), $\mathcal{S}^\star = \{1, 2\}$, and $c_{\mathcal{S}^\star}$ immediately dominates within only a few gradient steps.
- The second stage's training on the RPE turns the first attention layer into a ***copier*** by establishing the correspondence between the attention heads and the parents in the selected $\mathcal{S}^\star$. In Figure 4(b), the first two heads initialized towards the first two parents will deterministically copy parent -1 and -2 eventually while the third head is insignificant as $3 \notin \mathcal{S}^\star$. Also see Figure 3-(a) and (b).
- Given that the previous two stages have prepared the feature $\psi_{\mathcal{S}^\star}$ defined in (3.3), the last stage enforces the GIH mechanism by increasing the scalar weight a in the second attention layer, which serves as an exponential kernel ***classifier***. The two sub-stages with distinct growth rates can be clearly seen from Figure 4(c), where $\partial a(t)$ is initially large and gradually decays.

B.3. More Details on Layer Normalization

Recall that we have the normalization after the FFN layer as

$$u_l = \frac{\phi(v_l)}{C_D}.$$

To see this, 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 just copying some token in $x_{1:L}$ and since each token has unit norm, $\prod_{h\in\mathcal{S}}\langle v^{(h)}, v^{(h)}\rangle = 1$ and $\|\phi(v_l)\|_2 = \sqrt{C_D}$. Thus, u_l is close to the layer normalization $\phi(v_l)/\|\phi(v_l)\|_2$ (without trainable parameters). Such normalization is for simplifying the analysis and in later experiments in §D, we directly use this ℓ_2 layer normalization.

B.4. Assumptions for The Main Theorem

We introduce the following assumptions for our main theorem. We define the information gap within the D-degree parent set $[H]_{\le D}$ as $\Delta\widetilde{I}_{\chi^2} = \widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \max_{\mathcal{S}\in[H]_{\le D}\setminus\{\mathcal{S}^\star\}} \widetilde{I}_{\chi^2}(\mathcal{S})$, where we recall that $\mathcal{S}^\star$ maximizes the modified chi-squared mutual information as is defined in (3.2).

Assumption B.1 (Initialization)**.** *We assume that the following holds at initialization:*

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 related to the modified mutual information by*

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

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 boost the correspondence between parents and heads during the training by 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 $[L]$. This alignment enables 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.

Next, we present our assumptions on the Markov chain in the data generation process. To proceed, we define a $d^{r_n} \times d^{r_n}$ transition matrix P_π for the Markov chain as follows: Each row/column of P_π is indexed by the value of a length-r_n sequence of tokens $Z = (z_{-r_n}, \ldots, z_{-1})$ and each element indexed by tuple (Z', Z) is defined as $P_\pi(Z', Z) = \pi(z'_{-1} \mid Z_{\mathtt{pa}}) \cdot \mathbb{1}(Z'_{-r_n:-2} = Z_{-(r_n-1):-1})$. Note that P_π is a stochastic matrix but with zero entries due to the indicator. We need the following notion of the primitive matrix to state our assumption on P_π.

Definition B.2 (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 §E.2. By the celebrated Perron-Frobenius theorem, if P_π is 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 B.3 (Markov Chain)**.** *For any $\pi \in \mathrm{supp}(\mathcal{P})$, we assume that:*

1. *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 $|\lambda_2(P_\pi)| \le \lambda$.*
2. *There exists $\gamma > 0$ such that the transition kernel satisfies $\pi(x \mid X_{\mathtt{pa}}) \ge \gamma$ for any $(x, X_{\mathtt{pa}})$.*

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. In addition, the second assumption implies a lower bound on the probability for any $\mathcal{S} \subseteq [M]$ under the stationary distribution, i.e., $\mu^\pi(X_{-\mathcal{S}}) \ge \gamma^{|\mathcal{S}|}$.

B.5. Further Discussions on The Main Theorem

On the Modified Mutual Information. Now that we have shown how gradient flow approaches the desired GIH model, it is then natural to ask what is the optimal subset $\mathcal{S}^\star$ that the model selects and how well the model performs. For the purpose of illustration, let us consider a special 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 as

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

where $I_{\chi^2}(\mathcal{S})$ is the standard chi-squared 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., $\mathcal{S}^\star = \mathtt{pa}$.

1. If $n = 1$, i.e., each token only has one parent, then $\mathcal{S}^\star = \mathtt{pa}$. This is because $\mathcal{S}^\star$ simultaneously maximizes both terms in (B.2), 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 $\mathcal{S}^\star = \mathtt{pa}$. 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 behave poorly for small L, as the exact subsequence $X_{l-\mathcal{S}} = X_{L+1-\mathcal{S}}$ may appear rarely in the history.

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 B.4. $|\mathcal{S}^\star| \le n$ *regardless of the degree* D.

The rationale is that any $\mathcal{S}$ such that $|\mathcal{S}| > n$ has the mutual information no larger than the exact parent set $\mathtt{pa}$, while incurring a larger penalty on the model complexity. In other words, when $D > n$, minimizing $\log \widetilde{I}_{\chi^2}(\mathcal{S})$ encourages the model to become simpler, meanwhile solving the ICL task. Furthermore, if $D < n$, minimizing the modified chi-squared mutual information will instead become a constrained optimization problem.

C. 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 capabilities in performing reasoning steps based on provided prompts, 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).

Recent works focus on the setting of ICL to develop a theoretical understanding of transformers from different perspectives. A key perspective is the Bayesian view, which explores how transformers can be understood through the lens of Bayesian inference (Xie et al., 2021; Muller et al., 2021; Zhang et al., 2022; 2023b; Ahuja et al., 2023; Jeon et al., 2024). Another significant area of investigation examines how transformers internally execute specific algorithms to solve ICL tasks. This line of work uncovers the intricate mechanisms through which transformers perform these tasks (Akyürek et al., 2023; Von Oswald et al., 2023; Bai et al., 2023; Fu et al., 2023; Ahn et al., 2023; Mahankali et al., 2023; Giannou et al., 2024).

Furthermore, researchers study the statistical complexities of in-context learning (ICL), focusing on how transformers manage various statistical challenges (Wu et al., 2023; Cheng et al., 2023; Guo et al., 2023; Collins et al., 2024). There is also substantial interest in understanding how ICL operates over data drawn from Markov chains, providing insights into transformer behaviors in these specific data environments (Collins et al., 2024; Edelman et al., 2024; Makkuva et al., 2024; Chen & Zou, 2024), and with extension to in-context decision making (Lin et al., 2023; Sinii et al., 2023). 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 & Zhong, 2023; Deora et al., 2023; Chen & Li, 2024; Rajaraman et al., 2024).

On the other hand, understanding training dynamics from an optimization perspective is crucial for comprehending how transformers implement the ICL algorithm. The training dynamics for one layer attention are 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 & 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.

Induction Head. (Elhage et al., 2021) introduces the concept of "induction heads" as the mechanism underlying the ICL capabilities of transformers. At a high level, the induction head mechanism works by matching the history of the current token with those have been seen previously in the sequence and then predicting the next token based on the matched historical sub-sequences. (Olsson et al., 2022) provides 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. Also, a wider range of functionalities exhibited by induction heads that interact with various other mechanisms has been observed by (Wang et al., 2022). On the theory side, (Nichani et al., 2024) studies the ICL of first-order Markov chains using a two-layer transformer and demonstrates the formation of the induction head mechanism.

Most related to our work is the recent paper by 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. In comparison, the analysis in Nichani et al. (2024) applies to Markov chains where each token has at most one parent, while our setting encompasses general n-gram Markov chains where each token can have multiple parent tokens. 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.

D. Details of Experiments

In this section, we present the simulation results of $\texttt{TF}(M, H, d, D)$ in (2.5) which performs ICL on the n-gram Markov chain model introduced in §2.1.

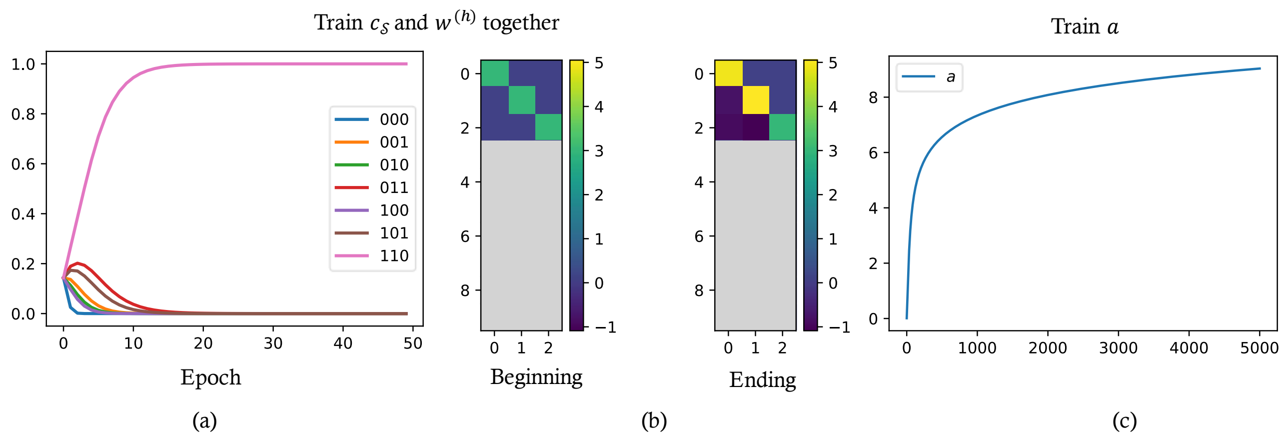

Figure 5. The evolution of gradient descent dynamics where we first train the first attention layer and the FFN together and then train the second attention layer. We plot the evolution of parameter $\{c_{\mathcal{S}}, \mathcal{S} \in [H]_{\le D}\}$, $\{W_P^{(h)}\}_{h\in[H]}$, and a respectively. Here we train a transformer $\texttt{TF}(M, H, d, D)$ with $M = H = 3$, $d = 3$, and $D = 2$, the number of input token is $L = 100$, and Markov chain has parent set $\texttt{pa} = \{-1, -2\}$. (a) A dominating $c_{\mathcal{S}^\star}$ was learned for $\mathcal{S}^\star = \mathrm{pa}$, i.e., the model selects the true parent set. This can be seen by observing that the line with the label "110" increases to about 1.0 while other lines decrease to nearly zero. (b) The first two attention heads, corresponding to the first two rows in the plotted matrix, became concentrated on the -1 and -2 parents, respectively. While the third attention head stays insignificant as the parent set pa contains only two elements. This can be seen by noticing that the top two diagonal entries after training have larger values than their initial values as well as those of all other entries. (c) The weight of the second attention layer, a, increased monotonically. In particular, it grew rapidly during the initial steps, and then the growth slowed down.

Data generation. The dataset for the ICL task was generated using n-gram Markov Chains as described in §2.1. We randomly sampled 10,000 Markov Chains with $L = 100$ from the prior distribution $\mathcal{P}$; 9,000 were used for training and 1,000 for validation. Each Markov Chain has 2 parents, *i.e.*, $|\mathrm{pa}| = 2$. Each token was embedded to $d = 3$. The prior

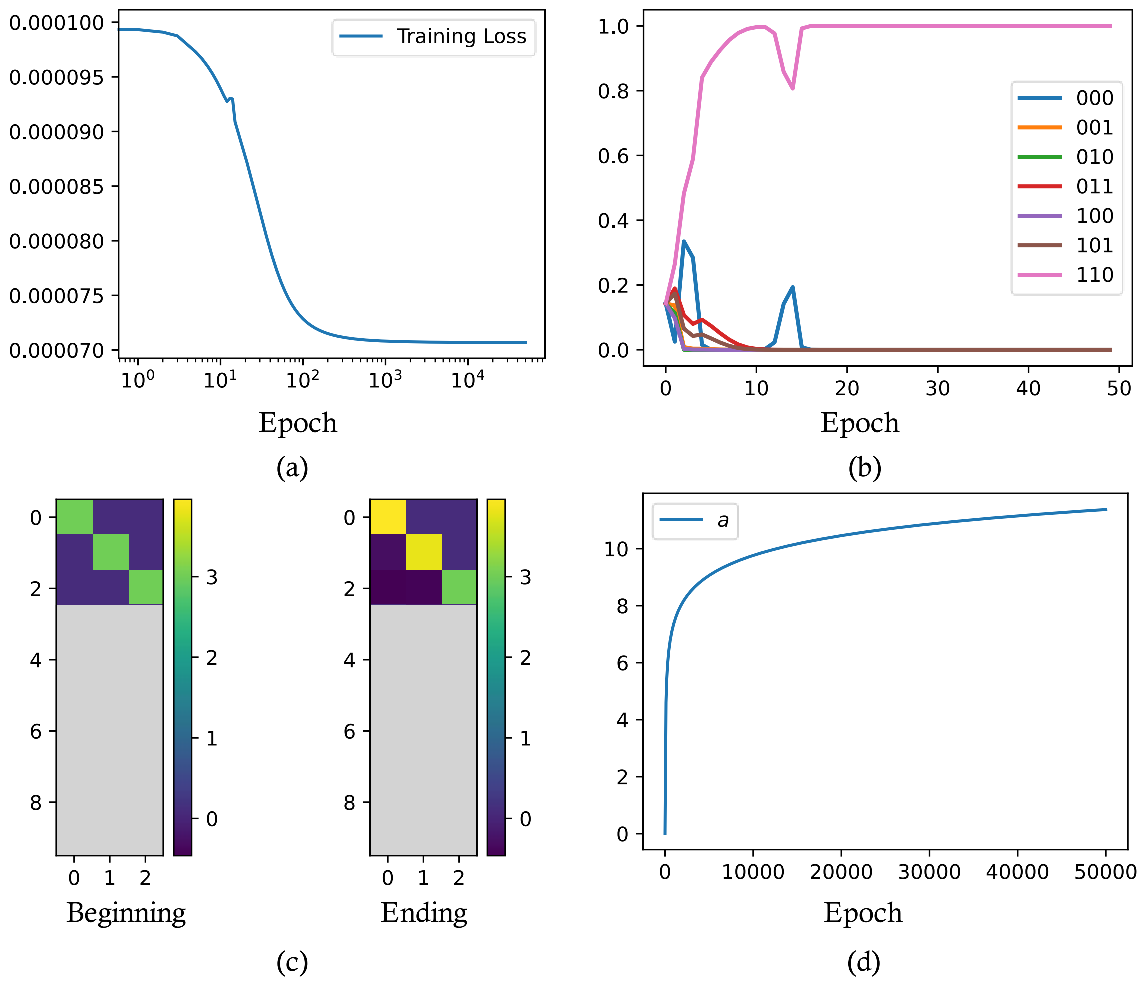

Figure 6. The evolution of gradient descent dynamics where we train the whole limiting model directly. We plot the training loss, the evolution of parameter $\{c_{\mathcal{S}}, \mathcal{S} \in [H]_{\le D}\}$, $\{W_P^{(h)}\}_{h\in[H]}$, and a respectively. Here we train a transformer $\texttt{TF}(M, H, d, D)$ with $M = H = 3$, $d = 3$, and $D = 2$, the number of input token is $L = 100$, and Markov chain has parent set $\texttt{pa} = \{-1, -2\}$. (a) The training loss curve of the model. (b) A dominating $c_{\mathcal{S}^\star}$ was learned for $\mathcal{S}^\star = \texttt{pa}$, i.e., the model selects the true parent set. This can be seen by observing that the line with the label "110" increases to about 1.0 while other lines decrease to nearly zero. (c) The first two attention heads, corresponding to the first two rows in the plotted matrix, became concentrated on the -1 and -2 parents, respectively. While the third attention head stays insignificant as the parent set $\texttt{pa}$ contains only two elements. This can be seen by noticing that the top two diagonal entries after training have larger values than their initial values as well as those of all other entries. (d) The weight of the second attention layer, a, increased monotonically. In particular, it grew rapidly during the initial steps, and then the growth slowed down.

distribution $\mathcal{P}$ is defined such that each row of the transition matrix of kernel π is independently drawn from a Dirichlet distribution with parameter $\alpha = 0.01$, i.e., $\pi(\cdot|x_{\mathrm{pa}(l)}) \sim \mathrm{Dir}(\alpha \cdot \mathbf{1}_{d^n})$.

Model initialization. We configured the model with three heads ($H = 3$) and window size ($M = 3$). The relative position encoding (RPE) weight matrix $W_P^{(h)}$ was initialized such that the $(-i)$-th diagonal of $W_P^{(h)}$ was set to $w_{-i}^{(h)}$ for $i = 1, 2, \ldots, M$, while all other entries were initialized to $-\infty$. We set $w^{(h)} = \rho e_h$, using a large positive value $\rho = 3$ to ensure that the h-th head focuses on the $-h$-th position. For other entries not set to $-\infty$, we assigned a value of 0.01. All $c_{\mathcal{S}}$ were initialized to 0.01. The initial value of a was set to 0.01.

Training settings. The models were trained using gradient descent with the cross-entropy loss function and a constant learning rate ($\lambda = 1$) for all stages. We trained the model in Stage I ($c_{\mathcal{S}}$) for 2000 epochs, in Stage II ($w^{(h)}$) for 50,000 epochs, and in Stage III (a) for 5000 epochs, respectively. The training was performed at a low degree ($D = 2$). All experiments were conducted using a single Nvidia A100 GPU.

Upon convergence, the weights of the trained disentangled transformers exhibited consistent structures, as shown in Figure 3. Specifically, $c_{\mathcal{S}^\star}$ dominated the ratio of $c_{\mathcal{S}^\star}^2 / \sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2$. Furthermore, heads $w^{(1)}$ and $w^{(2)}$ converged to the relative positions of the Markov parents.

In addition to separately training the first attention layer and the FFN (Figure 4), we demonstrate that these two components can be trained together, as illustrated in Figure 5. We remark that the learning behavior of the model under these two distinct paradigms is similar.

D.1. Additional Experiments

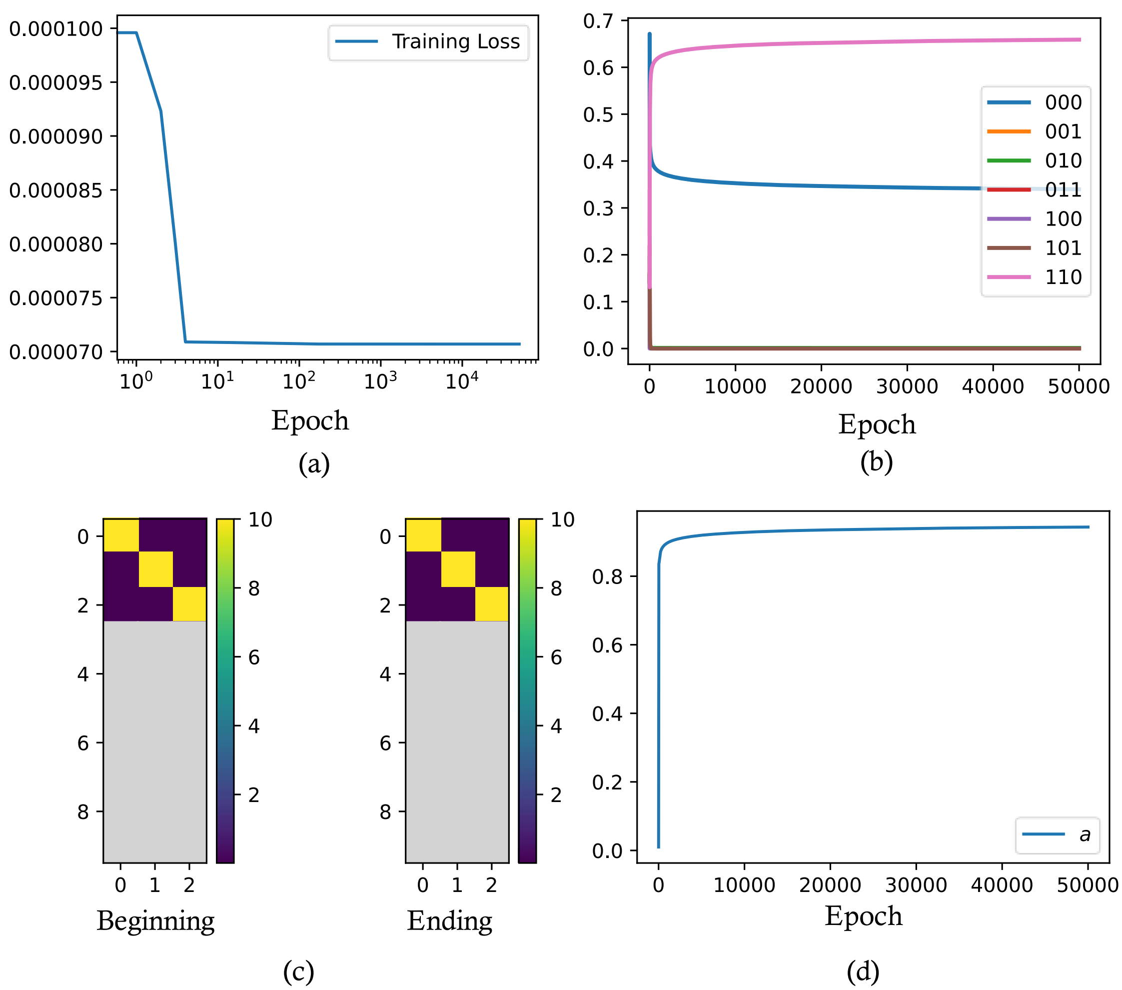

Figure 7. The evolution of gradient descent dynamics where we directly train the modified full model. We plot the training loss, evolution of parameter $\{c_{\mathcal{S}}, \mathcal{S} \in [H]_{\le D}\}$, $\{W_P^{(h)}\}_{h\in[H]}$, and a respectively. Here we train a transformer $\mathtt{TF}(M, H, d, D)$ with $M = H = 3$, $d = 3$, and $D = 2$, the number of input token is $L = 100$, and Markov chain has parent set $\mathtt{pa} = \{-1, -2\}$. (a) The training loss of the model. (b) A relatively dominating $c_{\mathcal{S}^\star}$ was learned for $\mathcal{S}^\star = \mathtt{pa}$, i.e., the model selects the true parent set. This can be seen by observing that the line with the label "110" increases to above 0.6 while other lines decrease to nearly zero. (c) The first two attention heads, corresponding to the first two rows in the plotted matrix, became concentrated on the -1 and -2 parents, respectively. While the third attention head stays insignificant as the parent set $\mathtt{pa}$ contains only two elements. This can be seen by noticing that the top two diagonal entries after training have larger values than their initial values as well as those of all other entries. (d) The weight of the second attention layer, a, increased monotonically. In particular, it grew rapidly during initial steps, and then the growth slowed down.

Previously, we show the simulation results on the simplified model in §D. Now we demonstrate additional experiments on the full model as follows.

$$\begin{aligned}
&\textbf{First Attention:} && \widetilde{V}^{(h)} = \sigma\big(\widetilde{X} W_{QK}^{(h)} \widetilde{X}^\top + W_P^{(h)}\big) \widetilde{X} W_{OV}^{(h)\top} && \in \mathbb{R}^{(L+1)\times d} \\
&\textbf{Concat \& Norm:} && V = \mathrm{LN}\big([\widetilde{V}^{(1)}, \ldots, \widetilde{V}^{(H)}, \widetilde{X}]\big) && \in \mathbb{R}^{(L+1)\times (H+1)d} \\
&\textbf{Feed Forward:} && \widetilde{U} = \phi(V) && \in \mathbb{R}^{(L+1)\times d_e} \\
&\textbf{Concat \& Norm:} && \widetilde{X}' = \mathrm{LN}([\widetilde{U}, V]) && \in \mathbb{R}^{(L+1)\times ((H+1)d + d_e)} \\
&\textbf{Second Attention:} && Y = \sigma\big(\widetilde{X}' W_{QK} \widetilde{X}'^\top\big) \widetilde{X}' W_{OV}^\top && \in \mathbb{R}^{(L+1)\times d}
\end{aligned}$$

For the second attention layer, we only use the last row $y = Y_{L+1}$ as the output. Here, LN denotes the ℓ_2 layer normalization without trainable parameters. For head h of the first attention layer, $W_P^{(h)}$ is the relative positional embedding matrix, $W_{QK}^{(h)}$ and $W_{OV}^{(h)}$ are the weight matrices for the query-key, value, and output projections, respectively; $\phi : \mathbb{R}^{(H+1)d} \to \mathbb{R}^{d_e}$ is a feed-forward network; and finally, W_{QK} and W_{OV} are the query-key matrix and output projection matrix for the second attention layer. In comparison to the simplified model in (2.5), here we incorporate all query, key, and output projections as in a standard transformer architecture. Also, we replace the normalization by a $\sqrt{C_D}$ factor with the usual ℓ_2 layer normalization, though they have similar functionality.

Our training setup is similar to that in §D. We used the same dataset and training settings (except the number of training epochs).

We initially attempted to train the full model directly, but this approach was ineffective. Consequently, we adopted an alternative strategy. Specifically, for the first layer, we used all components of the full model together except for the query-key projection weight $W_{QK}^{(h)}$. For the second layer, we utilized a simplified version similar to the one with polynomial kernel weights, but we incorporated an additional ReLU operation to avoid negative values for each product due to the use of value and output projection $W_{OV}^{(h)}$. Both $W_{QK}^{(h)}$ and $W_{QK}^{(h)}$ were initialized as identity matrices scaled by 0.001. Unlike the simplified model, we initialized the RPE vector $w^{(h)}$ deterministically as $w^{(h)} = \rho e_h$ with $\rho = 10$. We trained the full model with all parameters together for 50,000 epochs. As illustrated in Figure 7, the full model converged to a state comparable to our simplified model.

E. Additional Background and Discussions

E.1. Feed-Forward Network for Polynomial Kernel

Lemma E.1. *Recall that we define the feed-forward network (FFN) in (2.3), which maps a vector in $z \in \mathbb{R}^{dH}$ 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 $h \in [H]$. Then we can explicitly write $\phi(\cdot)$ by letting*

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

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

$$d_e = \sum_{\mathcal{S}\in[H]_{\le D}} d^{|\mathcal{S}|}. \tag{E.2}$$

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 (E.2).

It remains to verify (2.3) with $\phi(\cdot)$ defined in (E.1). To this end, we note that for any $u, v \in \mathbb{R}^{dH}$ and any $\mathcal{S} \in [H]_{\le D}$, we have

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

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

E.2. 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 matrix.

Definition E.2 (Irreducible Matrix). *A nonnegative square matrix $P \in \mathbb{R}_+^{d\times d}$ is called irreducible if the induced directed graph $\mathcal{G}$ is strongly connected, i.e., there always exists a directed path that connects any two given nodes within the graph. Here, the induced graph $\mathcal{G}$ is defined 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 a 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 our study, 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, which enables us to learn with a finite sequence length L. To achieve that, one typically needs the second largest magnitude of the eigenvalues of P_π, which we denote by λ, to be bounded below 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 P_π is has all positive entries, then it 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. 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 E.3 (Primitive Matrix). *A nonnegative and irreducible square matrix P is called primitive if there exists an integer k such that P^k has all positive entries.*

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

Theorem E.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.*

E.3. Sequential CE Loss

We define the sequential CE loss as

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

One can equivalently view this sequential CE loss as a mixing of the CE loss defined in (2.1) with different sequence length. Note that by Assumption B.3, the chain is sufficiently mixed for large L and changing the sequence length does not influence the stationary distribution. Intuitively, this means that if we pick another large L' different from L and look back at all the history up to L', the history will be very similar to that at L in distribution. In fact, the gradient on the transformer weights will converge fast (as long as we have spectral gap in the transition matrix P_π) to a limiting value independent of L. Suppose the mixing time is $L_0 \ll L$. Then, for $l = L_0, \ldots, L$, our analysis still holds, and for $l < L_0$, it suffices to sacrifice an additional $L_0/L = o(1)$ error. In the proof, however, we only consider the last token's CE loss in (2.1) to simplify the analysis.

E.4. Standard Chi-squared Divergence and Mutual Information

The chi-squared divergence (or chi-squared distance) between two probability distributions P and Q over the same probability space is defined as:

$$D_{\chi^2}(P\|Q) = \sum_x \frac{(P(x) - Q(x))^2}{Q(x)},$$

where the summation is taken over all elements x in the sample space where $Q(x) > 0$. The chi-squared 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 chi-squared mutual information satisfies the data processing inequality

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

which follows from the observation that chi-squared divergence is also an f-divergence.

F. Proof Sketch

In this section, we discuss the main ingredients of analysis of gradient flow. First, we show in §F.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 in Appendices F.2 to F.4. At a high level, the gradient flow dynamics can be decomposed into three stages, which feature one of the following behaviors respectively.

- 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 $\mathcal{O}(\log L)$. In this case, the learned model approximately implements the GIH mechanism $\texttt{GIH}(x_{1:L}; M, D, \tau)$ with $\tau = +\infty$.

F.1. Simplification of the Transformer Model at Initialization

In the following, we simplify the expression of the transformer model under Assumption B.1 for initialization. Specifically, we will show that the attention scores of the second attention layer admit a simpler form.

For the second attention layer, we write the output as $y^\top = \sigma(a \cdot s^\top)X$ where $s := u_{L+1}^\top \texttt{Mask}(U_{1:L}^\top)$ is the vector of similarity scores. Recall from (2.5) that $U = \phi(V)/\sqrt{C_D}$. Hence, the l-th row of U is given 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 $-\infty$. From (2.3) we have

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

Note that under Assumption B.1, by the definition of Δw in (B.1), we have $w_{-h}^{(h)} \gg w_{-j}^{(h)}$ for $j \neq h$ at initialization. Thus, the output of the first attention layer satisfies

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

Here we use the fact that Δw is sufficiently large, which makes the softmax function collapse to a one-hot vector approximately. This further implies that for $l = M+1, \ldots, L$, we have

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

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

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

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

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

F.2. Stage I: Optimal Subset Selection

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 transformer output is $y = (\sigma(a \cdot s^\top)X)^\top$ 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 $\ell(\Theta) = \langle x_{L+1}, \log(y + \varepsilon\mathbf{1})\rangle$. By direct calculation, we have

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

where $x_{L+1}/(y + \varepsilon\mathbf{1})$ is obtained by element-wise division and $\sigma_l(\cdot)$ denotes the l-th entry of the softmax function. Furthermore, by the expression of s_l in (F.1), we have

$$\frac{\partial s_l}{\partial c_{\mathcal{S}}} = \frac{2c_{\mathcal{S}} \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle}{\sum_{\mathcal{S}'\in[H]_{\le D}} c_{\mathcal{S}'}^2} - \frac{2c_{\mathcal{S}} s_l}{\sum_{\mathcal{S}'\in[H]_{\le D}} c_{\mathcal{S}'}^2}, \quad \text{for each } \mathcal{S} \in [H]_{\le D}. \tag{F.4}$$

By applying the chain rule and combining (F.3) and (F.4), we get

$$\begin{aligned}
\partial_t \log c_{\mathcal{S}}^2 &= \frac{2}{c_{\mathcal{S}}} \partial_t c_{\mathcal{S}} = -\frac{2}{c_{\mathcal{S}}} \frac{\partial \mathcal{L}}{\partial c_{\mathcal{S}}} = -\frac{2}{c_{\mathcal{S}}} \sum_{l=M+1}^{L} \mathbb{E}\left[\frac{\partial \ell}{\partial s_l} \frac{\partial s_l}{\partial c_{\mathcal{S}}}\right] \\
&= \frac{4a}{\sum_{\mathcal{S}'\in[H]_{\le D}} c_{\mathcal{S}'}^2} \sum_{l=M+1}^{L} \mathbb{E}\left[\sigma_l(a\cdot s^\top)\left(\prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle - s_l\right)\left(\frac{x_{L+1}}{y+\varepsilon\mathbf{1}}\right)^\top (x_l - y)\right].
\end{aligned}$$

Note that $C_D = \sum_{\mathcal{S}'\in[H]_{\le D}} c_{\mathcal{S}'}^2$. Also note that y is a vector in $\mathbb{R}^d$. We let $y(k)$ denote the k-th entry of y for all $k \in [d]$. Now utilizing the approximation in (F.2) and expanding $(x_{L+1}/(y + \epsilon\mathbf{1}))^\top (x_l - y)$, the above dynamics can be further simplified as

$$\begin{aligned}
\partial_t \log c_{\mathcal{S}}^2 &\approx \frac{4a}{C_D} \sum_{l=M+1}^{L} \mathbb{E}\left[\sigma_l(as^\top)\left(\prod_{h\in\mathcal{S}} \mathbb{1}\{x_{l-i} = x_{L+1-i}\} - s_l\right)\left(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - 1\right)\right] \\
&\approx \frac{4a}{(L-M)C_D} \sum_{l=M+1}^{L} \mathbb{E}\left[\left(\prod_{h\in\mathcal{S}} \mathbb{1}\{x_{l-i} = x_{L+1-i}\}\right)\left(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - 1\right)\right] \\
&\quad + \underbrace{\frac{4a}{(L-M)C_D} \sum_{l=M+1}^{L} \mathbb{E}\left[s_l \cdot \left(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - 1\right)\right]}_{f(t)}
\end{aligned} \tag{F.5}$$

where in the first line we use the fact that both x_{L+1} and x_l are one-hot vectors and that ϵ is small, and the second approximation is due to the fact that $\sigma_l(as^\top) \approx 1/(L-M)$ when a is small. We will prove that the first term in the resulting approximation can be further approximated using the modified chi-squared mutual information $\widetilde{I}_{\chi^2}(\mathcal{S})$ when L is large, which is introduced in **??**.

Therefore, it follows from (F.5) that

$$\partial_t \log c_{\mathcal{S}}^2(t) \approx \frac{4a}{C_D(t)} \widetilde{I}_{\chi^2}(\mathcal{S}) - f(t). \tag{F.6}$$

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 $\mathcal{S}^*$ which maximizes the modified chi-square mutual information within $[H]_{\le D}$, i.e.,

$$\mathcal{S}^\star = \operatorname*{argmax}_{\mathcal{S}\in[H]_{\le D}} \widetilde{I}_{\chi^2}(\mathcal{S}).$$

Correspondingly, by normalization, we have $p_{\mathcal{S}^\star}$ goes to one at t increases. To determine the growth rate of $p_{\mathcal{S}^\star}$, we first note that $C_D(t) \equiv C_D(0)$ due to the normalization (see Lemma G.1). Combining this fact with the definition $p_{\mathcal{S}^\star} = c_{\mathcal{S}^\star}/C_D$, we can derive a lower bound for the growth rate of $p_{\mathcal{S}^\star}(t)$ from the dynamics of $\log c_{\mathcal{S}}^2(t)$ in (F.6):

$$\partial_t \log(1 - p_{\mathcal{S}^\star}) \le -\Omega\biggl(\frac{a\cdot \Delta\widetilde{I}_{\chi^2}}{C_D(0)}\biggr), \quad \text{where} \quad \Delta\widetilde{I}_{\chi^2} = \min_{\mathcal{S}\in[H]_{\le D}\setminus\{\mathcal{S}^\star\}} I_{\chi^2}(\mathcal{S}^\star) - I_{\chi^2}(\mathcal{S}).$$

Thus, the error $1 - p_{\mathcal{S}^\star}$ will decay to zero exponentially fast.

F.3. Stage II: Convergence of $\sigma(w^{(h)})$ to One-Hot Vector

As we proceed to the second stage after $p_{\mathcal{S}^\star}$ approaches one, we will prove how $\sigma(w^{(h)})$ converges to a one-hot vector e_{M+1-h} for each $h \in \mathcal{S}^\star$. Recall that we denote $X = (x_1, \dots, x_L) \in \mathbb{R}^{L\times d}$. 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 $l-M, \dots, l-1$ for any l. Recall that

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

To begin with, by chain rule, differentiating s_l with respect to $w_{-i}^{(h)}$ yields

$$\begin{aligned}
\frac{\partial s_l}{\partial w_{-i}^{(h)}} &= \sum_{\mathcal{S}\in[H]_{\le D}} p_{\mathcal{S}} \cdot \frac{\partial}{\partial w_{-i}^{(h)}} \prod_{h'\in\mathcal{S}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle \\
&= \sum_{\substack{\mathcal{S}\in[H]_{\le D}\\ \text{s.t } h\in\mathcal{S}}} p_{\mathcal{S}} \cdot \frac{\partial}{\partial w_{-i}^{(h)}} \prod_{h'\in\mathcal{S}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle \\
&= \sum_{\substack{\mathcal{S}\in[H]_{\le D}\\ \text{s.t } h\in\mathcal{S}}} p_{\mathcal{S}} \cdot \biggl(\prod_{h'\in\mathcal{S}, h'\neq h} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle\biggr) \cdot \frac{\partial \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle}{\partial w_{-i}^{(h)}} \\
&= \sum_{\substack{\mathcal{S}\in[H]_{\le D}\\ \text{s.t } h\in\mathcal{S}}} p_{\mathcal{S}} \prod_{\substack{h'\in\mathcal{S}\\ h'\neq h}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle b_l^\top (e_{M+1-i} - (\sigma^{(h)})^\top)\sigma_{-i}^{(h)},
\end{aligned} \tag{F.7}$$

where the second equality is because if $h \notin \mathcal{S}$, then $\prod_{h'\in\mathcal{S}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle$ does not depend on $w^{(h)}$; for the third equality, we define $b_l := X_{(l-M):(l-1)} v_{L+1}^{(h)} + X_{(L+1-M):L} v_{l+1}^{(h)}$ and and $\sigma^{(h)} = (\sigma_{-M}^{(h)}, \dots, \sigma_{-1}^{(h)}) \in \mathbb{R}^{1\times M}$ to simplify the notation. Moreover, here the outer summation indicates summation over all subsets in $[H]_{\le D}$ containing h. Then, similar to the derivation of $\partial_t \log c_{\mathcal{S}}^2(t)$, it follows from direct calculation that

$$
\begin{aligned}
\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} &= \sum_{l=M+1}^{L} \mathbb{E}\biggl[\frac{\partial \ell}{\partial s_l}\biggl(\frac{\partial s_l}{\partial w_{-h}^{(h)}} - \frac{\partial s_l}{\partial w_{-i}^{(h)}}\biggr)\biggr] \\
&= a \sum_{l=M+1}^{L} \mathbb{E}\biggl[\sigma_l(as^\top) \sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1}=e_k)\cdot(\mathbb{1}(x_l=e_k)-y(k))}{y(k)+\varepsilon}\biggl(\frac{\partial s_l}{\partial w_{-h}^{(h)}} - \frac{\partial s_l}{\partial w_{-i}^{(h)}}\biggr)\biggr]. \qquad \text{(F.8)}
\end{aligned}
$$

Furthermore, to simplify the notation, we define

$$
g_h := \sum_{l=1}^{L} \sum_{\substack{\mathcal{S}\in[H]_{\le D}\\ \text{s.t } h\in\mathcal{S}}} p_{\mathcal{S}} \cdot \mathbb{E}\biggl[\sigma_l(as^\top) \sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1}=e_k)\cdot(\mathbb{1}(x_l=e_k)-y(k))}{y(k)+\varepsilon} \prod_{\substack{h'\in\mathcal{S}\\ h'\neq h}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle b_l\biggr].
$$

Here, we absorb the inner product $\prod_{h'\in\mathcal{S},h'\neq h}\langle v_l^{(h')}, v_{L+1}^{(h')}\rangle$ into the definition of g_h. Combining (F.7), (F.8), and the definition of g_h, we have

$$
\begin{aligned}
&\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} \qquad \text{(F.9)}\\
&\quad = a\cdot g_h^\top \biggl(\sigma_{-i}^{(h)}\,(e_{M+1-h}-e_{M+1-i}) + \Bigl(\sigma_{-h}^{(h)} - \sigma_{-i}^{(h)}\Bigr)\sum_{j\neq h}\sigma_{-j}^{(h)}(e_{M+1-h}-e_{M+1-j})\biggr),
\end{aligned}
$$

We again apply the approximation in (F.2) and replace the sum over l by the expectation over the stationary distribution of the Markov chain (which is valid because L is large), which yields

$$
g_h \approx \mathbb{E}\biggl[(Zx_{-h}+Xz_{-h})\cdot \prod_{h'\in\mathcal{S}^\star\backslash\{h\}} \mathbb{1}(z_{-h'}=x_{-h'})\cdot\biggl(\sum_{k\in[d]}\frac{\mathbb{1}(z=x=e_k)}{\mu^\pi(e_k)} - 1\biggr)\biggr] \in \mathbb{R}^M, \qquad \text{(F.10)}
$$

where (Z,z) and (X,x) are independent samples from μ^π, the stationary distribution of the Markov chain over a window of size $M+1$. To simplify the notation, we treat Z and X as matrices, denoted by $Z=[z_{-M},\ldots,z_{-1}]^\top\in\mathbb{R}^{M\times d}$ and $X=[x_{-M},\ldots,x_{-1}]^\top\in\mathbb{R}^{M\times d}$. Here, each row in Z and X corresponds to a vector sampled from μ^π, representing the state of the Markov chain at different time steps within the window.

Next, we derive the lower bound of the $g_h^\top(e_{M+1-h}-e_{M+1-i})$ for all $i\neq h$ in (F.9). By (F.10), we have $g_h^\top e_{M+1-h}\approx \widetilde{I}_{\chi^2}(\mathcal{S}^\star)$. It further follows from the Cauchy-Schwarz inequality that

$$
\begin{aligned}
g_h^\top e_{M+1-i} &\approx \mathbb{E}\biggl[\mathbb{1}(x_{-i}=z_{-h})\prod_{h'\in\mathcal{S}^\star\backslash\{h\}}\mathbb{1}(x_{-h'}=z_{-h'})\biggl(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{y(k)} - 1\biggr)\biggr]\\
&\le \frac{\widetilde{I}_{\chi^2}(\mathcal{S}^\star)+\widetilde{I}_{\chi^2}((\mathcal{S}^\star\backslash\{h\})\cup\{i\})}{2} \le \widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \frac{1}{2}\cdot\Delta\widetilde{I}_{\chi^2}.
\end{aligned}
$$

Here recall that we define $\Delta\widetilde{I}_{\chi^2}=\widetilde{I}_{\chi^2}(\mathcal{S}^\star)-\max_{\mathcal{S}\in[H]_{\le D}\backslash\{\mathcal{S}^\star\}}\widetilde{I}_{\chi^2}(\mathcal{S})$, which is the gap between the information-optimal set $\mathcal{S}^*$ and any other subset of $[H]_{\le D}$ in terms of the modified chi-squared mutual information. Plugging this back to the gradient difference, we conclude that

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

Thus, so long as $\sigma_{-h}^{(h)}>\sigma_{-i}^{(h)}$ when the second stage starts, $w_{-h}^{(h)}$ will thereafter grow faster than $w_{-i}^{(h)}$ and $\sigma(w^{(h)})$ will converge to a one-hot vector e_{-h}. The convergence rate is given by

$$
1-\prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)}(t))^2 \le \frac{2|\mathcal{S}^\star|\cdot(M-1)}{a(0)\cdot\Delta\widetilde{I}_{\chi^2}\cdot\sigma_{\min}(0)\cdot t_2/2+\exp(\Delta w)+(M-1)}\wedge 1,
$$

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

F.4. Stage III: Growth of a

In the last stage, we turn to the training of a given that $\sigma(w^{(h)})$ has converged to one-hot vectors for all $h \in \mathcal{S}^\star$. The following approximation of the dynamics of $a(t)$ is performed in the region $a \le O(\log L)$, where the signal term in the dynamics dominates the approximation error.

After Stages I and II, the output is approximated as $y(k) \approx y^\star(k) := \sum_{l=1}^{L} \sigma_l^\star \, \mathbb{1}(x_l = e_k)$, where we define

$$\sigma_l^\star = \frac{\exp\left(a \cdot \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})\right)}{\sum_{l'=M+1}^{L} \exp\left(a \cdot \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l'-h} = x_{L+1-h})\right)},$$

By approximating the empirical distribution $y^\star(k)$ with the population distribution $\widetilde{\mu}^\pi(e_k|X_{-\mathcal{S}^\star})$, the gradient on a is given by

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

where the underlying joint distribution of $(x, X, \widetilde{z}, \widetilde{Z})$ is given by

$$q^\pi = \mu^\pi(x, X_{-\mathcal{S}^\star}) \cdot \widetilde{\mu}^\pi(\widetilde{z}, \widetilde{Z}_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star}),$$

and $\widetilde{\mu}^\pi(\widetilde{Z} \,|\, X)$ is defined as

$$\widetilde{\mu}^\pi(\widetilde{z}, \widetilde{Z} \,|\, X) \propto \mu^\pi(z, Z) \cdot \exp\left(a \cdot \mathbb{1}(X_{-\mathcal{S}^\star} = \widetilde{Z}_{-\mathcal{S}^\star})\right).$$

As a result, the gradient on a can be rewritten as

$$\partial_t a(t) \approx \mathbb{E}_{\pi\sim\mathcal{P},(x,X_{-\mathcal{S}^\star})\sim\mu^\pi}\left[\left(\sum_{k\in[d]} \frac{\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star})^2}{\widetilde{\mu}^\pi(\widetilde{z} = e_k \,|\, X_{-\mathcal{S}^\star})} - 1\right)\widetilde{\mu}^\pi(\widetilde{Z}_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star})\right] \tag{F.11}$$

As we consider the cases where a is sufficiently small or large, the lower and upper bounds of (F.11) can be derived respectively. For small values of a, it undergoes super-exponential growth until it reaches a critical "elbow" time, $e^{-a(0)} \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\left[\sum_{X_{\mathcal{S}^\star}} D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}), |, \mu^\pi(\cdot))\mu^\pi(X_{-\mathcal{S}^\star})^2\right]^{-1}$. For large values of a, it grows logarithmically until it reaches $\Omega(\log L)$.

G. Dynamics Analysis

Additional Notation. To simplify the notation, we ignore the Mask in the simplified model (2.5) and let the index l runs from 1 to L. If the out of range issue occurs, e.g., we have x_{l-M} for $l \le M$, we can safely treat those out-of-range tokens as zero vectors. In a summation with respect to l for $l \in [L]$, the total number of the occurrence of the out-of-range issues is no larger than $O(M)$. Thus, as long as $L \gg M$, it just gives an $O(M/L)$ additional error term, which does not influence our results. Recall the error $\Delta_1(t_1)$ and $\Delta_2(t_2)$ defined in Theorem 3.1. We further denote by Δ_1 the value of $\Delta_1(t_1)$ at the end of Stage 1. And Δ_2 is defined similarly.

G.1. Analysis for Stage I

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

Proof Strategy. At a high level, the strategy is to analyze $\partial_t \log c_{\mathcal{S}^\star}^2 - \partial_t \log c_{\mathcal{S}}^2 > 0$ for all $\mathcal{S} \ne \mathcal{S}^\star$ via the following steps:

1. **Dynamics Calculation.** First, we calculate the dynamics of $\log c_{\mathcal{S}}^2$ for a 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 $\log c_{\mathcal{S}}^2$ can be approximated using the modified mutual information $\widetilde{I}_{\chi^2}(\mathcal{S})$.

2. **Lower Bound for The Growth Rate.** Consequently, we are able to lower bound $\partial_t \log c_{\mathcal{S}^\star}^2 - \partial_t \log c_{\mathcal{S}}^2$ 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 derived lower bound on $\partial_t \log c_{\mathcal{S}^\star}^2 - \partial_t \log c_{\mathcal{S}}^2$.

The detailed proof is provided below.

G.1.1. Calculation of The Dynamics of $\log c_{\mathcal{S}}^2$

Recall that our simplified model is given by

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

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 $\pi \sim \mathcal{P}$.

In the sequel, we first consider a fixed $\mathcal{S} \in [H]_{\le D}$ and derive the dynamics of $c_{\mathcal{S}}^2$. We abbreviate $\sigma \equiv \sigma(as^\top)$ for convenience. By direct calculation, we have

$$\frac{\partial y}{\partial \sigma} = X^\top, \quad \frac{\partial \sigma}{\partial s_l} = a \cdot \sigma_l\left(a \cdot s^\top\right) \cdot (e_l^\top - \sigma), \quad \frac{\partial s_l}{\partial c_{\mathcal{S}}} = \frac{2c_{\mathcal{S}} \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle}{\sum_{\mathcal{S}'\in[H]_{\le D}} c_{\mathcal{S}'}^2} - \frac{2c_{\mathcal{S}} s_l}{\sum_{\mathcal{S}'\in[H]_{\le D}} c_{\mathcal{S}'}^2}.$$

Then applying the chain rule, we can calculate $\partial \ell / \partial s_l$ as follows

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

Further using the chain rule $\partial \ell / \partial c_{\mathcal{S}} = \sum_{l=1}^{L} \partial \ell / \partial s_l \cdot \partial s_l / \partial c_{\mathcal{S}}$ and the gradient flow formula that $\partial_t c_{\mathcal{S}}^2 = -2c_{\mathcal{S}} \cdot \partial \mathcal{L} / \partial c_{\mathcal{S}}$, we obtain the following dynamics for $c_{\mathcal{S}}^2$

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

Recall the notations $C_D := \sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2$ and $p_{\mathcal{S}} := c_{\mathcal{S}}^2 / \sum_{\mathcal{S}'\in[H]_{\le D}} c_{\mathcal{S}'}^2$. 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. As a result, $\mathbb{E}$ means the expectation of the sequence X for a fixed π if it is not specified. To simplify the expression of $\partial_t c_{\mathcal{S}}^2$, we define quantities $g_{0,\mathcal{S}}$ and f as

$$\begin{aligned} g_{0,\mathcal{S}} &:= \sum_{l=1}^{L} \mathbb{E}\left[\sigma_l\left(a \cdot s^\top\right) \cdot \sum_{k\in[d]} \left(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k) + \varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k) + \varepsilon} \right) \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle \right], \\ f &:= \sum_{l=1}^{L} \mathbb{E}\left[\sigma_l\left(a \cdot s^\top\right) \cdot \sum_{k\in[d]} \left(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k) + \varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k) + \varepsilon} \right) \cdot s_l \right]. \end{aligned} \tag{G.1}$$

Based on the definition, we can rewrite the dynamics as follows:

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

One can notice that C_D does not change during the train as described in Lemma G.1 and f does not depend on $\mathcal{S}$.

G.1.2. Preservation of C_D along The Gradient Flow

Lemma G.1. *The quantity $C_D = \sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2$ is preserved under the dynamics, i.e., $\partial_t C_D \equiv 0$.*

Proof of Lemma G.1. Plugging the definition of $g_{0,\mathcal{S}}$ and f into the dynamics of $c_{\mathcal{S}}^2$, we have

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

Then, we can derive the dynamics of C_D in the following.

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

where $p_{\mathcal{S}} := c_{\mathcal{S}}^2/\sum_{\mathcal{S}'\in[H]_{\le D}} c_{\mathcal{S}'}^2$. Thus, the quantity C_D is preserved under the dynamics. □

G.1.3. Approximation of $g_{0,\mathcal{S}}$

For the analysis of the dynamics of $c_{\mathcal{S}}^2$, we need to understand the quantities $g_{0,\mathcal{S}}$ and f. To approximiate $g_{0,\mathcal{S}}$, we introduce the following quantities:

$$g_{1,\mathcal{S}} := \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\left[\left(\sum_{k\in[d]}\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\bar{y}(k)+\varepsilon} - \frac{\bar{y}(k)\,\mathbb{1}(x_{L+1}=e_k)}{\bar{y}(k)+\varepsilon}\right)\cdot\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\right], \tag{G.3}$$

$$g_{2,\mathcal{S}} := \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\left[\left(\sum_{k\in[d]}\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\mu^\pi(e_k)} - 1\right)\cdot\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\right], \tag{G.4}$$

$$g_{3,\mathcal{S}} := \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\left(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(e_k)} - 1\right)\cdot\prod_{h\in\mathcal{S}}\langle v^{(h)}(Z), v^{(h)}(X)\rangle,\right], \tag{G.5}$$

where $Z=(z_{-M},\ldots,z_{-1})$ is independent of $X=(x_{-M},\ldots,x_{-1})$ and we define $v^{(h)}(X) := \sum_{i_h\in[M]}\sigma_{-i_h}^{(h)} x_{-i_h}$, $v^{(h)}(Z) := \sum_{i_h\in[M]}\sigma_{-i_h}^{(h)} z_{-i_h}$, and $\bar{y} := \sum_{l=1}^{L} x_l/L$. Recall that the modified chi-squared mutual information is

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

In the following, we draw a connection between $g_{0,\mathcal{S}}$ and the modified chi-squared mutual information. Specifically, we demonstrate that $\max\{|g_{0,\mathcal{S}}-g_{1,\mathcal{S}}|, |g_{1,\mathcal{S}}-g_{2,\mathcal{S}}|, |g_{2,\mathcal{S}}-g_{3,\mathcal{S}}|\} \le O(1/\sqrt{L(1-\lambda)\mu_{\min}^\pi})$, provided that a and ε are sufficiently small. This holds under Assumption B.1, alongside the property that the Markov chain sequence over a window mixes as L increases.

Closeness between $g_{0,\mathcal{S}}$ and $g_{1,\mathcal{S}}$. Let us first consider the approximation of $g_{0,\mathcal{S}}$ by $g_{1,\mathcal{S}}$. If we select a to be sufficiently small, the attention scores of the second layer approach uniformity, meaning $\sigma_l(a\cdot s^\top)\approx 1/L$. Hence, it follows from Lemma H.2 that

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

Closeness between $g_{1,\mathcal{S}}$ and $g_{2,\mathcal{S}}$. For the approximation of $g_{1,\mathcal{S}}$ by $g_{2,\mathcal{S}}$, we leverage the approximation $\bar{y}(k)\approx\mu^\pi(e_k)$ for large L. The result in Lemma H.3 implies that

$$\begin{aligned}|g_{1,\mathcal{S}} - g_{2,\mathcal{S}}| &\le 2\cdot\sqrt{\mathbb{E}_X\left[D_{\chi^2}(\pi(\cdot\,|\,X_{\mathtt{pa}(L+1)})\,\|\,\mu^\pi(\cdot))+1\right]\cdot\left(\frac{D_{\chi^2}(\mu_0(\cdot)\,\|\,\mu^\pi(\cdot))+1}{L(1-\lambda)\cdot\mu_{\min}^\pi} + \frac{r_n}{L\mu_{\min}^\pi}\right)} \\ &\quad + \frac{r_n}{L\mu_{\min}^\pi} + \frac{\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}}{L(1-\lambda)\mu_{\min}^\pi} + \frac{\varepsilon}{\mu_{\min}^\pi}, \\ &\le O\left(\frac{1+\varepsilon}{\sqrt{L(1-\lambda)\mu_{\min}^\pi}} + \frac{\varepsilon}{\mu_{\min}^\pi}\right).\end{aligned}$$

Closeness between $g_{2,\mathcal{S}}$ and $g_{3,\mathcal{S}}$. Finally, we approximate $g_{2,\mathcal{S}}$ by $g_{3,\mathcal{S}}$ owing to the mixing property of the Markov chain. Lemma H.4 states that

$$|g_{2,\mathcal{S}} - g_{3,\mathcal{S}}| \le \left(\frac{4(M \vee r_n)}{L} + \frac{4\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda)} \right) \le O\left(\frac{1}{L(1-\lambda)} \right).$$

Combining the above results, and letting $a = a(0) \le O(1/L^{3/2})$, $\varepsilon = 1/\sqrt{L}$, we obtain

$$|g_{0,\mathcal{S}} - g_{3,\mathcal{S}}| \le O\left(\frac{1}{\sqrt{L(1-\lambda)\mu^\pi_{\min}}} \right).$$

Then, the dyanmics of $c_{\mathcal{S}}^2$ in (G.2) can be approximated as follows

$$\partial_t \log c_{\mathcal{S}}^2 = \frac{4a}{C_D} \cdot \mathbb{E}_{\pi\sim\mathcal{P}} \left(g_{3,\mathcal{S}} - f \pm O\left(\frac{1}{\sqrt{L(1-\lambda)\mu^\pi_{\min}}} \right) \right). \tag{G.6}$$

Connection between $\mathbb{E}_{\pi\sim\mathcal{P}}[g_{3,\mathcal{S}}]$ and $\widetilde{I}_{\chi^2}(\mathcal{S})$. For the next step, we establish the connection between $\mathbb{E}_{\pi\sim\mathcal{P}}[g_{3,\mathcal{S}}]$ and the modified chi-squared mutual information $\widetilde{I}_{\chi^2}(\mathcal{S})$. It follows from Lemma H.5 that $\mathbb{E}_{\pi\sim\mathcal{P}}[g_{3,\mathcal{S}}]$ can be approximated as follows:

$$\left| \mathbb{E}_{\pi\sim\mathcal{P}}\left[g_{3,\mathcal{S}}\right] - \prod_{h\in\mathcal{S}} (\sigma_{-h}^{(h)})^2 \cdot I_{\chi^2}(\mathcal{S}) \right| \le \left(1 - \prod_{h\in\mathcal{S}} (\sigma_{-h}^{(h)})^2 \right) I_{\chi^2}(\mathcal{S}^\star). \tag{G.7}$$

G.1.4. Lower Bound for The Difference $\partial_t \log c_{\mathcal{S}^\star}^2 - \partial_t \log c_{\mathcal{S}}^2$

Then, by (G.6) and (G.7), the difference between the dynamics of $c_{\mathcal{S}}^2$ and $c_{\mathcal{S}^\star}^2$ can be lower bounded by

$$\begin{aligned}
&\partial_t \log c_{\mathcal{S}^\star}^2 - \partial_t \log c_{\mathcal{S}}^2 \\
&= \mathbb{E}_{\pi\sim\mathcal{P}} \left[\frac{4a}{C_D} \cdot (g_{3,\mathcal{S}^\star} - g_{3,\mathcal{S}}) \pm O\left(\frac{a}{\sqrt{L(1-\lambda)\mu^\pi_{\min}}} \right) \right] \\
&\ge \frac{4a}{C_D} \left(\prod_{h\in\mathcal{S}^\star} (\sigma_{-h}^{(h)})^2 \cdot \widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \prod_{h\in\mathcal{S}} (\sigma_{-h}^{(h)})^2 \cdot \widetilde{I}_{\chi^2}(\mathcal{S}) - \left(2 - \prod_{h\in\mathcal{S}^\star} (\sigma_{-h}^{(h)})^2 - \prod_{h\in\mathcal{S}} (\sigma_{-h}^{(h)})^2 \right) \widetilde{I}_{\chi^2}(\mathcal{S}^\star) \right) \\
&\quad - O\left(\frac{a}{\sqrt{L(1-\lambda)\gamma}} \right),
\end{aligned}$$

where the inequality follows from the assumption that $\pi(\cdot \,|\, X_{\mathtt{pa}}) > \gamma$ uniformly. This can be further lower bounded as follows:

$$\begin{aligned}
&\partial_t \log c_{\mathcal{S}^\star}^2 - \partial_t \log c_{\mathcal{S}}^2 \\
&\ge \frac{4a}{C_D} \left(\left(2 \prod_{h\in\mathcal{S}^\star} (\sigma_{-h}^{(h)})^2 + \prod_{h\in\mathcal{S}} (\sigma_{-h}^{(h)})^2 \right) \widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \prod_{h\in\mathcal{S}} (\sigma_{-h}^{(h)})^2 \widetilde{I}_{\chi^2}(\mathcal{S}) - 2\widetilde{I}_{\chi^2}(\mathcal{S}^\star) \right) - \mathrm{err}. \\
&\ge \frac{4a}{C_D} \left(2 \prod_{h\in\mathcal{S}^\star} (\sigma_{-h}^{(h)})^2 \cdot \widetilde{I}_{\chi^2}(\mathcal{S}^\star) + \prod_{h\in\mathcal{S}} (\sigma_{-h}^{(h)})^2 \cdot \Delta\widetilde{I}_{\chi^2} - 2\widetilde{I}_{\chi^2}(\mathcal{S}^\star) \right) - \mathrm{err}. \\
&\ge \frac{4a}{C_D} \left(2 \prod_{h\in[H]} (\sigma_{-h}^{(h)})^2 \cdot \widetilde{I}_{\chi^2}(\mathcal{S}^\star) + \prod_{h\in[H]} (\sigma_{-h}^{(h)})^2 \cdot \Delta\widetilde{I}_{\chi^2} - 2\widetilde{I}_{\chi^2}(\mathcal{S}^\star) \right) - \mathrm{err},
\end{aligned} \tag{G.8}$$

where we define $\Delta\widetilde{I}_{\chi^2} = \min_{S\in[H]_{\le D}\setminus\{\mathcal{S}^\star\}} \widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \widetilde{I}_{\chi^2}(\mathcal{S})$ and $\mathrm{err} := O\left(a/\sqrt{L(1-\lambda)\gamma}\right)$. Here, the second inequality follows form the definition of $\Delta\widetilde{I}_{\chi^2}$, and the last inequality follows by replacing $\prod_{h\in\mathcal{S}}(\sigma_{-h}^{(h)})^2$ with $\prod_{h\in[H]}(\sigma_{-h}^{(h)})^2$.

G.1.5. Exponential Growth of $c_{\mathcal{S}^\star}^2$

In the following, we show that the first term in (G.8) dominates the err term and leads to the exponential growth of $c_{\mathcal{S}^\star}^2$.

Note that by Assumption B.1, it holds that $w_{-h}^{(h)} \geq w_{-j}^{(h)} + \Delta w$ for all $j \neq h$, and $h \in [H]$, where

$$\Delta w \geq \log(M-1) - \log\left(\left(1 + \Delta\widetilde{I}_{\chi^2}/(14\widetilde{I}_{\chi^2}(\mathcal{S}^\star))\right)^{\frac{1}{2H}} - 1\right),$$

Since we have dominant $w_{-h}^{(h)} \gg w_{-j}^{(h)}$ at initialization, $\prod_{h\in[H]}(\sigma_{-h}^{(h)})^2$ is sufficiently large. More precisely, we can check that $\prod_{h\in[H]}(\sigma_{-h}^{(h)})^2 \geq (2\widetilde{I}_{\chi^2}(\mathcal{S}^\star) + \frac{2}{3}\Delta\widetilde{I}_{\chi^2})/(2\widetilde{I}_{\chi^2}(\mathcal{S}^\star) + \Delta\widetilde{I}_{\chi^2})$, which yields

$$2\prod_{h\in[H]}(\sigma_{-h}^{(h)})^2 \cdot \widetilde{I}_{\chi^2}(\mathcal{S}^\star) + \prod_{h\in[H]}(\sigma_{-h}^{(h)})^2 \cdot \Delta\widetilde{I}_{\chi^2} - 2\widetilde{I}_{\chi^2}(\mathcal{S}^\star) \geq \frac{2}{3}\Delta\widetilde{I}_{\chi^2}. \tag{G.9}$$

By (G.8), and (G.9), we conclude that

$$\partial_t \log c_{\mathcal{S}^\star}^2 - \partial_t \log c_{\mathcal{S}}^2 \geq \frac{a\Delta\widetilde{I}_{\chi^2}}{2C_D}.$$

It implies that $c_{\mathcal{S}^\star}^2$ grows exponentially fast, dominating all the other $c_{\mathcal{S}}^2$ at the end of the first stage. Consequently, $p_{\mathcal{S}^\star}$ converges to 1.

G.1.6. Convergence of $p_{\mathcal{S}^\star}^2$

Now, let us derive the convergence rate of $p_{\mathcal{S}^\star}$. Since for all $\mathcal{S} \neq \mathcal{S}^\star$, $\partial_t\left(c_{\mathcal{S}}^2/c_{\mathcal{S}^\star}^2\right) < 0$, it holds that $\partial_t\left(\log C_D/c_{\mathcal{S}^\star}^2\right) = \partial_t\left(\log\sum_{\mathcal{S}\in[H]_{\leq D}} c_{\mathcal{S}}^2/c_{\mathcal{S}^\star}^2\right) < 0$. Together with Lemma G.1, we have $\partial_t \log c_{\mathcal{S}^\star}^2 > 0$. Furthermore,

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

Then, we can derive the convergence rate of $p_{\mathcal{S}^\star}$ as follows:

$$\partial_t\log(1 - p_{\mathcal{S}^\star}) = \partial_t\log\left(\frac{\sum_{\mathcal{S}\in[H]_{\leq D}\backslash\mathcal{S}^\star} c_{\mathcal{S}}^2}{C_D}\right) \leq -\frac{a\Delta\widetilde{I}_{\chi^2}}{2C_D}.$$

Applying the Grönwall's inequality, we have

$$1 - p_{\mathcal{S}^\star}(t) \leq (1 - p_{\mathcal{S}^\star}(0))\cdot\exp\left(-\frac{a\Delta\widetilde{I}_{\chi^2}}{2C_D}\cdot t\right). \tag{G.10}$$

G.2. Analysis for Stage II

In this section, we provide the analysis of the dynamics of $\sigma^{(h)} \equiv \sigma(w^{(h)})$ for $h \in \mathcal{S}^\star$. For $h \notin \mathcal{S}^\star$, it follows from the results in Stage I that the dynamics of $w_{-h}^{(h)}$ exponentially decay to zero. 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 $t \to \infty$, along with the corresponding convergence rate.

Proof Strategy. Similar to the proof of Stage I, our analysis characterizes the difference dynamics, $\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)}$ 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 fixed index i and h.

2. **Dynamics Approximation** Subsequently, we approximate the dynamics in terms of the modified chi-squared mutual information $\widetilde{I}_{\chi^2}(\mathcal{S}^\star)$, considering sufficiently small a, ε, and large L.

3. **Lower Bound for The Growth Rate** By comparing the corresponding modified chi-squared mutual information, we establish a lower bound for the difference dynamics.

4. **Convergence.** Finally, we derive the convergence rate of $\sigma_{-h}^{(h)}$ to 1 as $t \to \infty$ from the obtained lower bound.

Now we are ready to provide the proof of Stage II.

G.2.1. Calculation of The Dynamics of $\partial_t w^{(h)}$

For convenience, we recall the following notations:

$$s_l = \sum_{\mathcal{S}\in[H]_{\le D}} p_{\mathcal{S}} \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle,\ p_{\mathcal{S}} = \frac{c_{\mathcal{S}}^2}{\sum_{\mathcal{S}\in[H]_{\le D}} c_{\mathcal{S}}^2},\ v_l^{(h)} = \sum_{i\in[M]} \sigma_{-i}^{(h)} x_{l-i} = \sigma^{(h)} X_{l-M:l-1},$$

where $X_{l-M:l-1} \in \mathbb{R}^{M\times d}$ is the submatrix of X with rows $l-M,\dots,l-1$ and $\sigma^{(h)} = (\sigma_{-M}^{(h)},\dots,\sigma_{-1}^{(h)}) \in \mathbb{R}^{1\times M}$. The gradients are given by

$$\frac{\partial v_l^{(h)}}{\partial \sigma^{(h)}} = X_{l-M:l-1}^\top, \qquad \frac{\partial \sigma^{(h)}}{\partial w^{(h)}} = \operatorname{diag}(\sigma^{(h)}) - (\sigma^{(h)})^\top \sigma^{(h)},$$

$$\frac{\partial v_l^{(h)}}{\partial w^{(h)}} = X_{l-M:l-1}^\top \left(\operatorname{diag}(\sigma^{(h)}) - (\sigma^{(h)})^\top \sigma^{(h)}\right), \qquad \frac{\partial s_l}{\partial v_l^{(h)}} = \sum_{\substack{\mathcal{S}\in[H]_{\le D}\\ \text{s.t } h\in\mathcal{S}}} p_{\mathcal{S}} \cdot \prod_{h'\in\mathcal{S}\setminus\{h\}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle v_{L+1}^{(h)},$$

$$\frac{\partial s_l}{\partial v_{L+1}^{(h)}} = \sum_{\substack{\mathcal{S}\in[H]_{\le D}\\ \text{s.t } h\in\mathcal{S}}} p_{\mathcal{S}} \cdot \prod_{h'\in\mathcal{S}\setminus\{h\}} \langle v_l^{(h')}, v_{L+1}^{(h')}\rangle v_l^{(h)}, \qquad \frac{\partial \ell}{\partial s_l} = a\left(\frac{x_{L+1}}{y+\varepsilon\mathbf{1}}\right)^\top (x_l - y)\cdot \sigma_l\left(a\cdot s\right).$$

To simplify the notation, we define $b_l := X_{(l-M):(l-1)}(v_{L+1}^{(h)}) + X_{(L+1-M):L}(v_{l+1}^{(h)}) \in \mathbb{R}^M$. By the chain rule, we have

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

where we note that $e_i \in \mathbb{R}^M$ is the i-th standard basis vector. To proceed, we define the quantity $g_{h,0}$ as follows:

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

By the gradients and the definition of $g_{h,0}$, the gradient flow dynamics of $w_{-i}^{(h)}$ is given by

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

The difference of the dynamics of $w_{-h}^{(h)}$ and $w_{-i}^{(h)}$ can be written as

$$\begin{aligned}&\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} \\ &= a \cdot \sum_{l=1}^{L} \mathbb{E}\left[\sigma_l\left(a \cdot s\right) \cdot \sum_{k \in [d]}\left(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon}\right) \cdot \left(\frac{\partial s_l}{\partial w_{-h}^{(h)}} - \frac{\partial s_l}{\partial w_{-i}^{(h)}}\right)\right]. \\ &= a \cdot \mathbb{E}_{\pi \sim \mathcal{P}}\left[g_{h,0}^{\top}\left(\sigma_{-i}^{(h)}\left(e_{M+1-h} - e_{M+1-i}\right) + \left(\sigma_{-h}^{(h)} - \sigma_{-i}^{(h)}\right)\sum_{j \neq h} \sigma_{-j}^{(h)}(e_{M+1-h} - e_{M+1-j})\right)\right]. \qquad \text{(G.11)}\end{aligned}$$

G.2.2. Approximation of $g_{h,0}$

To further analyze the dynamics of $w_{-h}^{(h)} - w_{-i}^{(h)}$, we define the following quantities that are used for approximating $g_{h,0}$.

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

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

$$v^{(h)}(X) := \sum_{i_h \in [M]} \sigma_{-i_h}^{(h)} x_{-i_h}, \quad v^{(h)}(Z) := \sum_{i_h \in [M]} \sigma_{-i_h}^{(h)} z_{-i_h},$$

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

To simplify the notation, we treat X and Z as matrices, where each row in Z and X reflects a vector sampled from μ^π, indicating the state of the Markov chain at different steps within the window.

One can observe from (G.11) that the lower bounded of $g_{h,0}^{\top}(e_{M+1-h} - e_{M+1-i})$ for all $i \neq h$ is required to show $\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} > 0$. To achieve this, we first approximate $g_{h,0}^{\top}(e_{M+1-h} - e_{M+1-i})$ by $g_{h,4}^{\top}(e_{M+1-h} - e_{M+1-i})$, similar to our approach in Stage I.

Closeness between $g_{h,0}$ and $g_{h,1}$. Due to the rapid exponential dominance of $p_{\mathcal{S}^\star}$ from Stage I, the coefficients $p_{\mathcal{S}}$ for $\mathcal{S} \neq \mathcal{S}^\star$ in $g_{h,0}$ are negligible. Moreover, note that $b_l^\top (e_{M+1-h} - e_{M+1-i}) = \langle v_{L+1}^{(h)}, x_{l-h} - x_{l-i}\rangle - \langle v_l^{(h)}, x_{L+1-h} - x_{L+1-i}\rangle$. By similar argument in (G.17), we have

$$\left|(g_{h,0} - g_{h,1})^\top (e_{M+1-h} - e_{M+1-i})\right| \lesssim (1 - p_{\mathcal{S}^\star}) =: \Delta_1$$

for all $i \neq h$. Given $(1 - p_{\mathcal{S}^\star}(t)) \leq \exp\left(-at\Delta\widetilde{I}_{\chi^2}/(2C_D)\right)$ from (G.10), we consider $t \gtrsim \log(L\log L)/(a\Delta\widetilde{I}_{\chi^2})$ to ensure that $\Delta_1 \leq O\left(1/(L\log L)\right)$.

Closeness between $g_{h,1}$ and $g_{h,2}$. Next, since a is chosen to be a sufficiently small value, we have $\sigma_l(a \cdot s^\top) \approx 1/L$, and $g_{h,1}^\top (e_{M+1-h} - e_{M+1-i})$ can be approximated by $g_{h,2}^\top (e_{M+1-h} - e_{M+1-i})$. By Lemma H.2, it holds that

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

Closeness between $g_{h,2}$ and $g_{h,3}$. In addition, as $\bar{y}(k) \approx \mu^\pi(e_k)$, for large L, we can approximate $g_{h,2}^\top (e_{M+1-h} - e_{M+1-i})$ by $g_{h,3}^\top (e_{M+1-h} - e_{M+1-i})$. More precisely, it follows from Lemma H.3 that

$$\begin{aligned}
&\left|(g_{h,2} - g_{h,3})^\top (e_{M+1-h} - e_{M+1-i})\right| \\
&\lesssim \sqrt{\mathbb{E}_X\left[D_{\chi^2}(\pi(\cdot \,|\, X_{\mathrm{pa}(L+1)}) \,\|\, \mu^\pi(\cdot)) + 1\right] \cdot \left(\frac{D_{\chi^2}(\mu_0(\cdot) \,\|\, \mu^\pi(\cdot)) + 1}{L(1-\lambda)\cdot \mu^\pi_{\min}} + \frac{r_n}{L\mu^\pi_{\min}}\right)} \\
&\quad + \frac{r_n}{L\mu^\pi_{\min}} + \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda)\mu^\pi_{\min}} + \frac{\varepsilon}{\mu^\pi_{\min}}.
\end{aligned}$$

Closeness between $g_{h,3}$ and $g_{h,4}$. Finally, the mixing of the Markov chain implies the approximation of $g_{h,3}^\top (e_{M+1-h} - e_{M+1-i})$ by $g_{h,4}^\top (e_{M+1-h} - e_{M+1-i})$. This is described in Lemma H.4, which states

$$\left|(g_{h,3} - g_{h,4})^\top (e_{M+1-h} - e_{M+1-i})\right| \lesssim \frac{(M \vee r_n)}{L} + \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda)}.$$

Combining the above results and setting $\varepsilon = 1/\sqrt{L}$, and $a = a(0) \leq O(1/L^{3/2})$, we obtain

$$\left|(g_{h,0} - g_{h,4})^\top (e_{M+1-h} - e_{M+1-i})\right| \leq O\left(\frac{1}{\sqrt{L(1-\lambda)\mu^\pi_{\min}}}\right).$$

G.2.3. Lower Bound for The Difference $\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)}$

Then, we can rewrite (G.11) as

$$\begin{aligned}
&\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} \\
&= a \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\left[g_{h,0}^\top \left(\sigma_{-i}^{(h)} (e_{M+1-h} - e_{M+1-i}) + \left(\sigma_{-h}^{(h)} - \sigma_{-i}^{(h)}\right)\sum_{j\neq h} \sigma_{-j}^{(h)}(e_{M+1-h} - e_{M+1-j})\right)\right] \\
&\geq a \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\left[g_{h,4}^\top \left(\sigma_{-i}^{(h)} (e_{M+1-h} - e_{M+1-i}) + \left(\sigma_{-h}^{(h)} - \sigma_{-i}^{(h)}\right)\sum_{j\neq h} \sigma_{-j}^{(h)}(e_{M+1-h} - e_{M+1-j})\right)\right] \\
&\quad - O\left(\frac{a}{\sqrt{L(1-\lambda)\gamma}}\right).
\end{aligned} \tag{G.12}$$

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

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

where we define

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

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

Hence, it suffices to analyze the difference between $\tau_{h,1}$ and $\tau_{h,2}$. Drawing on similar reasoning as in the proof of Lemma H.5, we can approximate $\tau_{h,1}$ and $\tau_{h,2}$ as follows:

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

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

where

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

To establish the lower bound for $\tau_{h,1} - \tau_{h,2}$, let's begin by finding an upper bound for ψ, which is approximately equal to $\tau_{h,2}$. We consider the two cases: $i \in \mathcal{S}^\star$ and $i \notin \mathcal{S}^\star$. If $i \notin \mathcal{S}^\star$, we invoke Lemma H.6 with $\mathcal{S} = \mathcal{S}^\star$ and $\mathcal{S}' = \mathcal{S}^\star\setminus\{h\}\cup\{i\}$. Then,

$$
\psi = \frac{1}{2}\widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \frac{1}{2}\widetilde{I}_{\chi^2}((\mathcal{S}^\star\setminus\{h\})\cup\{i\}) \le \widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \frac{1}{2}\cdot\Delta\widetilde{I}_{\chi^2}.
$$

On the other hand, if $i \in \mathcal{S}^\star$, we apply Lemma H.7 with $\mathcal{S} = \mathcal{S}^\star\setminus\{h\}$ and $\mathcal{S}' = \mathcal{S}^\star$ and obtain

$$
\psi < \frac{1}{2}\widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \frac{1}{2}\widetilde{I}_{\chi^2}((\mathcal{S}^\star\setminus\{h\})) \le \widetilde{I}_{\chi^2}(\mathcal{S}^\star) - \frac{1}{2}\cdot\Delta\widetilde{I}_{\chi^2}.
$$

In both cases, we have the same upper bound for ψ. Thus,

$$
\begin{aligned}
2\tau_{h,1} - 2\tau_{h,2} &\ge \prod_{h'\in\mathcal{S}^\star\setminus\{h\}}(\sigma_{-h'}^{(h')})^2\cdot\sigma_{-h}^{(h)}\cdot\Delta\widetilde{I}_{\chi^2} - 4\left(1 - \prod_{h'\in\mathcal{S}^\star\setminus\{h\}}(\sigma_{-h'}^{(h')})^2\cdot\sigma_{-h}^{(h)}\right)\widetilde{I}_{\chi^2}(\mathcal{S}^\star) \\
&\ge \prod_{h\in[H]}(\sigma_{-h}^{(h)})^2\cdot\Delta\widetilde{I}_{\chi^2} - 4\left(1 - \prod_{h\in[H]}(\sigma_{-h}^{(h)})^2\right)\widetilde{I}_{\chi^2}(\mathcal{S}^\star). \qquad \text{(G.13)}
\end{aligned}
$$

Note that it follows from Assumption B.1, that

$$\prod_{h\in[H]} (\sigma_{-h}^{(h)})^2 \geq \frac{4\widetilde{I}_{\chi^2}(\mathcal{S}^\star) + \frac{2}{3}\Delta\widetilde{I}_{\chi^2}}{4\widetilde{I}_{\chi^2}(\mathcal{S}^\star) + \Delta\widetilde{I}_{\chi^2}}. \tag{G.14}$$

Consequently, by (G.13) and (G.14), it holds that

$$2\tau_{h,1} - 2\tau_{h,2} \geq \frac{2}{3}\Delta\widetilde{I}_{\chi^2}. \tag{G.15}$$

Together with (G.12) and (G.15) implies that $\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} > 0$ for all $i \neq h$ and $w_{-h}^{(h)}$ will grow faster than $w_{-i}^{(h)}$ for all $i \neq h$ if $h \in \mathcal{S}^\star$.

G.2.4. Convergence of $\sigma^{(h)}$

Next, we characterize the convergence rate of $\sigma^{(h)}$. Since $\partial_t \sigma_{-h}^{(h)} > 0$ for all $h \in \mathcal{S}^\star$, the lower bound for $\sigma_{-i}^{(h)}$ is given by

$$\sigma_{-i}^{(h)} \geq \sigma_{-h}^{(h)} \cdot \exp(-(w_{-h}^{(h)} - w_{-i}^{(h)})) \geq \sigma_{-h}^{(h)}(0) \cdot \exp(-(w_{-h}^{(h)} - w_{-i}^{(h)})). \tag{G.16}$$

Then, by (G.12), (G.15) and (G.16), the following lower bound is obtained.

$$\begin{aligned}
&\partial_t w_{-h}^{(h)} - \partial_t w_{-i}^{(h)} \\
&\geq a \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\left[g_{h,4}^\top \left(\sigma_{-i}^{(h)}\left(e_{M+1-h} - e_{M+1-i}\right) + \left(\sigma_{-h}^{(h)} - \sigma_{-i}^{(h)}\right)\sum_{j\neq h}\sigma_{-j}^{(h)}(e_{M+1-h} - e_{M+1-j})\right)\right] \\
&\qquad\qquad - O\left(\frac{a}{\sqrt{L(1-\lambda)\gamma}}\right) \\
&\geq a \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\left[\sigma_{-i}^{(h)} g_{h,4}^\top\left(e_{M+1-h} - e_{M+1-i}\right)\right] - O\left(\frac{a}{\sqrt{L(1-\lambda)\gamma}}\right) \\
&\geq \frac{a\Delta\widetilde{I}_{\chi^2}}{2} \cdot \sigma_{-h}^{(h)}(0) \cdot \exp(-(w_{-h}^{(h)} - w_{-i}^{(h)})).
\end{aligned}$$

Rearranging the terms, the dynamics of $w_{-h}^{(h)} - w_{-i}^{(h)}$ can be characterized as follows:

$$\begin{aligned}
\partial_t \exp(w_{-h}^{(h)} - w_{-i}^{(h)}) &\geq \frac{a\Delta\widetilde{I}_{\chi^2} \cdot \sigma_{-h}^{(h)}(0)}{2} > 0, \\
\exp(w_{-h}^{(h)}(t) - w_{-i}^{(h)}(t)) &\geq \frac{a\Delta\widetilde{I}_{\chi^2} \cdot \sigma_{-h}^{(h)}(0)}{2} \cdot t + \exp(\Delta w),
\end{aligned}$$

where we use the assumption that $w_{-h}^{(h)}(0) - w_{-i}^{(h)}(0) \geq \Delta w$. As a result, during the Stage II, $\sigma^{(h)}$ becomes a hot one vector e_{M+1-h} and the following upper bound goes to zero as t goes to infinity.

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

where we define $\sigma_{\min}(0) := \min_{h\in\mathcal{S}^\star}\sigma_{-h}^{(h)}(0)$ use the inequality $(1-x)^n \geq 1 - nx$ for $x \in [0, 1/n]$ and $n \geq 1$.

G.3. Proof of Stage III

Additional Notation. For a set $\mathcal{S} \subseteq [M]$, we let $X_{l-\mathcal{S}}$ denote the set of tokens $\{X_{l-s} | s \in \mathcal{S}\}$. If $l = 0$, we will ignore l in the subscript and simply use $X_{-\mathcal{S}}$.

In this section, we derive the dynamics of the second 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.

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

1. **Dynamics Calculation.** First, we derive 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 the dynamics of a when a is either sufficiently small or large.

G.3.1. Calculation of The Dynamics of a

Let us consider the time-derivative of a at Stage III. By taking the gradient through the softmax operation, we have

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

Therefore,

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

Here, $y = \sum_{l=1}^{L} \sigma_l x_l$ is the predicted output, which is a vector function of (x_{L+1}, X). Also, we abbreviate $\sigma_l(a \cdot s^\top)$ as σ_l and denote by σ the vector $(\sigma_1, \ldots, \sigma_L)^\top$. We denote the above quantity by f_0.

G.3.2. Approximation of $\partial_t a$

Approximation of f_0 by f_1. Our first step is to remove the summation over $[H]_{\le D} \backslash \{\mathcal{S}^\star\}$ where $\mathcal{S}^\star$ is the optimal set that maximizes the modified mutual information defined in (**??**) and $c_{\mathcal{S}^\star}$ dominates according to the training of Stage I. To this end, we define f_1 as

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

It follows that

$$|f_0 - f_1| \le 2(1 - p_{\mathcal{S}^\star}) =: 2\Delta_1.$$

Here, the inequality holds by noting that for any $\mathcal{S} \in [H]_{\le D}$,

$$\begin{aligned}
&\left| \sum_{l=1}^{L} \sigma_l \cdot \sum_{k\in[d]} \left(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} - \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon} \right) \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle \right| \\
&\quad \le \left| \sum_{k\in[d]} \sum_{l=1}^{L} \sigma_l \cdot \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y(k)+\varepsilon} \right| + \left| \sum_{k\in[d]} \sum_{l=1}^{L} \sigma_l \cdot \frac{y(k)\,\mathbb{1}(x_{L+1} = e_k)}{y(k)+\varepsilon} \right| \\
&\quad \le \left| \sum_{k\in[d]} \sum_{l=1}^{L} \sigma_l \cdot \frac{\mathbb{1}(x_l = e_k)}{y(k)+\varepsilon} \right| + \left| \sum_{k\in[d]} \sum_{l=1}^{L} \sigma_l \cdot \frac{y(k)}{y(k)+\varepsilon} \right| \le 2 \left| \sum_{k\in[d]} \frac{y(k)}{y(k)+\varepsilon} \right| \le 2.
\end{aligned} \tag{G.17}$$

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

Approximation of f_1 by f_2. Next, we use the results from Stage II to control the difference between $\prod_{h\in\mathcal{S}^\star} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle$ and $\prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})$ as

$$\left| \prod_{h\in\mathcal{S}^\star} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle - \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h}) \right| \le 1 - \prod_{h\in\mathcal{S}^\star} (\sigma_{-h}^{(h)})^2 := \Delta_2.$$

Note that these two error terms also influence our definition of f_1 through σ_l as the second layer's softmax score is given by

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

Let us define $s_l^\star = \mathbb{1}_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})$. Then, we have

$$|s_l - s_l^\star| \le \Delta_1 + \Delta_2, \quad \forall l \in [L].$$

To proceed, we define $\sigma_l^\star$ as

$$\sigma_l^\star = \frac{\exp\left(a \cdot \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})\right)}{\sum_{l'=1}^{L} \exp\left(a \cdot \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l'-h} = x_{L+1-h})\right)},$$

and define $y^\star(k) = \sum_{l=1}^{L} \sigma_l^\star\, \mathbb{1}(x_l = e_k)$. As a result, by Lemma 5.1 of Chen et al. (2022),

$$\begin{aligned}
\left\| \log \frac{\sigma_l^\star}{\sigma_l} \right\|_\infty &\le 2a \cdot \|s - s^\star\|_\infty \le 2a \cdot (\Delta_1 + \Delta_2), \\
\|\sigma - \sigma^\star\|_1 &\le 4a \cdot \|s - s^\star\|_\infty \le 4a \cdot (\Delta_1 + \Delta_2), \\
\|y^\star - y\|_1 &\le \|\sigma - \sigma^\star\|_1 \le 4a \cdot (\Delta_1 + \Delta_2).
\end{aligned}$$

To this end, we also define

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

The approximation error is then given by

$$|f_1 - f_2| \le \mathrm{err}_1 + \mathrm{err}_2 + \mathrm{err}_3,$$

where the three error terms are give respectively by

$$\mathrm{err}_1 := \left| \mathbb{E}\left[\sum_{l=1}^{L} \sigma_l \cdot \sum_{k\in[d]} \left(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y^\star(k)+\varepsilon} - \frac{y^\star(k)\,\mathbb{1}(x_{L+1}=e_k)}{y^\star(k)+\varepsilon} \right) \cdot \left(\prod_{h\in\mathcal{S}^\star} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle - \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h}) \right) \right] \right|,$$

$$\mathrm{err}_2 := \left| \mathbb{E}\left[\sum_{l=1}^{L} (\sigma_l^\star - \sigma_l) \cdot \sum_{k\in[d]} \left(\frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y^\star(k)+\varepsilon} - \frac{y^\star(k)\,\mathbb{1}(x_{L+1}=e_k)}{y^\star(k)+\varepsilon} \right) \cdot \prod_{h\in\mathcal{S}^\star} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle \right] \right|,$$

$$\mathrm{err}_3 := \left| \mathbb{E}\left[\sum_{l=1}^{L} \sigma_l \cdot \sum_{k\in[d]} \left(\frac{1}{y^\star(k)+\varepsilon} - \frac{1}{y(k)+\varepsilon} \right) \cdot \mathbb{1}(x_{L+1} = x_l = e_k) \cdot \prod_{h\in\mathcal{S}^\star} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle \right] \right| + \left| \mathbb{E}\left[\sum_{l=1}^{L} \sigma_l \cdot \sum_{k\in[d]} \left(\frac{y^\star(k)}{y^\star(k)+\varepsilon} - \frac{y(k)}{y(k)+\varepsilon} \right) \cdot \mathbb{1}(x_{L+1} = e_k) \cdot \prod_{h\in\mathcal{S}^\star} \langle v_l^{(h)}, v_{L+1}^{(h)} \rangle \right] \right|.$$

It then holds that

$$\begin{aligned} \mathrm{err}_1 + \mathrm{err}_2 + \mathrm{err}_3 &\le \varepsilon^{-1} \left(\Delta_2 + 4a(\Delta_1 + \Delta_2) \right) + \sum_{k\in[d]} \frac{|y^\star(k) - y(k)| y(k)}{(y^\star(k)+\varepsilon)(y(k)+\varepsilon)} \cdot (1+\varepsilon) \\ &\le \varepsilon^{-1} \left(\Delta_2 + 4a(\Delta_1+\Delta_2) + 4a \cdot (\Delta_1+\Delta_2) \cdot (1+\varepsilon) \right) \\ &= O(\varepsilon^{-1}(1+a)(\Delta_1+\Delta_2)). \end{aligned}$$

In summary, this error terms captures the difference between the ideal weights and the actual converging weights from Stage II.

Approximation of f_2 by f_3. Next, we approximate f_2 by f_3 where we replace $y^\star = \sum_{l=1}^{L} \sigma_l^\star x_l$ by its population counterpart

$$\widetilde{\mu}_X^\pi(z, Z) = \frac{\mu^\pi(z,Z) \exp\left(a \cdot \prod_{h\in\mathcal{S}^\star} \mathbb{1}(z_{-h} = x_{L+1-h}) \right)}{\sum_{z,Z} \mu^\pi(z,Z) \exp\left(a \cdot \prod_{h\in\mathcal{S}^\star} \mathbb{1}(z_{-h} = x_{L+1-h}) \right)},$$

where $Z = (z_{-M}, \ldots, z_{-1})$ and $\mu^\pi(z, Z)$ is the joint distribution of a length-$(M+1)$ window of the Markov chain. 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. We define f_3 as

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

One can immediately draw a connection to Lemma H.3 as both targets characterize the gap between the empirical and population distributions. The only difference is that this time we have the distribution reweighted by some exponential term. For completeness, we provide the following lemma.

Lemma G.2. *The difference between f_2 and f_3 is bounded by*

$$|f_2 - f_3| \lesssim \frac{\gamma^{-1}}{\min_{z, Z_{-\mathcal{S}^\star}} \mu^\pi(z, Z_{-\mathcal{S}^\star})} \cdot \sqrt{\frac{D_{\chi^2}(\mu_0(\cdot) \,\|\, \mu^\pi(\cdot)) + 1}{L(1-\lambda)} + \frac{3M}{L}} + \frac{d\varepsilon}{\gamma},$$

where $\lesssim$ hides some universal constant.

Proof of Lemma G.2. The proof follows the same arguments as Lemma H.3. We use $y_X^\star(k)$ in place of $y_X(k)$ to remind the reader that $y^\star(k)$ is also a function of the whole chain. We note that

$$
\begin{aligned}
|f_3 - f_2| &= \bigg|\mathbb{E}\bigg[\sum_{l=1}^{L} \sigma_l^\star \bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{y_X^\star(k) + \varepsilon} - \sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\widetilde{\mu}_X^\pi(e_k)} \\
&\qquad - \sum_{k\in[d]} \frac{y_X^\star(k)\,\mathbb{1}(x_{L+1} = e_k)}{y_X^\star(k) + \varepsilon} + 1\bigg) \cdot \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})\bigg]\bigg| \\
&= \bigg|\mathbb{E}\bigg[\sum_{l=1}^{L} \sigma_l^\star \bigg(\sum_{k\in[d]} \Big(\frac{\widetilde{\mu}_X^\pi(e_k) - y_X^\star(k)}{(y_X^\star(k)+\varepsilon)\widetilde{\mu}_X^\pi(e_k)} - \frac{\varepsilon}{(y_X^\star(k)+\varepsilon)\widetilde{\mu}_X^\pi(e_k)}\Big) \cdot \mathbb{1}(x_{L+1} = x_l = e_k) \\
&\qquad - \sum_{k\in[d]} \frac{\varepsilon\,\mathbb{1}(x_{L+1} = e_k)}{y_X^\star(k) + \varepsilon}\bigg) \cdot \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})\bigg]\bigg|.
\end{aligned}
$$

We also define three error terms as

$$
\begin{aligned}
\mathrm{err}_1 &:= \left|\mathbb{E}\left[\sum_{k\in[d]} \frac{\widetilde{\mu}_X^\pi(e_k) - y_X^\star(k)}{(y_X^\star(k)+\varepsilon)\widetilde{\mu}_X^\pi(e_k)} \cdot \sum_{l=1}^{L} \sigma_l^\star\,\mathbb{1}(x_{L+1} = x_l = e_k) \cdot \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})\right]\right|, \\
\mathrm{err}_2 &:= \left|\mathbb{E}\left[\sum_{k\in[d]} \frac{\varepsilon}{(y_X^\star(k)+\varepsilon)\widetilde{\mu}_X^\pi(e_k)} \cdot \sum_{l=1}^{L} \sigma_l^\star\,\mathbb{1}(x_{L+1} = x_l = e_k) \cdot \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})\right]\right|, \\
\mathrm{err}_3 &:= \left|\mathbb{E}\left[\sum_{k\in[d]} \frac{\varepsilon}{y_X^\star(k)+\varepsilon} \cdot \mathbb{1}(x_{L+1} = e_k) \cdot \sum_{l=1}^{L} \sigma_l^\star \prod_{h\in\mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})\right]\right|.
\end{aligned}
$$

For the first error term, we have have that

$$
\begin{aligned}
\mathrm{err}_1 &\le \mathbb{E}\left[\sum_{k\in[d]} \frac{|\widetilde{\mu}_X^\pi(e_k) - y_X^\star(k)|}{(y_X^\star(k)+\varepsilon)} \cdot \sum_{l=1}^{L} \frac{\sigma_l^\star\,\mathbb{1}(x_l = e_k)}{\widetilde{\mu}_X^\pi(e_k)}\right] \\
&= \mathbb{E}\left[\sum_{k\in[d]} \frac{|\widetilde{\mu}_X^\pi(e_k) - y_X^\star(k)|}{(y_X^\star(k)+\varepsilon)} \cdot \frac{y_X^\star(k)}{\widetilde{\mu}_X^\pi(e_k)}\right] \\
&\le \gamma^{-1} \cdot \mathbb{E}\left[\sum_{k\in[d]} |\widetilde{\mu}_X^\pi(e_k) - y_X^\star(k)|\right],
\end{aligned}
$$

where we recall that by assumption, γ provides a lower bound for $\pi(\cdot \,|\, X_{\mathtt{pa}})$, hence also lower bound for $\widetilde{\mu}_X^\pi(e_k)$. The following proposition provides a bound for the 1-norm of the difference between the empirical and population distributions.

Proposition G.3. *It holds that*

$$
\mathbb{E}_X\left[\|\widetilde{\mu}_X^\pi(e_k) - y_X^\star(k)\|_1\right] \lesssim \frac{2}{\min_{z, Z_{-\mathcal{S}^\star}} \mu^\pi(z, Z_{-\mathcal{S}^\star})} \cdot \sqrt{\frac{D_{\chi^2}(\mu_0(\cdot)\,\|\,\mu^\pi(\cdot)) + 1}{L(1-\lambda)} + \frac{3M}{L}}.
$$

Hence, we control the first error term.

For the second error term, we follow the same procedure and obtain an upper bound as

$$
\mathrm{err}_2 \le \mathbb{E}\left[\sum_{k\in[d]} \frac{\varepsilon}{\widetilde{\mu}_X^\pi(e_k)} \cdot \sum_{l=1}^{L} \frac{\sigma_l^\star\,\mathbb{1}(x_l = e_k)}{(y_X^\star(k)+\varepsilon)}\right] \le \gamma^{-1} d\varepsilon.
$$

For the last error term, it holds that

$$
\begin{aligned}
\mathrm{err}_3 &\le \mathbb{E}\left[\sum_{k\in[d]} \frac{\varepsilon}{y_X^\star(k)+\varepsilon}\cdot \mathbb{1}(x_{L+1}=e_k)\right] \\
&\le \left|\mathbb{E}\left[\sum_{k\in[d]} \frac{\varepsilon\,\mathbb{1}(x_{L+1}=e_k)}{\widetilde{\mu}_X^\pi(e_k)+\varepsilon}\right]\right| + \left|\sum_{k\in[d]} \mathbb{E}\left[\frac{\varepsilon(y_X^\star(k)-\widetilde{\mu}_X^\pi(e_k))\cdot \mathbb{1}(x_{L+1}=e_k)}{(\widetilde{\mu}_X^\pi(e_k)+\varepsilon)(y_X^\star(k)+\varepsilon)}\right]\right| \\
&\le \frac{\varepsilon}{\gamma+\varepsilon} + \mathbb{E}\left[\sum_{k\in[d]} \frac{|y_X^\star(k)-\widetilde{\mu}_X^\pi(e_k)|}{\gamma+\varepsilon}\right].
\end{aligned}
$$

We further have the last term controlled by the upper bound in Proposition G.3.

In summary, the difference between f_2 and f_3 is bounded by

$$
\begin{aligned}
|f_2-f_3| &\le \mathrm{err}_1+\mathrm{err}_2+\mathrm{err}_3 \\
&\lesssim \frac{\gamma^{-1}}{\min_{z,Z_{-\mathcal{S}^\star}} \mu^\pi(z,Z_{-\mathcal{S}^\star})}\cdot\sqrt{\frac{D_{\chi^2}(\mu_0(\cdot)\,\|\,\mu^\pi(\cdot))+1}{L(1-\lambda)}+\frac{3M}{L}}+\frac{d\varepsilon}{\gamma},
\end{aligned}
$$

which completes our proof. □

Approximation of f_3 by f_4. Note that in the expression for f_3, we still have $\sigma_l^\star$ that implicitly depends on the whole sequence. We define f_4 by replacing $\sigma_l^\star$ by its population counterpart $\sigma^\star(X_{l-\mathcal{S}^\star})$ which is defined as

$$
\sigma^\star(X_{l-\mathcal{S}^\star}) := \frac{\mu^\pi(X_{l-\mathcal{S}^\star})\exp(a\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(X_{l-h}=X_{L+1-h}))}{\sum_{X'_{l-\mathcal{S}^\star}}\mu^\pi(X'_{l-\mathcal{S}^\star})\exp(a\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(X'_{l-h}=X_{L+1-h}))}.
$$

And we define f_4 as

$$
f_4 := \mathbb{E}_{(z,Z)\sim\widetilde{\mu}_X^\pi,X}\left[\sum_{k\in[d]}\left(\frac{\mathbb{1}(x_{L+1}=z=e_k)}{\widetilde{\mu}_X^\pi(e_k)} - \mathbb{1}(x_{L+1}=e_k)\right)\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(z_{l-h}=x_{L+1-h})\right].
$$

We only need to characterize the difference between $\sigma_l^\star$ and $\sigma^\star(X_{l-\mathcal{S}^\star})$. We have the following proposition.

Lemma G.4. *The difference between f_3 and f_4 is bounded by*

$$
|f_3-f_4| \lesssim \frac{2\gamma^{-1}}{\min_{z,Z_{-\mathcal{S}^\star}}\mu^\pi(z,Z_{-\mathcal{S}^\star})}\cdot\sqrt{\frac{D_{\chi^2}(\mu_0(\cdot)\,\|\,\mu^\pi(\cdot))+1}{L(1-\lambda)}+\frac{3M}{L}}.
$$

Proof of Lemma G.4. We follow the same notation as in the proof of Proposition G.3 and let

$$
R(Z_{-\mathcal{S}^\star}, X_{L+1-\mathcal{S}^\star}) = \exp\left(a\cdot\prod_{h\in\mathcal{S}^\star}\mathbb{1}(Z_{-h}=X_{L+1-h})\right)
$$

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 note that

$$
\begin{aligned}
\sum_{l=1}^{L}\sigma_l^\star\,\mathbb{1}(x_l=z, X_{l-\mathcal{S}^\star}=Z_{-\mathcal{S}^\star}) &= \frac{\widehat{\mu}_X^\pi(z,Z_{-\mathcal{S}^\star})R(Z_{-\mathcal{S}^\star},X_{-\mathcal{S}^\star})}{\widehat{\Phi}}, \\
\widetilde{\mu}_X^\pi(z,Z_{-\mathcal{S}^\star}) &= \frac{\mu^\pi(z,Z_{-\mathcal{S}^\star})R(Z_{-\mathcal{S}^\star},X_{-\mathcal{S}^\star})}{\Phi}.
\end{aligned}
$$

We further define

$$
\phi(z,Z_{-\mathcal{S}^\star}) = \mu^\pi(z,Z_{-\mathcal{S}^\star})R(Z_{-\mathcal{S}^\star},X_{-\mathcal{S}^\star}),\quad \widehat{\phi}(z,Z_{-\mathcal{S}^\star}) = \widehat{\mu}_X^\pi(z,Z_{-\mathcal{S}^\star})R(Z_{-\mathcal{S}^\star},X_{-\mathcal{S}^\star}).
$$

Therefore, the difference of f_3 and f_4 is given by

$$|f_3 - f_4| \le \gamma^{-1} \cdot \mathbb{E}_X \left[\sum_{z, Z_{-\mathcal{S}^\star}} \left| \frac{\phi(z, Z_{-\mathcal{S}^\star})}{\Phi} - \frac{\widehat{\phi}(z, Z_{-\mathcal{S}^\star})}{\widehat{\Phi}} \right| \right].$$

Following the same procedure of (H.14) and (H.15) in the proof of Proposition G.3, we have

$$\begin{aligned} \sum_{z, Z_{-\mathcal{S}^\star}} \left| \frac{\phi(z, Z_{-\mathcal{S}^\star})}{\Phi} - \frac{\widehat{\phi}(z, Z_{-\mathcal{S}^\star})}{\widehat{\Phi}} \right| &\le 2 \cdot \sum_{z, Z_{-\mathcal{S}^\star}} |\mu^\pi(z, Z_{-\mathcal{S}^\star}) - \widehat{\mu}^\pi_X(z, Z_{-\mathcal{S}^\star})| \\ &\quad + 2 \cdot \frac{\sum_z |\mu^\pi(z, X_{-\mathcal{S}^\star}) - \widehat{\mu}^\pi_X(z, X_{-\mathcal{S}^\star})|}{\mu^\pi(Z_{-\mathcal{S}^\star})}. \end{aligned} \tag{G.18}$$

The second term of the right-hand side of (G.18) has an upper bound

$$\frac{2}{\min_E \mu^\pi(Z_{-\mathcal{S}^\star} = E)} \cdot \sqrt{\frac{D_{\chi^2}(\mu_0(\cdot) \,\|\, \mu^\pi(\cdot)) + 1}{L(1-\lambda)} + \frac{3M}{L}}$$

as we have established in (H.16), (H.17), and (H.18). For the first term, we have by the Cauchy-Schwarz inequality that

$$\begin{aligned} &\mathbb{E}_X \left[\sum_z |\mu^\pi(z, Z_{-\mathcal{S}^\star}) - \widehat{\mu}^\pi_X(z, Z_{-\mathcal{S}^\star})| \right] \\ &\le \sqrt{\mathbb{E}_X \left[\sum_{z, Z_{-\mathcal{S}^\star}} \frac{(\mu^\pi(z, Z_{-\mathcal{S}^\star}) - \widehat{\mu}^\pi_X(z, Z_{-\mathcal{S}^\star}))^2}{\mu^\pi(z, Z_{-\mathcal{S}^\star})} \right]} \\ &\le \frac{1}{\sqrt{\min_{e,E} \mu^\pi(z = e, Z_{-\mathcal{S}^\star} = E)}} \cdot \sqrt{\frac{D_{\chi^2}(\mu_0(\cdot) \,\|\, \mu^\pi(\cdot)) + 1}{L(1-\lambda)} + \frac{3M}{L}}, \end{aligned}$$

where Lemma H.10 is used in the last inequality. □

Approximation of f_4 by f_5. Now that we have z, Z distributed according $\widetilde{\mu}^\pi_X$, which depends only on $X_{L+1-\mathcal{S}^\star}$. In the sequel, we abbreviate $(x_{L+1}, X_{L+1-\mathcal{S}^\star})$ as $(x, X_{-\mathcal{S}^\star})$ where $X_{-\mathcal{S}^\star} = (x_{-h})_{h \in \mathcal{S}^\star}$. The joint distribution for $(x, X_{-\mathcal{S}^\star}, z, Z_{-\mathcal{S}^\star})$ is given by

$$\widetilde{p}^\pi(x, X_{-\mathcal{S}^\star}, z, Z_{-\mathcal{S}^\star}) = p^\pi_{L+1}(x, X_{-\mathcal{S}^\star}) \cdot \widetilde{\mu}^\pi(z, Z_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star}),$$

where we use $\widetilde{\mu}^\pi(z, Z_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star})$ to replace $\widetilde{\mu}^\pi_X(z, Z_{-\mathcal{S}^\star})$ for a clearer notation of the dependency. Here, p^π_{L+1} is the distribution for $(x_{L+1}, X_{L+1-\mathcal{S}^\star})$ and $\widetilde{\mu}^\pi(\cdot)$ is defined as

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

For our convenience, we define

$$q^\pi = \mu^\pi(x, X_{-\mathcal{S}^\star}) \cdot \widetilde{\mu}^\pi(z, Z_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star}),$$

and let

$$f_5 = \mathbb{E}_{(x, X_{-\mathcal{S}^\star}, z, Z_{-\mathcal{S}^\star}) \sim q^\pi} \left[\sum_{k \in [d]} \left(\frac{\mathbb{1}(x = z = e_k)}{\widetilde{\mu}^\pi(e_k \,|\, X_{-\mathcal{S}^\star})} - \mathbb{1}(x = e_k) \right) \cdot \prod_{h \in \mathcal{S}^\star} \mathbb{1}(z_{-h} = x_{-h}) \right].$$

One can rewrite f_5 as

$$f_5 = \mathbb{E}_{(x, X_{-\mathcal{S}^\star}) \sim \mu^\pi} \Bigg[\sum_{k \in [d]} \frac{\mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star}) \widetilde{\mu}^\pi(z = e_k, Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star})}{\widetilde{\mu}^\pi(z = e_k \,|\, X_{-\mathcal{S}^\star})} - \widetilde{\mu}^\pi(Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star} \,|\, X_{-\mathcal{S}^\star}) \Bigg].$$

And f_4 is given by replacing the distribution of $(x, X_{-\mathcal{S}^\star})$ by p_{L+1}^π in f_5. The difference between f_4 and f_5 is thus bounded by the results in (H.11) of Lemma H.8:

$$|f_4 - f_5| \le \big\| \mu^\pi(x, X_{-\mathcal{S}^\star}) - p_{L+1}^\pi(x, X_{-\mathcal{S}^\star}) \big\|_1 \le \lambda^{L-M} \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}.$$

Collecting all the approximation results, we have

$$|f_0 - f_5| \lesssim \Delta_1 + \varepsilon^{-1}(1+a)(\Delta_1 + \Delta_2) + \lambda^{L-M}\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1} + \frac{\gamma^{-1}}{\min_{z, Z_{-\mathcal{S}^\star}} \mu^\pi(z, Z_{-\mathcal{S}^\star})} \cdot \sqrt{\frac{D_{\chi^2}(\mu_0(\cdot) \,\|\, \mu^\pi(\cdot)) + 1}{L(1-\lambda)} + \frac{3M}{L}} + \frac{d\varepsilon}{\gamma}.$$

Here, we split the error into two parts where the first part is constant error and the second part is the error that also depends on a:

$$\xi = \Delta_1 + \lambda^{L-M}\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1} + \frac{\gamma^{-1}}{\min_{z, Z_{-\mathcal{S}^\star}} \mu^\pi(z, Z_{-\mathcal{S}^\star})} \cdot \sqrt{\frac{D_{\chi^2}(\mu_0(\cdot) \,\|\, \mu^\pi(\cdot)) + 1}{L(1-\lambda)} + \frac{3M}{L}} + \frac{d\varepsilon}{\gamma}$$

$$\psi(a) = \varepsilon^{-1}(1+a)(\Delta_1 + \Delta_2).$$

G.3.3. Lower and Upper Bound for The Dynamics of a

Now, we can safely work with f_5. By definition, we have

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

where we note that $\widetilde{\mu}^\pi(z = e_k \,|\, Z_{-\mathcal{S}^\star} = X_{-\mathcal{S}^\star}, X_{-\mathcal{S}^\star}) = \mu^\pi(x = e_k \,|\, X_{-\mathcal{S}^\star})$ as fixing $Z_{-\mathcal{S}^\star}$ makes z independent of $X_{-\mathcal{S}^\star}$. We can rewrite $\widetilde{\mu}^\pi(z \,|\, X_{-\mathcal{S}^\star})$ as

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

For our convenience, we let $r(X_{-\mathcal{S}^\star}) = (1+\mu^\pi(X_{-\mathcal{S}^\star})\cdot(e^a-1))^{-1}$. We then have

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

Consequently, we have for f_5 that

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

We see that f_5 is bounded below as

$$f_5 \ge \sum_{X_{-\mathcal{S}^\star}}\left(\min_{\alpha\in[\rho_-(a),\rho_+(a)]}\sum_{k\in[d]}\frac{(\mu^\pi(x=e_k\,|\,X_{-\mathcal{S}^\star})-\mu^\pi(e_k))^2}{(1-\alpha)\mu^\pi(x=e_k\,|\,X_{-\mathcal{S}^\star})+\alpha\mu^\pi(e_k)}\right)\cdot e^a r(X_{-\mathcal{S}^\star})^3\cdot\mu^\pi(X_{-\mathcal{S}^\star})^2,$$

where

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

which are given by the upper and lower bound of $r(X_{-\mathcal{S}^\star})$, respectively. Let us define

$$\widetilde{D}_{\chi^2,\rho(a)}(P\,\|\,Q) = \min_{\alpha\in[0,\rho(a)]}\sum_{x\in\mathcal{X}}\frac{(P(x)-Q(x))^2}{(1-\alpha)Q(x)+\alpha P(x)}.$$

In the sequel, as we study the dynamics of a, we will denote f_5 as $f_5(a)$. Then, the lower bound for f_5 can be also written as

$$f_5(a) \ge \sum_{X_{-\mathcal{S}^\star}}\underbrace{\widetilde{D}_{\chi^2,\rho(a)}\left(\mu^\pi(\cdot)\,\|\,\mu^\pi(\cdot\,|\,X_{-\mathcal{S}^\star})\right)}_{J_-(X_{-\mathcal{S}^\star};a)}\cdot\frac{e^a\mu^\pi(X_{-\mathcal{S}^\star})}{(1+\mu^\pi(X_{-\mathcal{S}^\star})\cdot(e^a-1))^3}\cdot\mu^\pi(X_{-\mathcal{S}^\star}).$$

Also, since

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

is a convex function of α (by noting that the second derivative is non-negative), we have

$$\begin{aligned}
f_5(a) &\le \big(D_{\chi^2}(\mu^\pi(\cdot)\,\|\,\mu^\pi(\cdot\,|\,X_{-\mathcal{S}^\star}))\cdot(1-r(X_{-\mathcal{S}^\star})) + D_{\chi^2}(\mu^\pi(\cdot\,|\,X_{-\mathcal{S}^\star})\,\|\,\mu^\pi(\cdot))\cdot r(X_{-\mathcal{S}^\star})\big)\\
&\qquad\cdot e^a r(X_{-\mathcal{S}^\star})^3\cdot\mu^\pi(X_{-\mathcal{S}^\star})^2\\
&\le \sum_{X_{-\mathcal{S}^\star}}\underbrace{\max_{\alpha\in[\rho_-(a),\rho_+(a)]}\big((1-\alpha)\cdot D_{\chi^2}(\mu^\pi(\cdot)\,\|\,\mu^\pi(\cdot\,|\,X_{-\mathcal{S}^\star}))+\alpha\cdot D_{\chi^2}(\mu^\pi(\cdot\,|\,X_{-\mathcal{S}^\star})\,\|\,\mu^\pi(\cdot))\big)}_{J_+(X_{-\mathcal{S}^\star};a)}\\
&\qquad\cdot\frac{e^a\mu^\pi(X_{-\mathcal{S}^\star})}{(1+\mu^\pi(X_{-\mathcal{S}^\star})\cdot(e^a-1))^3}\cdot\mu^\pi(X_{-\mathcal{S}^\star}).
\end{aligned}$$

Note that both $J_+(X_{-\mathcal{S}^\star}; a)$ and $J_-(X_{-\mathcal{S}^\star}; a)$ are of constant scale, i.e., uniformly upper and lower bounded regardless of a. Also, the time derivative of a is given by

$$\partial_t a = \mathbb{E}_{\pi\sim\mathcal{P}}[f_5] \pm (\xi + \psi(a)).$$

G.3.4. Convergence of a

Here, we abuse the notation and denote by $\xi = \mathbb{E}_{\pi\sim\mathcal{P}}[\xi]$ and $\psi(a) = \mathbb{E}_{\pi\sim\mathcal{P}}[\psi(a)]$. Thus, a continues to increase until it reaches a point where f_5 no longer dominates the error. We denote by $a^\star$ the threshold where $f_5(a^\star) = \xi + \psi(a^\star)$. Note that $a^\star$ can be as large as $\log L$ since we could make $\psi(a)$ arbitrarily small by letting the first and second stages to be sufficiently long and $\xi = O(L^{-1/2})$ will be the elbow. In the following, we only characterize the dynamics of a for $a \le a^\star$. We also use $x = o(1)$ to denote that a term is much smaller than 1, e.g., $x = (\log\log L)^{-1}$. We use $x = x_0 \pm \delta$ to represent the fact that x is bounded around x_0 by δ error.

Small a. Consider the case where a is small in the sense that $\mu^\pi(X_{-\mathcal{S}^\star})e^a \le \delta, \forall X_{-\mathcal{S}^\star}, \forall \pi \in \mathrm{supp}(\mathcal{P})$ for some small constant δ. Then, we have for the gradient that

$$\partial_t a = (1 \pm O(\delta)) \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\left[\sum_{X_{-\mathcal{S}^\star}} J(X_{-\mathcal{S}^\star}; a) \cdot e^a \mu^\pi(X_{-\mathcal{S}^\star})^2 \pm (\xi + \psi(a))\right].$$

Here, we recall that

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

with lower bound $J_-(X_{-\mathcal{S}^\star}; a)$ and upper bound $J_+(X_{-\mathcal{S}^\star}; a)$. We notice that $\rho_-(a) \ge 1-\delta$. Thus, both $J_-(X_{-\mathcal{S}^\star}; a)$ and $J_+(X_{-\mathcal{S}^\star}; a)$ are controlled within $(1 \pm O(\delta)) D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot))$. Here, we use the condition that

$$\begin{aligned}
\xi + \psi(\log L) &= O(L^{-1/2} \cdot \gamma^{-|\mathcal{S}^\star|-2} \cdot (1-\lambda)^{-1/2}) \\
&\le \delta \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\left[\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot)) \cdot \mu^\pi(X_{-\mathcal{S}^\star})^2\right],
\end{aligned}$$

which gives us

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

With the result, we have

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

which implies that for small a, the growth follows

$$\begin{aligned}
a(t) &\le -\log\left(e^{-a(0)} - (1 + O(\delta)) \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\left[\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot))\mu^\pi(X_{-\mathcal{S}^\star})^2\right] \cdot t\right), \\
a(t) &\ge -\log\left(e^{-a(0)} - (1 - O(\delta)) \cdot \mathbb{E}_{\pi\sim\mathcal{P}}\left[\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}(\mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star}) \,\|\, \mu^\pi(\cdot))\mu^\pi(X_{-\mathcal{S}^\star})^2\right] \cdot t\right).
\end{aligned}$$

Therefore, at the beginning, a grows super exponentially fast.

Large a. As a grows large such that $\mu^\pi(X_{-\mathcal{S}^\star})e^a \geq \delta^{-1}, \forall X_{-\mathcal{S}^\star}, \forall \pi \in \mathrm{supp}(\mathcal{P})$ with δ being the same as in the previous case, we have for the gradient that

$$\partial_t a = (1 \pm O(\delta)) \cdot \mathbb{E}_{\pi \sim \mathcal{P}} \left[\sum_{X_{-\mathcal{S}^\star}} J(X_{-\mathcal{S}^\star}; a) \cdot \frac{e^{-2a}}{\mu^\pi(X_{-\mathcal{S}^\star})} \right] \pm (\xi + \psi(a)).$$

Notice that $\rho_+(a) = (1 + \min_{X_{-\mathcal{S}^\star}} \mu^\pi(X_{-\mathcal{S}^\star})(e^a - 1))^{-1} \leq \delta$ this time, which implies that

$$J(X_{-\mathcal{S}^\star}; a) = (1 \pm O(\delta)) \cdot D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})).$$

To ensure that the signal in the gradient dominates the error, we require

$$\sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})) \cdot \frac{e^{-2a}}{\mu^\pi(X_{-\mathcal{S}^\star})} = \omega(\xi + \psi(a)).$$

A sufficient condition for this to be true is $a \leq (1 - \delta)\log L/4$ with

$$\delta \cdot \sum_{X_{-\mathcal{S}^\star}} D_{\chi^2}(\mu^\pi(\cdot) \,\|\, \mu^\pi(\cdot \,|\, X_{-\mathcal{S}^\star})) \cdot L^{\delta/2} \geq O(\gamma^{-2} \cdot (1 - \lambda)^{-1/2})$$

given the fact that $\xi = O(L^{-1/2})$ and $\psi(a) < O(L^{-1/2})$ by letting the first two stages run long enough such that $\Delta_1 + \Delta_2 \leq O(L^{-1/2}/\log L)$. Thus,

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

which gives us

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

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

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

until it reaches the value $(1 - \delta) \log L/4$.

G.4. 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 G.5. *Consider $H = M$ and Assumption B.3 holds. Suppose the error $\Delta_1, \Delta_2 \lesssim L^{-1/2}$ after the first two stages' training, and $a \geq (1 - \delta) \log L/4$ for some small constant $\delta < 1$ 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 $\mathtt{GIH}(x_{1:L}; M, D)$. Then with high probability $1 - O(L^{-1})$, it holds that*

$$\|y^\star - y\|_1 \leq O(L^{-(1-\delta)/4}).$$

Proof of Lemma G.5. Let $s_l^\star = \prod_{h \in \mathcal{S}^\star} \mathbb{1}(x_{l-h} = x_{L+1-h})$ and $s_l = \langle u_{L+1}, u_l \rangle$. Let us invoke Lemma H.1 to obtain the model misspecification error as

$$\max_{l \in [L]} |s_l^\star - s_l| \leq 2(\Delta_1 + \Delta_2) := \Delta.$$

We note that the second layer's attention weight a can be as large as $(1-\delta)\log L/4$. 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

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

We define

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

with $\sigma^\star = (\sigma_l^\star)_{l>M}$. Therefore, 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 $\Gamma^c = \{L \ge l > M : \prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h} = x_{L+1-h}) = 0\}$. We then have

$$\|\sigma^\star - \sigma\|_1 \le \sum_{l\in\Gamma}|\sigma_l^\star - \sigma_l| + \sum_{l\in\Gamma^c}\sigma_l$$

For $l\in\Gamma$, we have $1\ge s_l \ge 1-\Delta$ and for $l\in\Gamma^c$, we have $0\le s_l\le\Delta$. Consider the normalization factor in the softmax operator.

$$\mathcal{Z} := \sum_{l>M}\exp(a\cdot s_l).$$

The normalization factor is lower and upper bounded by

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

We then have for $l\in\Gamma$ that

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

The right hand side is upper bounded by $O(a\Delta/N) + O(L\exp(-a)/N^2)$. For $l\in\Gamma^c$, we have

$$\sigma_l \le \frac{\exp(a\Delta)}{\mathcal{Z}_-} \le \frac{\exp(a\cdot(2\Delta-1))}{N}.$$

In summary,

$$\|y^\star - y\|_1 \le \|\sigma^\star - \sigma\|_1 \le O(a\Delta) + O\left(L\exp(-a)/N\right). \tag{G.19}$$

The above inequality holds whenever $N\ge 1$, where we use the condition that $a\Delta \le \log L\cdot\Delta \ll 1$. By Lemma H.10, we have the second moment

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

Therefore, by the Chebyshev'sinequality, 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 $\mathcal{X}^{|\mathcal{S}^\star|}$, we conclude that with high probability (say 0.99) it holds that $N\geq tL=L\cdot\min_{E\in\mathcal{X}^{|\mathcal{S}^\star|}}\mu^\pi(E)/2$. Thus, it follows from (G.19) that with high probability

$$\|y^\star-y\|_1\lesssim a\Delta+\exp(-a)\lesssim L^{-1/2}\log L+L^{-(1-\delta)/4}.$$

where in the last inequality we use $a\geq(1-\delta)\log L/4$. □

H. Auxiliary Lemmas

H.1. Useful Inequalities

Lemma H.1 (Model Misspecification). *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 $2(\Delta_1+\Delta_2)$, where Δ_1 and Δ_2 are the errors after the first and second stage's training, 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. Let us consider the output feature u_l after the FFN & Normalization layer, where the inner product is given by

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

Since each $v_l^{(h)}$ is a convex combination of $x_{\mathcal{M}(l)}$ where $\mathcal{M}(l)=\{l-M,\ldots,l-1\}$, we have $v_l^{(h)}$ having norm at most 1. Thus,

$$\begin{aligned}\left|\langle u_{L+1},u_l\rangle-\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)},v_{L+1}^{(h)}\rangle\right|&\leq(1-p_{\mathcal{S}^\star})+\sum_{\mathcal{S}\in[H]_{\leq D}\backslash\{\mathcal{S}^\star\}}p_{\mathcal{S}}\cdot\prod_{h\in\mathcal{S}}\langle v_l^{(h)},v_{L+1}^{(h)}\rangle\\&\leq2(1-p_{\mathcal{S}^\star})=:2\Delta_1,\end{aligned}$$

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

$$\langle v_l^{(h)},v_{L+1}^{(h)}\rangle=\sum_{i,j\in[M]^2}\sigma_{-i}^{(h)}\sigma_{-j}^{(h)}\langle x_{l-i},x_{L+1-j}\rangle=\sum_{i,j\in[M]^2}\sigma_{-i}^{(h)}\sigma_{-j}^{(h)}\,\mathbb{1}(x_{l-i}=x_{L+1-j}).$$

Hence, we have that

$$\begin{aligned}&\left|\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)},v_{L+1}^{(h)}\rangle-\prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)})^2\,\mathbb{1}(x_{l-h}=x_{L+1-h})\right|\\&\quad=\left|\sum_{\{i_h,j_h\}_{h\in\mathcal{S}^\star}\neq\{h,h\}_{h\in\mathcal{S}^\star}}\prod_{h\in\mathcal{S}^\star}\sigma_{-i_h}^{(h)}\sigma_{-j_h}^{(h)}\,\mathbb{1}(x_{l-i_h}=x_{L+1-j_h})\right|\\&\quad\leq\sum_{\{i_h,j_h\}_{h\in\mathcal{S}^\star}\neq\{h,h\}_{h\in\mathcal{S}^\star}}\prod_{h\in\mathcal{S}^\star}\sigma_{-i_h}^{(h)}\sigma_{-j_h}^{(h)}\leq1-\prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)})^2=:\Delta_2,\end{aligned}$$

where Δ_2 is the error after the second stage's training. As a result,

$$\left|\prod_{h\in\mathcal{S}^\star}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle - \prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\right| \le 2\left(1-\prod_{h\in\mathcal{S}^\star}(\sigma_{-h}^{(h)})^2\right) = 2\Delta_2.$$

In summary, we have that

$$\left|\langle u_{L+1}, u_l\rangle - \prod_{h\in\mathcal{S}^\star}\mathbb{1}(x_{l-h}=x_{L+1-h})\right| \le 2(\Delta_1+\Delta_2).$$

Hence, the model misspecification error is bounded by $2(\Delta_1+\Delta_2)$. We finish the proof. □

Lemma H.2. *Consider $g_{0,\mathcal{S}}$ in (G.1) with $a=a_0=a(0)$ and $g_{1,\mathcal{S}}$ in (G.3), which is equivalent to $g_{0,\mathcal{S}}$ when $a=0$. Then, for $a_0\le 1$, it holds that*

$$|g_{0,\mathcal{S}}-g_{1,\mathcal{S}}| \le \frac{8a_0 d}{\varepsilon^2}.$$

Proof of Lemma H.2. By triangular inequality, we have

$$\begin{aligned}|g_{0,\mathcal{S}}-g_{1,\mathcal{S}}| \le \sum_{l=1}^{L}\mathbb{E}\Bigg[\sum_{k\in[d]}\Bigg\{&\left|\sigma\left(a_0\cdot s^\top\right)_l-\frac{1}{L}\right|\left|\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y(k)+\varepsilon}\right| \\ &+\frac{1}{L}\left|\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y(k)+\varepsilon}-\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\bar{y}(k)+\varepsilon}\right| + \left|\sigma\left(a_0\cdot s^\top\right)_l-\frac{1}{L}\right|\left|\frac{y(k)\,\mathbb{1}(x_{L+1}=e_k)}{y(k)+\varepsilon}\right|, \\ &+\frac{1}{L}\left|\frac{y(k)\,\mathbb{1}(x_{L+1}=e_k)}{y(k)+\varepsilon}-\frac{\bar{y}(k)\,\mathbb{1}(x_{L+1}=e_k)}{\bar{y}(k)+\varepsilon}\right|\Bigg\}\cdot\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Bigg].\end{aligned}$$

Note that $0\le s_l\le 1$ for all $l\in[L]$ thanks to the layer normalization. Then, for the softmax operation, we have

$$\frac{1}{1+(L-1)\exp(a_0)} \le \sigma\left(a_0\cdot s^\top\right)_l \le \frac{\exp(a_0)}{L-1+\exp(a_0)},$$

which implies that

$$\left|\sigma\left(a_0\cdot s^\top\right)_l-\frac{1}{L}\right| \le \max\left\{\frac{1}{L}-\frac{1}{1+(L-1)\exp(a_0)}, \frac{\exp(a_0)}{L-1+\exp(a_0)}-\frac{1}{L}\right\} \le \frac{\exp(a_0)-1}{L}. \tag{H.1}$$

Since indicator functions are bounded above by 1, we have

$$\left|\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y(k)+\varepsilon}\right| \le \frac{1}{\varepsilon}, \quad \left|\frac{y(k)\,\mathbb{1}(x_{L+1}=e_k)}{y(k)+\varepsilon}\right| \le \frac{1}{\varepsilon}, \tag{H.2}$$

For the second term, we have

$$\begin{aligned}\left|\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{y(k)+\varepsilon}-\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\bar{y}(k)+\varepsilon}\right| &\le \frac{|\bar{y}(k)-y(k)|}{\varepsilon^2} \le \frac{\sum_{l=1}^{L}|\sigma\left(a_0\cdot s^\top\right)_l-\frac{1}{L}|}{\varepsilon^2} \\ &\le \frac{\exp(a_0)-1}{\varepsilon^2},\end{aligned} \tag{H.3}$$

where the last inequality follows from (H.1). Similary, the following bound can be derived.

$$\left|\frac{y(k)\,\mathbb{1}(x_{L+1}=e_k)}{y(k)+\varepsilon}-\frac{\bar{y}(k)\,\mathbb{1}(x_{L+1}=e_k)}{\bar{y}(k)+\varepsilon}\right| \le \frac{\exp(a_0)-1}{\varepsilon}. \tag{H.4}$$

Combining (H.1), (H.2), (H.3) and (H.4), it holds that

$$|g_{0,\mathcal{S}} - g_{1,\mathcal{S}}| \le \sum_{l=1}^{L} \mathbb{E}\bigg[4 \sum_{k\in[d]} \frac{\exp(a_0)-1}{\varepsilon^2 L} \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\bigg] \le \frac{4d(\exp(a_0)-1)}{\varepsilon^2} \le \frac{8a_0 d}{\varepsilon^2},$$

where the last inequality follows from $\exp(x) - 1 \le 2x$ for $0 \le x \le 1$. □

Lemma H.3. *Consider $g_{1,\mathcal{S}}$ in (G.3) and $g_{3,\mathcal{S}}$ in (G.4). Then, it holds that*

$$|g_{1,\mathcal{S}} - g_{2,\mathcal{S}}| \le 2\sqrt{\mathbb{E}_X\left[D_{\chi^2}(\pi(\cdot \mid X_{\mathtt{pa}(L+1)}) \,\|\, \mu^\pi(\cdot)) + 1\right] \cdot \left(\frac{D_{\chi^2}(\mu_0(\cdot)\,\|\,\mu^\pi(\cdot)) + 1}{L(1-\lambda)\cdot \mu^\pi_{\min}} + \frac{r_n}{L\mu^\pi_{\min}}\right)} + \frac{r_n}{L\mu^\pi_{\min}} + \frac{\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}}{L(1-\lambda)\mu^\pi_{\min}} + \frac{\varepsilon}{\mu^\pi_{\min}},$$

where $\mu_0(\cdot)$ is the initial distribution over the first r_n tokens, $\mu^\pi_{\min}$ is the minimum of the one-token stationary distribution.

Proof of Lemma H.3. Let us use $\bar{y}_X(\cdot)$ to remind ourself that $\bar{y}(\cdot)$ is also a function of X. By rearranging the terms, we have

$$\begin{aligned}
|g_{1,\mathcal{S}} - g_{2,\mathcal{S}}| &= \bigg|\frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\bigg[\bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\bar{y}_X(k)+\varepsilon} - \sum_{k\in[d]}\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\mu^\pi(e_k)} \\
&\qquad - \sum_{k\in[d]}\frac{\bar{y}_X(k)\,\mathbb{1}(x_{L+1}=e_k)}{\bar{y}_X(k)+\varepsilon} + 1\bigg)\cdot \prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\bigg]\bigg| \\
&= \bigg|\frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\bigg[\bigg(\sum_{k\in[d]}\Big(\frac{\mu^\pi(e_k)-\bar{y}_X(k)}{(\bar{y}_X(k)+\varepsilon)\mu^\pi(e_k)} - \frac{\varepsilon}{(\bar{y}_X(k)+\varepsilon)\mu^\pi(e_k)}\Big)\cdot \mathbb{1}(x_{L+1}=x_l=e_k) \\
&\qquad - \sum_{k\in[d]}\frac{\varepsilon\,\mathbb{1}(x_{L+1}=e_k)}{\bar{y}_X(k)+\varepsilon}\bigg)\cdot\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\bigg]\bigg|.
\end{aligned}$$

Here, we have three terms to control. For the first error term, we define

$$\mathrm{err}_1 := \bigg|\frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\bigg[\sum_{k\in[d]}\frac{\mu^\pi(e_k)-\bar{y}_X(k)}{(\bar{y}_X(k)+\varepsilon)\mu^\pi(e_k)}\cdot \mathbb{1}(x_{L+1}=x_l=e_k)\cdot\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\bigg]\bigg|$$

Using Cauchy-Schwarz inequality, we arrive at

$$\begin{aligned}
\mathrm{err}_1^2 &\le \mathbb{E}_X\left[\sum_{k\in[d]}\left(\frac{\mu^\pi(e_k)-\bar{y}_X(k)}{\sqrt{\mu^\pi(e_k)}}\right)^2\right] \\
&\quad \cdot \mathbb{E}_X\left[\sum_{k\in[d]}\left(\frac{1}{L}\sum_{l=1}^{L}\frac{\pi(e_k\mid X_{\mathtt{pa}(L+1)})\,\mathbb{1}(x_l=e_k)\cdot\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle}{(\bar{y}_X(k)+\varepsilon)\sqrt{\mu^\pi(e_k)}}\right)^2\right] \\
&\le \mathbb{E}_X\left[\sum_{k\in[d]}\left(\frac{\mu^\pi(e_k)-\bar{y}_X(k)}{\sqrt{\mu^\pi(e_k)}}\right)^2\right]\cdot \mathbb{E}_X\left[\sum_{k\in[d]}\left(\frac{\pi(e_k\mid X_{\mathtt{pa}(L+1)})\bar{y}(k)}{(\bar{y}_X(k)+\varepsilon)\sqrt{\mu^\pi(e_k)}}\right)^2\right] \\
&\le \mathbb{E}_X D_{\chi^2}(\bar{y}_X(\cdot)\,\|\,\mu^\pi(\cdot))\cdot \mathbb{E}_X\left[D_{\chi^2}(\pi(\cdot\mid X_{\mathtt{pa}(L+1)})\,\|\,\mu^\pi(\cdot))+1\right]
\end{aligned}$$

where in the first inequality, we also invoke the exchangeability of summation over L and the expectation. The second inequality holds by noting that $\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle \le 1$. Now, the problem boils down to controlling the chi-square divergence between the empirical distribution and the stationary distribution. Lastly, we invoke Lemma H.10 which indicates that the first chi-square distance is upper bounded by

$$\frac{D_{\chi^2}(\mu_0(B=\cdot)\,\|\,\mu^\pi(B=\cdot))+1}{L(1-\lambda)\cdot\mu^\pi_{\min}}+\frac{r_n}{L\mu^\pi_{\min}}.$$

For the second term, we have

$$\begin{aligned}
\mathrm{err}_2 &= \left|\frac{1}{L}\sum_{l=1}^{L}\sum_{k\in[d]}\mathbb{E}\left[\frac{\varepsilon}{(\bar{y}_X(k)+\varepsilon)\mu^\pi(e_k)}\cdot\mathbb{1}(x_{L+1}=x_l=e_k)\cdot\prod_{h\in\mathcal{S}}\langle v_l^{(h)},v_{L+1}^{(h)}\rangle\right]\right| \\
&\le \left|\frac{1}{L}\sum_{l=1}^{L}\sum_{k\in[d]}\mathbb{E}\left[\frac{\varepsilon}{(\bar{y}_X(k)+\varepsilon)\mu^\pi(e_k)}\cdot\mathbb{1}(x_{L+1}=x_l=e_k)\right]\right| \\
&\le \underbrace{\left|\frac{1}{L}\sum_{l=1}^{L}\sum_{k\in[d]}\mathbb{E}\left[\frac{\varepsilon}{(\bar{y}_X(k)+\varepsilon)}\cdot\frac{L^{-1}\sum_{l=1}^{L}\mathbb{1}(x_{L+1}=x_l=e_k)-\mu^\pi(e_k)\bar{y}_X(k)}{\mu^\pi(e_k)}\right]\right|}_{(i)} \\
&\quad+\underbrace{\left|\frac{1}{L}\sum_{l=1}^{L}\sum_{k\in[d]}\mathbb{E}\left[\frac{\varepsilon\bar{y}_X(k)}{(\bar{y}_X(k)+\varepsilon)}\right]\right|}_{(ii)}
\end{aligned}$$

We invoke (H.12) of Lemma H.9 for the first term, which gives us

$$\begin{aligned}
(i) &\le \left|\frac{1}{L}\sum_{l=1}^{L}\sum_{k\in[d]}\mathbb{E}\left[\frac{L^{-1}\sum_{l=1}^{L}\mathbb{1}(x_{L+1}=x_l=e_k)-\mu^\pi(e_k)\bar{y}_X(k)}{\mu^\pi(e_k)}\right]\right| \\
&\le \frac{r_n}{L\mu^\pi_{\min}}+\frac{\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}}{L(1-\lambda)\mu^\pi_{\min}}.
\end{aligned}$$

And the second term (ii) is directly upper bounded by ε. Lastly, we have the error term

$$\begin{aligned}
\mathrm{err}_3 := \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\left[\sum_{k\in[d]}\frac{\varepsilon\,\mathbb{1}(x_{L+1}=e_k)}{\bar{y}_X(k)+\varepsilon}\cdot\prod_{h\in\mathcal{S}}\langle v_l^{(h)},v_{L+1}^{(h)}\rangle\right] &\le \mathbb{E}\left[\sum_{k\in[d]}\frac{\varepsilon\,\mathbb{1}(x_{L+1}=e_k)}{\bar{y}_X(k)+\varepsilon}\right] \\
\le \left|\mathbb{E}\left[\sum_{k\in[d]}\frac{\varepsilon\,\mathbb{1}(x_{L+1}=e_k)}{\mu^\pi(e_k)+\varepsilon}\right]\right| + \left|\sum_{k\in[d]}\mathbb{E}\left[\frac{\varepsilon(\bar{y}_X(k)-\mu^\pi(e_k))\cdot\mathbb{1}(x_{L+1}=e_k)}{(\mu^\pi(e_k)+\varepsilon)(\bar{y}_X(k)+\varepsilon)}\right]\right|&.
\end{aligned}$$

Here, the first term is upper bounded by $\varepsilon/\mu^\pi_{\min}$, and for the second term we have by Cauchy-Schwartz that

$$\begin{aligned}
&\left|\sum_{k\in[d]}\mathbb{E}\left[\frac{\varepsilon(\bar{y}_X(k)-\mu^\pi(e_k))\cdot\mathbb{1}(x_{L+1}=e_k)}{(\mu^\pi(e_k)+\varepsilon)(\bar{y}_X(k)+\varepsilon)}\right]\right|^2 \\
&\quad\le \varepsilon^2\cdot\mathbb{E}\left[\sum_{k\in[d]}\frac{(\bar{y}_X(k)-\mu^\pi(e_k))^2}{\mu^\pi(e_k)}\right]\cdot\mathbb{E}\left[\sum_{k\in[d]}\frac{\pi(x_{L+1}=e_k\,|\,X_{\mathtt{pa}(L+1)})^2}{(\bar{y}_X(k)+\varepsilon)^2\mu^\pi(e_k)}\right] \\
&\quad\le \mathbb{E}_X D_{\chi^2}(\bar{y}_X(\cdot)\,\|\,\mu^\pi(\cdot))\cdot\mathbb{E}_X\left[D_{\chi^2}(\pi(\cdot\,|\,X_{\mathtt{pa}(L+1)})\,\|\,\mu^\pi(\cdot))+1\right],
\end{aligned}$$

which shares a similar upper bound as err_1. Hence, we complete our proof.

□

Lemma H.4. *Consider $g_{2,\mathcal{S}}$ in (G.4) and $g_{3,\mathcal{S}}$ in (G.5). Then, it holds that*

$$|g_{2,\mathcal{S}} - g_{3,\mathcal{S}}| \le \frac{4(M \vee r_n)}{L} + \frac{4\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda)},$$

where $\mu_0(\cdot)$ is the initial distribution over the first r_n tokens.

Proof of Lemma H.4. Recall that $v^{(h)}(X) := \sum_{i_h \in [M]} \sigma_{-i_h}^{(h)} X_{-i_h}$, $v^{(h)}(Z) := \sum_{i_h \in [M]} \sigma_{-i_h}^{(h)} Z_{-i_h}$. By triangular inequality, we have

$$\begin{aligned}|g_{2,\mathcal{S}} - g_{3,\mathcal{S}}| \le \Bigg| &\frac{1}{L}\sum_{l=1}^{L} \mathbb{E}\Bigg[\Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\mu^\pi(e_k)}\Bigg) \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Bigg] \\ &- \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\Bigg[\Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(e_k)} \cdot \Bigg(\prod_{h\in\mathcal{S}} \langle v^{(h)}(Z), v^{(h)}(X)\rangle\Bigg)\Bigg)\Bigg]\Bigg| \\ &+ \Bigg|\frac{1}{L}\sum_{l=1}^{L} \mathbb{E}\Bigg[\prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Bigg] - \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\Bigg[\Bigg(\prod_{h\in\mathcal{S}} \langle v^{(h)}(Z), v^{(h)}(X)\rangle\Bigg)\Bigg]\Bigg|,\end{aligned}$$

We can establish the upper bounds for each of the absolute value terms. Initially, we focus on bounding the first absolute value term. Since $v_l^{(h)} := \sum_{i_h\in[M]} \sigma_{-i_h}^{(h)} x_{l-i_h}$, we can write

$$\prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle = \sum_{\{i_h, j_h\}_{h\in\mathcal{S}}} \prod_{h\in\mathcal{S}} \sigma_{-i_h}^{(h)} \sigma_{-j_h}^{(h)}\, \mathbb{1}(x_{l-i_h} = x_{L+1-j_h}).$$

Then,

$$\begin{aligned}&\frac{1}{L}\sum_{l=1}^{L} \mathbb{E}\Bigg[\Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\mu^\pi(e_k)}\Bigg) \cdot \prod_{h\in\mathcal{S}} \langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Bigg] \\ &= \frac{1}{L}\sum_{l=1}^{L} \sum_{\{i_h,j_h\}_{h\in\mathcal{S}}} \mathbb{E}\Bigg[\sum_{k\in[d]} \sum_{\{k_h\}_{h\in\mathcal{S}}} \frac{\mathbb{1}(x_{L+1} = x_l = e_k)}{\mu^\pi(z = e_k)} \cdot \prod_{h\in\mathcal{S}} \sigma_{-i_h}^{(h)}\sigma_{-j_h}^{(h)}\, \mathbb{1}(x_{l-i_h} = x_{L+1-j_h} = e_{k_h})\Bigg] \\ &= \sum_{\{i_h,j_h\}_{h\in\mathcal{S}}} \sum_{k\in[d]} \sum_{\{k_h\}_{h\in\mathcal{S}}} \frac{\frac{1}{L}\sum_{l=1}^{L} p^\pi(x_{L+1} = x_l = e_k, x_{l-i_h} = x_{L+1-j_h} = e_{k_h}, \forall h\in\mathcal{S})}{\mu^\pi(z = e_k)} \\ &\qquad \cdot \prod_{h\in\mathcal{S}} \sigma_{-i_h}^{(h)}\sigma_{-j_h}^{(h)}\end{aligned} \tag{H.5}$$

Similarly,

$$\begin{aligned}&\mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\Bigg[\Bigg(\sum_{k\in[d]} \frac{\mathbb{1}(x = z = e_k)}{\mu^\pi(e_k)} \cdot \Bigg(\prod_{h\in\mathcal{S}} \langle v_z^{(h)}, v_x^{(h)}\rangle\Bigg)\Bigg)\Bigg] \\ &= \sum_{\{i_h,j_h\}_{h\in\mathcal{S}}} \sum_{k\in[d]} \sum_{\{k_h\}_{h\in\mathcal{S}}} \frac{\mu^\pi(x = e_k, x_{-i_h} = e_{k_h} \forall h\in\mathcal{S}) \cdot \mu^\pi(z = e_k, z_{-j_h} = e_{k_h} \forall h\in\mathcal{S})}{\mu^\pi(z = e_k)} \\ &\qquad \cdot \prod_{h\in\mathcal{S}} \sigma_{-i_h}^{(h)}\sigma_{-j_h}^{(h)}\end{aligned} \tag{H.6}$$

By Lemma H.9, we have

$$\begin{aligned}&\Bigg|\frac{1}{L}\sum_{l=1}^{L}\sum_{k\in[d]}\sum_{\{k_h\}_{h\in\mathcal{S}}} p^\pi(x_{L+1}=x_l=e_k, x_{l-i_h}=x_{L+1-j_h}=e_{k_h}, \forall h\in\mathcal{S})\\ &\qquad - \mu^\pi(x=e_k, x_{-i_h}=e_{k_h}\forall h\in\mathcal{S})\cdot\mu^\pi(z=e_k, z_{-j_h}=e_{k_h}\forall h\in\mathcal{S})\Bigg|\\ &\le \frac{2(M\vee r_n)}{L}+\frac{2\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}}{L(1-\lambda)},\end{aligned} \tag{H.7}$$

where $\mu_0(\cdot)$ is the initial distribution over the first r_n tokens. Then, by (H.5), (H.6), (H.7), and the triangular inequality, it holds that

$$\begin{aligned}&\Bigg|\frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\Bigg[\Bigg(\sum_{k\in[d]}\frac{\mathbb{1}(x_{L+1}=x_l=e_k)}{\mu^\pi(e_k)}\Bigg)\cdot\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Bigg]\\ &\qquad - \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\Bigg[\Bigg(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(e_k)}\cdot\Bigg(\prod_{h\in\mathcal{S}}\langle v_z^{(h)}, v_x^{(h)}\rangle\Bigg)\Bigg)\Bigg]\Bigg|\\ &\le \sum_{\{i_h,j_h\}_{h\in\mathcal{S}}}\prod_{h\in\mathcal{S}}\sigma_{-i_h}^{(h)}\sigma_{-j_h}^{(h)}\cdot\Bigg(\frac{2(M\vee r_n)}{L}+\frac{2\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}}{L(1-\lambda)}\Bigg)\\ &= \Bigg(\frac{2(M\vee r_n)}{L}+\frac{2\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}}{L(1-\lambda)}\Bigg).\end{aligned}$$

For the second absolute value term, the analagous argument can be applied. It follows form Lemma H.9 that

$$\begin{aligned}&\Bigg|\frac{1}{L}\sum_{l=1}^{L}\mathbb{E}\Bigg[\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle\Bigg] - \mathbb{E}_{(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\Bigg[\Bigg(\prod_{h\in\mathcal{S}}\langle v_z^{(h)}, v_x^{(h)}\rangle\Bigg)\Bigg]\Bigg|\\ &\le \Bigg(\frac{2(M\vee r_n)}{L}+\frac{2\sqrt{D_{\chi^2}(\mu_0\,\|\,\mu^\pi)+1}}{L(1-\lambda)}\Bigg).\end{aligned}$$

This completes the proof. □

Lemma H.5. *Consider $g_{3,\mathcal{S}}$ in (G.5). Then, it holds that*

$$\Bigg|\mathbb{E}_{\pi\sim\mathcal{P}}\left[g_{3,\mathcal{S}}\right]-\prod_{h\in\mathcal{S}}(\sigma_{-h}^{(h)})^2\cdot I_{\chi^2}(\mathcal{S})\Bigg|\le\Bigg(1-\prod_{h\in\mathcal{S}}(\sigma_{-h}^{(h)})^2\Bigg)\,I_{\chi^2}(\mathcal{S}^\star),$$

Proof of Lemma H.5. Since $v_l^{(h)}=\sum_{i_h\in[M]}\sigma_{-i_h}^{(h)}x_{l-i_h}$, we have

$$\prod_{h\in\mathcal{S}}\langle v_l^{(h)}, v_{L+1}^{(h)}\rangle=\sum_{\{i_h,j_h\}_{h\in\mathcal{S}}}\prod_{h\in\mathcal{S}}\sigma_{-i_h}^{(h)}\sigma_{-j_h}^{(h)}\,\mathbb{1}(x_{l-i_h}=x_{L+1-j_h}).$$

Recall that

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

Then the $\mathbb{E}_{\pi\sim\mathcal{P}}[g_{3,\mathcal{S}}]$ can be expressed as the summation of two terms:

$$\begin{aligned}
&\mathbb{E}_{\pi}[g_{3,\mathcal{S}}]\\
&= \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\sum_{\{i_h,j_h\}_{h\in\mathcal{S}}}\prod_{h\in\mathcal{S}}\sigma_{-i_h}^{(h)}\sigma_{-j_h}^{(h)}\,\mathbb{1}(x_{l-i_h}=z_{-j_h})\left(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(e_k)}-1\right)\right]\\
&= \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\prod_{h\in\mathcal{S}}(\sigma_{-h}^{(h)})^2\,\mathbb{1}(x_{-h}=z_{-h})\left(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(e_k)}-1\right)\right]\\
&+ \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\sum_{(i_h,j_h)\in\Gamma^c(\mathcal{S})}\prod_{h\in\mathcal{S}}\sigma_{-i_h}^{(h)}\sigma_{-j_h}^{(h)}\,\mathbb{1}(x_{-i_h}=z_{-j_h})\left(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(e_k)}-1\right)\right]
\end{aligned}$$

where the signal set is defined as $\Gamma(\mathcal{S}) := \{(i_h,j_h)\mid i_h=j_h=h,\forall h\in\mathcal{S}\}$ and the error set is $\Gamma^c(\mathcal{S}) := \{(i_h,j_h)\mid \forall h\in\mathcal{S}\}\setminus\Gamma(\mathcal{S})$. Note that we can upper bound the second term by Lemma H.6 as

$$\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\prod_{h\in\mathcal{S}}\mathbb{1}(x_{-i_h}=z_{-j_h})\left(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(z=e_k)}-1\right)\right]\le I_{\chi^2}(\mathcal{S}^\star).$$

Thus, the gradient is upper and lower bounded by

$$\prod_{h\in\mathcal{S}}(\sigma_{-h}^{(h)})^2\cdot I_{\chi^2}(\mathcal{S})\pm\left(1-\prod_{h\in\mathcal{S}}(\sigma_{-h}^{(h)})^2\right)I_{\chi^2}(\mathcal{S}^\star)$$

□

Lemma H.6. *Consider any $\mathcal{S}=\{i_1,\ldots,i_{|\mathcal{S}|}\},\mathcal{S}'=\{j_1,\ldots,j_{|\mathcal{S}'|}\}\in\mathcal{A}_H^{\le D}$ such that $|\mathcal{S}|=|\mathcal{S}'|$ It holds that*

$$\begin{aligned}
&\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\prod_{l\in[|\mathcal{S}|]}\mathbb{1}(x_{-i_l}=z_{-j_l})\left(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(z=e_k)}-1\right)\right]\\
&\qquad=\frac{1}{2}\left(\widetilde{I}_{\chi^2}(\mathcal{S})+\widetilde{I}_{\chi^2}(\mathcal{S}')\right)\le\widetilde{I}_{\chi^2}(\mathcal{S}^\star).
\end{aligned}$$

Proof of Lemma H.6. Note that

$$\begin{aligned}
&\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\prod_{l\in[|\mathcal{S}|]}\mathbb{1}(x_{-i_l}=z_{-j_l})\left(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(z=e_k)}-1\right)\right]\\
&= \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\sum_{\{k_l\}_{l\in|\mathcal{S}|}}\mathbb{1}(X_{-\mathcal{S}}=Z_{-\mathcal{S}'})\left(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(z=e_k)}-1\right)\right]\\
&= \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\left(\sum_{k\in[d]}\frac{\mu^\pi(x=e_k|X_{-\mathcal{S}})\cdot\mu^\pi(z=e_k|Z_{-\mathcal{S}})}{\mu^\pi(z=e_k)}-1\right)\right]\\
&= \mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\sum_{k\in[d]}\left(\frac{\mu^\pi(x=e_k|X_{-\mathcal{S}})}{\mu^\pi(z=e_k)}-1\right)\cdot\left(\frac{\mu^\pi(z=e_k|Z_{-\mathcal{S}})}{\mu^\pi(z=e_k)}-1\right)\cdot\mu^\pi(z=e_k)\right].
\end{aligned}\tag{H.8}$$

Then, we apply the inequality $ab \leq a^2 + b^2/2$ to the (H.8) and obtain the upper bound as follows:

$$
\begin{aligned}
&\frac{1}{2}\mathbb{E}_{\pi,(x,X)\sim\mu^\pi}\Bigg[\sum_{\{k_l\}_{l\in|S|}}\sum_{k\in[d]}\bigg(\frac{\mu^\pi(x=e_k|X_{-\mathcal{S}})}{\mu^\pi(z=e_k)}-1\bigg)^2\cdot\mu^\pi(z=e_k)\cdot\mu^\pi(X_{-\mathcal{S}})\Bigg] \\
&+\frac{1}{2}\mathbb{E}_{\pi,(z,Z)\sim\mu^\pi}\Bigg[\sum_{\{k_l\}_{l\in|S|}}\sum_{k\in[d]}\bigg(\frac{\mu^\pi(z=e_k|Z_{-\mathcal{S}})}{\mu^\pi(z=e_k)}-1\bigg)^2\cdot\mu^\pi(z=e_k)\cdot\mu^\pi(Z_{-\mathcal{S}})\Bigg] \\
&=\frac{1}{2}\widetilde{I}_{\chi^2}(\mathcal{S})+\frac{1}{2}\widetilde{I}_{\chi^2}(\mathcal{S}')\leq\widetilde{I}_{\chi^2}(\mathcal{S}^\star),
\end{aligned}
$$

where the equality follows from the definition of the modified mutual information and the last inequality follows from the definition of $\mathcal{S}^\star$. □

Lemma H.7. *Consider any $\mathcal{S}=\{i_1,\ldots,i_{|\mathcal{S}|}\},\mathcal{S}'=\{j_1,\ldots,j_{|\mathcal{S}|+1}\}\in\mathcal{A}_H^{\leq D}$ such that $|\mathcal{S}|+1=|\mathcal{S}'|$. Let $i_{|S|+1}=i_{l^*}$ for some $l^*\in[|S|]$. It holds that*

$$
\begin{aligned}
&\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\prod_{l\in[|\mathcal{S}|+1]}\mathbb{1}(x_{-i_l}=z_{-j_l})\bigg(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(z=e_k)}-1\bigg)\right] \\
&\qquad<\frac{1}{2}\widetilde{I}_{\chi^2}(\mathcal{S})+\frac{1}{2}\widetilde{I}_{\chi^2}(\mathcal{S}')\leq\widetilde{I}_{\chi^2}(\mathcal{S}^\star).
\end{aligned}
$$

Proof of Lemma H.7. The proof is similar to the proof of Lemma H.6. The left hand side of the inequality can be expressed as follows:

$$
\begin{aligned}
&\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\prod_{l\in[|\mathcal{S}|+1]}\mathbb{1}(x_{-i_l}=z_{-j_l})\bigg(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(z=e_k)}-1\bigg)\right] \\
&=\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\prod_{l\in[|\mathcal{S}|]}\mathbb{1}(x_{-i_l}=z_{-j_l})\,\mathbb{1}(z_{-i_{l^*}}=z_{-j_{|S|+1}})\bigg(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(z=e_k)}-1\bigg)\right] \\
&=\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\mathbb{1}(X_{-\mathcal{S}}=Z_{-\mathcal{S}'\setminus\{j_{|S|+1}\}})\,\mathbb{1}(z_{-j_{|S|+1}}=z_{-i_{l^*}})\bigg(\sum_{k\in[d]}\frac{\mathbb{1}(x=z=e_k)}{\mu^\pi(z=e_k)}-1\bigg)\right] \\
&=\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\left[\sum_{k\in[d]}\frac{\mu^\pi(x=e_k|X_{-\mathcal{S}})\cdot\mu^\pi(z=e_k|Z_{-\mathcal{S}'\setminus\{j_{|S|+1}\}},z_{-j_{|S|+1}}=z_{-i_{l^*}})}{\mu^\pi(z=e_k)}-1\right] \\
&=\mathbb{E}_{\pi,(x,X),(z,Z)\sim\mu^\pi\otimes\mu^\pi}\Bigg[\sum_{k\in[d]}\bigg(\frac{\mu^\pi(x=e_k|X_{-\mathcal{S}})}{\mu^\pi(z=e_k)}-1\bigg) \\
&\qquad\cdot\bigg(\frac{\mu^\pi(z=e_k|Z_{-\mathcal{S}'\setminus\{j_{|S|+1}\}},z_{-j_{|S|+1}}=z_{-i_{l^*}})}{\mu^\pi(z=e_k)}-1\bigg)\cdot\mu^\pi(z=e_k)\Bigg].
\end{aligned}
\tag{H.9}
$$

By the inequality $ab \le a^2 + b^2/2$ the upper bound of (H.9) can be derived as

$$\begin{aligned}
&\frac{1}{2}\mathbb{E}_{\pi,(x,X)\sim\mu^\pi}\Bigg[\sum_{k\in[d]}\bigg(\frac{\mu^\pi(x=e_k|X_{-\mathcal{S}})}{\mu^\pi(z=e_k)}-1\bigg)\cdot\mu^\pi(z=e_k)\cdot\mu^\pi(X_{-\mathcal{S}})\Bigg] \\
&+\frac{1}{2}\mathbb{E}_{\pi,(z,Z)\sim\mu^\pi}\Bigg[\sum_{k\in[d]}\bigg(\frac{\mu^\pi(z=e_k|Z_{-\mathcal{S}'\setminus\{j_{|S|+1}\}},z_{-j_{|S|+1}}=z_{-i_{l^*}})}{\mu^\pi(z=e_k)}-1\bigg)^2 \\
&\qquad\cdot\mu^\pi(z=e_k)\cdot\mu^\pi(Z_{-\mathcal{S}'\setminus\{j_{|S|+1}\}},z_{-j_{|S|+1}}=z_{-i_{l^*}})\Bigg] \\
&=\frac{1}{2}\widetilde{I}_{\chi^2}(\mathcal{S})+\frac{1}{2}\mathbb{E}_{\pi,(z,Z)\sim\mu^\pi}\Bigg[\sum_{k\in[d]}\bigg(\frac{\mu^\pi(z=e_k|Z_{-\mathcal{S}'})}{\mu^\pi(z=e_k)}-1\bigg)^2 \\
&\qquad\cdot\mu^\pi(z=e_k)\cdot\mu^\pi(Z_{-\mathcal{S}'})\cdot\mathbb{1}(z_{-j_{|S|+1}}=z_{-i_{l^*}})\Bigg] \\
&<\frac{1}{2}\widetilde{I}_{\chi^2}(\mathcal{S})+\frac{1}{2}\widetilde{I}_{\chi^2}(\mathcal{S}')\le\widetilde{I}_{\chi^2}(\mathcal{S}^\star),
\end{aligned}$$

where the equality follows from the definition of the modified mutual information and the last inequality follows from the definition of $\mathcal{S}^\star$. □

H.2. Lemmas on Concentration of Markov Chain

For simplicity, we denote by $\mathcal{M}(l)=\{l-1,l-2,\dots,l-M\}$ the length-M window before l. Also, recall that we have the parent set $\texttt{pa}(l)=\{-r_1,\dots,-r_n\}$ and we define $\mathcal{N}(l)=\{l-1,\dots,l-r_n\}$ as the minimal set of continuous indices that contains $\texttt{pa}(l)$. We denote by $p^\pi(\cdot)$ the joint distribution of the chain (X,x_{L+1}) under the Markov kernel π. For $\mathcal{M}(l)$ or $\mathcal{N}(l)$ that goes to the negative index, we extend $p^\pi(\cdot)$ to be

$$p^\pi(x_{L+1},X,X_{\mathcal{M}(1)})=p^\pi(x_{L+1},X)\cdot\prod_{l\in\mathcal{M}(1)}\mathbb{1}(x_l=\mathbf{0}),$$

where we extend the space of $\mathcal{X}$ to also include the zero vector $\mathbf{0}$.

Let us first introduce the notations to be used in the later proof. For more generality, let us take Y_{L+1} as a subset of (x_{L+1},X) such that the maximal index and minimal index within Y_{L+1} have difference at most $m+1$. Here, m is just an integer less than L. Two special cases of the definition is $Y_{L+1}=\{x_{L+1},X_{\mathcal{M}(L+1)}\}$ with $m=M$ and $Y_{L+1}=\{x_{L+1}\}$ with $m=0$ which will be studied extensively. Take Y_l as the the subset with indices shifted from Y_{L+1} by $-(L+1-l)$. Let $A=X_{\mathcal{N}(L-m+r_n+1)}$ and $B_l=X_{\mathcal{N}(l+1)}$. By the Markov property, we have

$$Y_{L+1}\perp\!\!\!\perp(B_l,Y_l)\,|\,A,\quad(Y_{L+1},A)\perp\!\!\!\perp Y_l\,|\,B_l,\quad\forall l\in[L-m+r_n],$$

The quantity of interest here is

$$\begin{aligned}
\widehat{p}^\pi(E,E') &:= \frac{1}{L}\sum_{l=1}^{L}p^\pi(Y_{L+1}=E,Y_l=E') \\
&=\frac{1}{L}\sum_{l=1}^{L}\sum_{A,b}p^\pi(Y_{L+1}=E\,|\,A)\cdot\pi^{(L-l-(m-r_n))}(A\,|\,B_l=b) \\
&\qquad\cdot p^\pi(Y_l=E'\,|\,B_l=b)\cdot p^\pi(B_l=b).
\end{aligned}\tag{H.10}$$

Here, we denote by $\pi^{(i)}$ the i-step transition kernel of the chain. In the matrix form, let K of shape $|\mathcal{X}^{r_n}|\times|\mathcal{X}^{r_n}|$ be the transition matrix such that $K_{ij}=\pi(j\,|\,i)$. Let μ denote the vector of the stationary distribution of the chain with element $\mu(i)=\mu^\pi(i)$. Let us consider the reweighted transition kernel

$$\widetilde{K}=\mathrm{diag}\big(\sqrt{\mu}^{-1}\big)\cdot K\cdot\mathrm{diag}\left(\sqrt{\mu}\right),$$

Since the transition matrix is *primitive* by assumption and having only one eigenvalue 1 on its spectral circle, we also have for $\widetilde{K}$ that the leading eigenvalue is 1 with eigenvector $\sqrt{\mu}$. However, the projection in the leading eigenspace (or the Perron projection) is not of our interest. We note that

$$K^i - \mu\mathbf{1}^\top = \mathrm{diag}\left(\sqrt{\mu}\right)\cdot\left(\widetilde{K} - \sqrt{\mu}\sqrt{\mu}^\top\right)^i \cdot \mathrm{diag}\left(\sqrt{\mu}^{-1}\right).$$

Thus, it is the eigenvalue of second largest magnitude that matters when studying the convergence of the chain. Let λ denote the second largest magnitude of the eigenvalues of $\widetilde{K}$. Before we proceed to study $\widehat{p}^\pi$, let us first study a simpler convergence result, which is to quantify the closeness between $\sum_{l=1}^L \eta^{L-l} p^\pi(B_l = b)/(\sum_{l=1}^L \eta^{L-l})$ and $\mu^\pi(b)$ for certain $\eta \in (0, 1]$.

Lemma H.8. *The following two inequalities hold for length-r_n window:*

$$\left\| \frac{\sum_{l=1}^L \lambda^{L-l} p^\pi(B_l = \cdot)}{\sum_{l=1}^L \lambda^{L-l}} - \mu^\pi(\cdot) \right\|_{\mathrm{TV}} \le \frac{L\cdot\lambda^{L-r_n}\cdot(1-\lambda)}{1-\lambda^L}\cdot\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1},$$

$$\left\| \frac{\sum_{l=1}^L p^\pi(B_l = \cdot)}{L} - \mu^\pi(\cdot) \right\|_{\mathrm{TV}} \le \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda)} + \frac{r_n}{L},$$

where $D_{\chi^2}(\mu_0 \,\|\, \mu^\pi)$ is the χ^2 divergence between the initial distribution μ_0 and the stationary distribution μ^π. For a set Y_l that can be covered by a length-m window, we have

$$\left\| \frac{\sum_{l=1}^L p^\pi(Y_l = \cdot)}{L} - \mu^\pi(\cdot) \right\|_{\mathrm{TV}} \le \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda)} + \frac{m\vee r_n}{L},$$

$$\|p^\pi(Y_{L+1} = \cdot) - \mu^\pi(Y_{L+1} = \cdot)\|_{\mathrm{TV}} \le \lambda^{L-m\vee r_n}\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}. \tag{H.11}$$

Proof of Lemma H.8. Let $c_l = \eta^{L-l}/\sum_{l=1}^L \eta^{L-l}$. Let $\mu_0 \in \mathcal{X}^{r_n}$ be the vector of the initial distribution of the chain. Using the matrix representation, we have

$$\begin{aligned}
\sum_{l=r_n}^L c_l\cdot(p^\pi(B_l = b) - \mu^\pi(b)) &= \sum_{l=r_n}^L c_l \cdot \mathbf{1}_B^\top \cdot K^{l-r_n}\cdot(\mu_0 - \mu) \\
&= \sum_{l=r_n}^L c_l \cdot \mathbf{1}_B^\top \cdot (K^{l-r_n} - \mu\mathbf{1}^\top)\cdot\mu_0 \\
&= \sum_{l=r_n}^L c_l \cdot \mathbf{1}_B^\top \cdot \mathrm{diag}\left(\sqrt{\mu}\right)\cdot\left(\widetilde{K} - \sqrt{\mu}\sqrt{\mu}^\top\right)^{l-r_n}\cdot\mathrm{diag}\left(\sqrt{\mu}^{-1}\right)\cdot\mu_0.
\end{aligned}$$

To conclude, we use the variational representation of the total variation distance and have for any test vector $u \in \{0,1\}^{|\mathcal{X}^{r_n}|}$ that

$$\begin{aligned}
u^\top\cdot\sum_{l=r_n}^L c_l\cdot(p^\pi(B_l = \cdot) - \mu^\pi(\cdot)) &\le \sum_{l=r_n}^L c_l \cdot \underbrace{u^\top\cdot\mathrm{diag}\left(\sqrt{\mu}\right)}_{\|\cdot\|_2\le 1}\cdot\left(\widetilde{K} - \sqrt{\mu}\sqrt{\mu}^\top\right)^{l-r_n}\cdot\mathrm{diag}\left(\sqrt{\mu}^{-1}\right)\cdot\mu_0 \\
&\le \sum_{l=r_n}^L c_l\cdot\lambda^{l-r_n}\cdot\left\|\mathrm{diag}\left(\sqrt{\mu}^{-1}\right)\cdot\mu_0\right\|_2 \\
&= \sum_{l=r_n}^L c_l\cdot\lambda^{l-r_n}\cdot\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}.
\end{aligned}$$

Plugging in the definition of c_l, we have

$$\left\| \frac{\sum_{l=1}^L \eta^{L-l} p^\pi(B_l = b)}{\sum_{l=1}^L \eta^{L-l}} - \mu^\pi(b) \right\|_{\mathrm{TV}} \le \frac{\sum_{l=r_n}^L \eta^{L-l}\cdot\lambda^{l-r_n}\cdot\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1} + \sum_{l=1}^{r_n-1}\eta^{L-l}}{\sum_{l=r_n}^L \eta^{L-l} + \sum_{l=1}^{r_n-1}\eta^{L-l}}.$$

We consider two special cases. In the first case, we set $\eta = \lambda$, which gives us

$$\left\| \frac{\sum_{l=1}^{L} \lambda^{L-l} p^\pi(B_l = b)}{\sum_{l=1}^{L} \lambda^{L-l}} - \mu^\pi(b) \right\|_{\mathrm{TV}} \le \frac{\sum_{l=r_n}^{L} \lambda^{L-r_n} \cdot \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1} + \sum_{l=1}^{r_n - 1} \lambda^{L-l}}{(1-\lambda^L)/(1-\lambda)}$$
$$\le \frac{L \cdot \lambda^{L-r_n} \cdot (1-\lambda)}{1-\lambda^L} \cdot \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}.$$

In the second case, we set $\eta = 1$, which gives us

$$\left\| \frac{\sum_{l=1}^{L} p^\pi(B_l = b)}{L} - \mu^\pi(b) \right\|_{\mathrm{TV}} \le \frac{\sum_{l=r_n}^{L} \lambda^{l-r_n} \cdot \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1} + r_n - 1}{L}$$
$$\le \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda)} + \frac{r_n}{L}.$$

Similar results can also be derived for a length-$(m+1)$ windows. Note that

$$\left\| \frac{\sum_{l=1}^{L} p^\pi(Y_l = \cdot)}{L} - \mu^\pi(\cdot) \right\|_{\mathrm{TV}} = \left\| \frac{\sum_{l=1}^{L} p^\pi(B_{l-(m-r_n)\vee 0} = \cdot)}{L} - \mu^\pi(\cdot) \right\|_{\mathrm{TV}}$$
$$\le \frac{m \vee r_n}{L} + \left\| \frac{\sum_{l=r_n}^{L-(m-r_n)\vee 0} p^\pi(B_l = \cdot)}{L} - \frac{L - m \vee r_n}{L} \cdot \mu^\pi(\cdot) \right\|_{\mathrm{TV}}$$
$$\le \frac{m \vee r_n}{L} + \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda)}.$$

Lastly, we consider the difference between $p^\pi(Y_{L+1} = \cdot)$ and $\mu^\pi(\cdot)$.

$$\begin{aligned}
&\|p^\pi(Y_{L+1} = \cdot) - \mu^\pi(Y_{L+1} = \cdot)\|_{\mathrm{TV}} \\
&\quad \le \|p^\pi(B_{L+1-(m-r_n)\vee 0} = \cdot) - \mu^\pi(B_{L+1-(m-r_n)\vee 0} = \cdot)\|_{\mathrm{TV}} \\
&\quad \le \max_{u \in \{0,1\}^{d^{r_n}}} u^\top \cdot \mathrm{diag}\left(\sqrt{\mu}\right) \cdot \left(\widetilde{K} - \sqrt{\mu}\sqrt{\mu}^\top\right)^{L - m \vee r_n} \cdot \mathrm{diag}\left(\sqrt{\mu}^{-1}\right) \cdot \mu_0 \\
&\quad \le \lambda^{L - m \vee r_n} \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}.
\end{aligned}$$

Hence, the proof is completed. □

Now that we know that the average of $p^\pi(B_l = \cdot)$ converges to $\mu^\pi(\cdot)$, which is "first-ordered" convergence. The next question is whether $\widehat{p}^\pi(\cdot,\cdot)$ converges to $\mu^\pi(\cdot) \cdot \mu^\pi(\cdot)$. The following lemma quantifies the total variation distance between the distribution $\widehat{p}^\pi$ and the product of two stationary distributions.

Lemma H.9. *For $\widehat{p}^\pi$ defined in (H.10), we have*

$$\|\widehat{p}^\pi(\cdot,\cdot) - \mu^\pi(\cdot)\mu^\pi(\cdot)\|_{\mathrm{TV}} \le \frac{2(M \vee r_n)}{L} + \frac{2\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda)}.$$

In particular,

$$\left\| \widehat{p}^\pi(E, E') - \mu^\pi(E) \cdot \left(\frac{1}{L} \sum_{l=1}^{L} p^\pi(Y_l = E') \right) \right\|_{\mathrm{TV}} \le \frac{M \vee r_n}{L} + \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi) + 1}}{L(1-\lambda)}. \tag{H.12}$$

Proof of Lemma H.9. We want to control the difference between $\widehat{p}^\pi$ and the averaged product distribution of Y_{L+1} and Y_l, which is given by

$$
\begin{aligned}
&\widehat{p}^{\pi}(E, E') - \mu^{\pi}(E) \cdot \left(\frac{1}{L}\sum_{l=1}^{L} p^{\pi}(Y_l = E')\right) \\
&\quad = \frac{1}{L}\sum_{l=1}^{L}\sum_{A,b} p^{\pi}(Y_{L+1} = E \mid A) \cdot \left(\pi^{(L-l-(M-r_n))}(A \mid B_l = b) - \mu^{\pi}(A)\right) \\
&\qquad\qquad \cdot p^{\pi}(Y_l = E' \mid B_l = b) \cdot p^{\pi}(B_l = b).
\end{aligned} \tag{H.13}
$$

We can also rewrite (H.13) in the matrix form as

$$
\begin{aligned}
&\widehat{p}^{\pi}(\cdot,\cdot) - \mu^{\pi}(\cdot) \cdot \left(\frac{1}{L}\sum_{l=1}^{L} p^{\pi}(Y_l = \cdot)\right) \\
&\quad = \frac{1}{L}\sum_{l=1}^{L} p^{\pi}(Y_{L+1} = \cdot \mid A = \cdot) \cdot \operatorname{diag}(\sqrt{\mu}) \cdot \left(\widetilde{K} - \sqrt{\mu}\sqrt{\mu}^{\top}\right)^{L-l-(M-r_n)} \cdot \operatorname{diag}\left(\sqrt{\mu}^{-1}\right) \\
&\qquad\qquad \cdot \operatorname{diag}(p^{\pi}(B_l = \cdot)) \cdot p^{\pi}(Y_l = \cdot \mid B_l = \cdot)^{\top}.
\end{aligned}
$$

When considering the ℓ_1-norm of the difference between the two distributions, we introduce a test matrix U of shape $|\mathcal{X}^M| \times |\mathcal{X}^M|$ with each element of U chosen from $\{0,1\}$. Then, we have

$$
\begin{aligned}
\mathrm{TV}_1 &:= \left\| \widehat{p}^{\pi}(\cdot,\cdot) - \mu^{\pi}(\cdot) \cdot \left(\frac{1}{L}\sum_{l=1}^{L} p^{\pi}(Y_l = \cdot)\right) \right\|_{\mathrm{TV}} \\
&\le \max_{U} \operatorname{Tr}\Bigg[\frac{1}{L}\sum_{l=1}^{L} p^{\pi}(Y_{L+1} = \cdot \mid A = \cdot) \cdot \operatorname{diag}(\sqrt{\mu}) \cdot \left(\widetilde{K} - \sqrt{\mu}\sqrt{\mu}^{\top}\right)^{L-l-(M-r_n)} \\
&\qquad\qquad \cdot \operatorname{diag}\left(\sqrt{\mu}^{-1}\right) \cdot \operatorname{diag}(p^{\pi}(B_l = \cdot)) \cdot p^{\pi}(Y_l = \cdot \mid B_l = \cdot)^{\top} \cdot U(\cdot,\cdot)^{\top}\Bigg].
\end{aligned}
$$

To upper bound this quantity, we consider each row of U as $U(Y_{L+1}, \cdot) = u(\cdot \mid Y_{L+1})^{\top}$. Note that $u(\cdot \mid Y_{L+1})$ is also a $\{0,1\}$-valued vector. In this spirit, we have

$$
\begin{aligned}
\mathrm{TV}_1 \le \sum_{E} \max_{u(\cdot \mid Y_{L+1}=E)} &\frac{1}{L}\sum_{l=1}^{L} p^{\pi}(Y_{L+1} = b \mid A = \cdot) \cdot \operatorname{diag}(\sqrt{\mu}) \cdot \left(\widetilde{K} - \sqrt{\mu}\sqrt{\mu}^{\top}\right)^{L-l-(M-r_n)} \\
&\cdot \operatorname{diag}\left(\sqrt{\mu}^{-1}\right) \cdot \operatorname{diag}(p^{\pi}(B_l = \cdot)) \cdot p^{\pi}(Y_l = \cdot \mid B_l = \cdot)^{\top} \cdot u(\cdot \mid Y_{L+1} = E).
\end{aligned}
$$

Note that the norm of the vector in the last line is at most

$$
\begin{aligned}
&\left\| \operatorname{diag}\left(\sqrt{\mu}^{-1}\right) \cdot \operatorname{diag}(p^{\pi}(B_l = \cdot)) \cdot p^{\pi}(Y_l = \cdot \mid B_l = \cdot)^{\top} \cdot u(\cdot \mid Y_{L+1} = E) \right\|_2 \\
&\quad \le \left\| \operatorname{diag}\left(\sqrt{\mu}^{-1}\right) \cdot \operatorname{diag}(p^{\pi}(B_l = \cdot)) \cdot \mathbf{1} \right\|_2 \\
&\quad \le \sqrt{D_{\chi^2}(p^{\pi}(B_l = \cdot) \,\|\, \mu^{\pi}(\cdot)) + 1} \le \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^{\pi}) + 1},
\end{aligned}
$$

where the first inequality holds by noting that $p^{\pi}(Y_l = \cdot \mid B_l = \cdot)^{\top} \cdot u(\cdot \mid Y_{L+1} = E)$ is a vector with element within $[0,1]$.

The last inequality is the data processing inequality. Consequently, we have for the TV distance that

$$\begin{aligned}\mathrm{TV}_1 &\le \frac{M-r_n}{L} + \frac{1}{L}\sum_{l=1}^{L-(M-r_n)} \lambda^{L-l-(M-r_n)} \cdot \sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi)+1} \\ &\quad \cdot \sum_{E,b} \max_{v(\cdot\,|\,E): \|v(\cdot\,|\,E)\|_2\le 1} p^\pi(Y_{L+1}=E\,|\,A=b)\cdot\sqrt{\mu^\pi(b)}\cdot v(b\,|\,E) \\ &\le \frac{M-r_n}{L} + \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi)+1}}{L(1-\lambda)} \cdot \max_{\{v(\cdot\,|\,E)\}_E:\, \|v(\cdot\,|\,E)\|_2\le 1} \sqrt{\sum_{E,b} v(b\,|\,E)^2 p^\pi(Y_{L+1}=E\,|\,A=b)} \\ &\le \frac{M\vee r_n}{L} + \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi)+1}}{L(1-\lambda)},\end{aligned}$$

where the first inequality follows from the spectral norm of the matrix $\widetilde{K} - \sqrt{\mu}\sqrt{\mu}^\top$, and the second inequality follows from the Cauchy-Schwarz inequality. Now, it remains to quantify the TV distance

$$\mathrm{TV}_2 := \left\| \mu^\pi(\cdot)\cdot\left(\frac{1}{L}\sum_{l=1}^{L} p^\pi(Y_l=\cdot)\right) - \mu^\pi(\cdot)\cdot\mu^\pi(\cdot)\right\|_{\mathrm{TV}} = \left\|\left(\frac{1}{L}\sum_{l=1}^{L} p^\pi(Y_l=\cdot)\right) - \mu^\pi(\cdot)\right\|_{\mathrm{TV}}.$$

Invoking Lemma H.8, we have this quantity upper bounded by

$$\mathrm{TV}_2 \le \frac{\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi)+1}}{L(1-\lambda)} + \frac{M}{L}.$$

Using the triangular inequality for the total variation distance, we have

$$\|\widehat{p}^\pi(\cdot,\cdot) - \mu^\pi(\cdot)\mu^\pi(\cdot)\|_{\mathrm{TV}} \le \mathrm{TV}_1 + \mathrm{TV}_2 \le \frac{2M}{L} + \frac{2\sqrt{D_{\chi^2}(\mu_0 \,\|\, \mu^\pi)+1}}{L(1-\lambda)}.$$

Hence, the proof is completed. □

In the following, we use a similar technique as in Lemma H.9 to derive a bound for the chi-square distance.

Lemma H.10. *Consider Y_l has a cover of size m, i.e., there exists a $j \in [L]$ and a successive sequence $\{x_j, \cdots, x_{j+m-1}\}$ such that Y_l is a subset of the sequence. Then, for the chi-square divergence between the empirical distribution $L^{-1}\sum_{l=1}^{L} \mathbb{1}(Y_l=\cdot)$ and the stationary distribution $\pi^\pi(\cdot)$, we have*

$$D_{\chi^2}\left(L^{-1}\sum_{l=1}^{L}\mathbb{1}(Y_l=\cdot)\,\Big\|\,\mu^\pi(\cdot)\right) \le \frac{D_{\chi^2}(\mu_0(B=\cdot)\,\|\,\mu^\pi(B=\cdot))+1}{L(1-\lambda)\cdot\min_E \mu^\pi(Y=E)} + \frac{2r_n\vee(3m-r_n)}{L\min_E \mu^\pi(Y=E)}.$$

Proof of Lemma H.10. What we aim to bound is just

$$\mathbb{E}\left[\sum_E \frac{\left(L^{-1}\sum_{l=1}^{L}\mathbb{1}(Y_l=E)-\mu^\pi(E)\right)^2}{\mu^\pi(E)}\right] = \mathbb{E}\left[\sum_E \frac{L^{-2}\sum_{l,l'\in[L]^2}\mathbb{1}(Y_l=Y_{l'}=E)-\mu^\pi(E)^2}{\mu^\pi(E)}\right],$$

Let us separate this term into two parts:

$$\begin{aligned}J_1 &:= \mathbb{E}\left[\sum_E \frac{L^{-2}\sum_{l,l'\in[L]^2}\mathbb{1}(Y_l=Y_{l'}=E) - L^{-1}\sum_{l\in[L]}\mathbb{1}(Y_l=E)\mu^\pi(E)}{\mu^\pi(E)}\right], \\ J_2 &:= \mathbb{E}\left[\sum_E \frac{L^{-1}\sum_{l\in[L]}\mathbb{1}(Y_l=E)\mu^\pi(E)-\mu^\pi(E)^2}{\mu^\pi(E)}\right] = \mathbb{E}\left[\sum_E\left(L^{-1}\sum_{l\in[L]}\mathbb{1}(Y_l=E)-\mu^\pi(E)\right)\right] = 0.\end{aligned}$$

Following our convention, we let B_l and $B_{l'}$ be two length-r_n window such that

$$Y_{l+1} \perp\!\!\!\perp (B_{l'}, Y_{l'}) \,|\, B_l, \quad (Y_{l+1}, B_l) \perp\!\!\!\perp Y_{l'} \,|\, B_{l'}.$$

For the first part J_1, let us fix an index $l \geq r_n \vee m + (m - r_n) \vee 0$ and take a summation over $m \vee r_n \leq l' \leq l - (m - r_n) \vee 0$. Let $\tau = (m - r_n) \vee 0$ and $\varrho = m \vee r_n$. This gives us

$$\begin{aligned} J_1(l) &:= \frac{1}{L^2} \sum_{l'=\varrho}^{l-\tau} \sum_{B_l, b} p^\pi(Y_l = E \,|\, B_l) \cdot \Big(\pi^{(l-l'-\tau)}(B_l \,|\, B_l = b) - \mu^\pi(B_l)\Big) \\ &\qquad \cdot p^\pi(Y_{l'} = E \,|\, B_{l'} = b) \cdot p^\pi(B_{l'} = b) \cdot \mu^\pi(E)^{-1} \\ &= \frac{1}{L^2} \sum_{l'=\varrho}^{l-\tau} \mathrm{Tr}\Big[p^\pi(Y_l = \cdot \,|\, B_l = \cdot) \cdot \mathrm{diag}\,(\sqrt{\mu}) \cdot \big(\widetilde{K} - \sqrt{\mu}\sqrt{\mu}^\top\big)^{l-l'-\tau} \\ &\qquad \cdot \mathrm{diag}\big(\sqrt{\mu}^{-1}\big) \cdot \mathrm{diag}(p^\pi(B_{l'} = \cdot)) \cdot p^\pi(Y_{l'} = \cdot \,|\, B_{l'} = \cdot)^\top \cdot \mathrm{diag}(\mu^\pi(Y_{l'} = \cdot)^{-1})\Big]. \end{aligned}$$

We next invoke the Cauchy-Schwarz inequality for trace, i.e., $\mathrm{Tr}(W^\top V)^2 \leq \mathrm{Tr}(W^\top W)\,\mathrm{Tr}(V^\top V)$, and take

$$\begin{aligned} W^\top &= \mathrm{diag}(\mu^\pi(Y_l = \cdot)^{-1/2}) \cdot p^\pi(Y_l = \cdot \,|\, A = \cdot) \cdot \mathrm{diag}\,(\sqrt{\mu}) \cdot \big(\widetilde{K} - \sqrt{\mu}\sqrt{\mu}^\top\big)^{l-l'-\tau}, \\ V &= \mathrm{diag}\big(\sqrt{\mu}^{-1}\big) \cdot \mathrm{diag}(p^\pi(B_{l'} = \cdot)) \cdot p^\pi(Y_{l'} = \cdot \,|\, B_{l'} = \cdot)^\top \cdot \mathrm{diag}(\mu^\pi(Y_{l'} = \cdot)^{-1/2}), \end{aligned}$$

which gives us

$$J_1(l) \leq \frac{1}{L^2} \sum_{l'=\varrho}^{l-\tau} \lambda^{l-l'-\tau} \cdot \sqrt{\left\langle \frac{p^\pi(Y_{l'} = \cdot, B_{l'} = \cdot)^2}{\mu^\pi(B_{l'} = \cdot)\mu^\pi(Y_{l'} = \cdot)} \right\rangle \cdot \left\langle \frac{p^\pi(Y_l = \cdot, B_l = \cdot)^2}{\mu^\pi(Y_l = \cdot)\mu^\pi(B_l = \cdot)} \right\rangle}.$$

Here, we use the bracket $\langle \cdot \rangle$ to denote summation over the variables represented by "$\cdot$". We further have

$$\begin{aligned} \left\langle \frac{p^\pi(Y_l = \cdot, B_l = \cdot)^2}{\mu^\pi(Y_l = \cdot)\mu^\pi(B_l = \cdot)} \right\rangle &\leq \max_{b,E} \frac{p^\pi(Y_l = E \,|\, B_l = b)}{\mu^\pi(Y_l = E)} \cdot \left\langle \frac{p^\pi(B_l = \cdot)^2}{\mu^\pi(B_l = \cdot)} \right\rangle \\ &\leq \frac{1}{\min_E \mu^\pi(Y_l = E)} \cdot \big(D_{\chi^2}(\mu_0(B = \cdot) \,\|\, \mu^\pi(B = \cdot)) + 1\big), \end{aligned}$$

where the second inequality holds by the data processing inequality. Therefore, we conclude that

$$J_1(l) \leq \frac{D_{\chi^2}(\mu_0(B = \cdot) \,\|\, \mu^\pi(B = \cdot)) + 1}{L^2(1 - \lambda) \cdot \min_E \mu^\pi(Y = E)}.$$

For the remaining term not included in J_1, we note that each term indexed by l, l' is at most $(L^2 \min_E \mu^\pi(Y = E))^{-1}$ in value and we have at most $L \cdot (2\tau + \varrho)$ of these terms. As a result, we conclude that

$$J_1 \leq \frac{D_{\chi^2}(\mu_0(B = \cdot) \,\|\, \mu^\pi(B = \cdot)) + 1}{L(1 - \lambda) \cdot \min_E \mu^\pi(Y = E)} + \frac{2r_n \vee (3m - r_n)}{L \min_E \mu^\pi(Y = E)}.$$

Since the second term is 0, we complete the proof. □

Proof of Proposition G.3. To unify the notations, we let $Z = (z_{-M}, \ldots, z_{-1})$ and define

$$\begin{aligned} \widehat{\mu}_X^\pi(z, Z) &= \frac{1}{L} \sum_{l=1}^{L} \mathbb{1}(x_l = z, X_{\mathcal{M}(l)} = Z), \\ R(Z, X) &= \exp\left(a \cdot \prod_{h \in \mathcal{S}^\star} \mathbb{1}(z_{-h} = x_{L+1-h})\right). \end{aligned}$$

Using these notations, we can define the normalizing factor in $\widetilde{\mu}_X^\pi$ and $y_X^\star$ respectively as

$$\Phi = \sum_{z,Z} \mu^\pi(z,Z) \cdot R(Z,X), \quad \widehat{\Phi} = \sum_{z,Z} \widehat{\mu}_X^\pi(z,Z) \cdot R(Z,X).$$

We also define

$$\phi(z) = \sum_{Z} \mu^\pi(z,Z) \cdot R(Z,X), \quad \widehat{\phi}(z) = \sum_{Z} \widehat{\mu}_X^\pi(z,Z) \cdot R(Z,X).$$

We can then rewrite the objective as

$$\begin{aligned}
\|\widetilde{\mu}_X^\pi(e_k) - y_X^\star(k)\|_1 &= \sum_z \left| \frac{\phi(z)}{\Phi} - \frac{\widehat{\phi}(z)}{\widehat{\Phi}} \right| \\
&\le \sum_z \frac{\widehat{\phi}(z) \cdot |\widehat{\Phi} - \Phi| + |\phi(z) - \widehat{\phi}(z)| \cdot \widehat{\Phi}}{\Phi \cdot \widehat{\Phi}} \\
&= \frac{|\widehat{\Phi} - \Phi| + \sum_z |\phi(z) - \widehat{\phi}(z)|}{\Phi} \le \frac{2\sum_z |\phi(z) - \widehat{\phi}(z)|}{\Phi}.
\end{aligned} \tag{H.14}$$

Furthermore, notice that

$$\begin{aligned}
\frac{\sum_z |\phi(z) - \widehat{\phi}(z)|}{\Phi} &= \frac{\sum_z |\sum_Z (\mu^\pi(z,Z) - \widehat{\mu}_X^\pi(z,Z)) \cdot R(Z,X)|}{\sum_{z,Z} \mu^\pi(z,Z) \cdot R(Z,X)} \\
&\le \frac{\sum_z |\sum_Z (\mu^\pi(z,Z) - \widehat{\mu}_X^\pi(z,Z))| + (e^a - 1)\sum_z \left|\sum_{Z\in\Gamma_X} (\mu^\pi(z,Z) - \widehat{\mu}_X^\pi(z,Z))\right|}{1 + (e^a-1) \cdot \sum_z \sum_{Z\in\Gamma_X} \mu^\pi(z,Z)} \\
&\le \sum_z |\mu^\pi(z) - \widehat{\mu}_X^\pi(z)| + \frac{\sum_z \left|\sum_{Z\in\Gamma_X} (\mu^\pi(z,Z) - \widehat{\mu}_X^\pi(z,Z))\right|}{\sum_z \sum_{Z\in\Gamma_X} \mu^\pi(z,Z)}.
\end{aligned} \tag{H.15}$$

where we define $\Gamma_X = \{Z : Z_{-\mathcal{S}^\star} = X_{L+1-\mathcal{S}^\star}\}$. For the first term, we have by Cauchy-Schwarz that

$$\begin{aligned}
\mathbb{E}_X\left[\sum_z |\mu^\pi(z) - \widehat{\mu}_X^\pi(z)|\right] &\le \sqrt{\mathbb{E}_X\left[\sum_z \frac{(\mu^\pi(z) - \widehat{\mu}_X^\pi(z))^2}{\mu^\pi(z)}\right]} \\
&\le \sqrt{\frac{D_{\chi^2}(\mu_0(\cdot) \,\|\, \mu^\pi(\cdot)) + 1}{L(1-\lambda) \cdot \mu^\pi_{\min}} + \frac{r_n}{L \cdot \mu^\pi_{\min}}},
\end{aligned}$$

where in the last inequality, we invoke Lemma H.10 for a length-1 window. For the second term, we note that

$$\begin{aligned}
&\mathbb{E}_X\left[\frac{\sum_z \left|\sum_{Z\in\Gamma_X} (\mu^\pi(z,Z) - \widehat{\mu}_X^\pi(z,Z))\right|}{\sum_z \sum_{Z\in\Gamma_X} \mu^\pi(z,Z)}\right] \\
&= \sum_{E,z} \mathbb{E}_X\left[\frac{|\mu^\pi(z, Z_{-\mathcal{S}^\star} = E) - \widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star} = E)|}{\mu^\pi(Z_{-\mathcal{S}^\star} = E)} \mathbb{1}(X_{L+1-\mathcal{S}^\star} = E)\right] \\
&\le \sum_{E,z} \sqrt{\mathbb{E}_X\left[\left(\frac{\mu^\pi(z, Z_{-\mathcal{S}^\star} = E) - \widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star} = E)}{\sqrt{\mu^\pi(Z_{-\mathcal{S}^\star} = E)}}\right)^2\right] \cdot \frac{p^\pi(X_{L+1-\mathcal{S}^\star} = E)}{\mu^\pi(Z_{-\mathcal{S}^\star} = E)}} \\
&\le \sqrt{\mathbb{E}_X\left[\sum_{E,z} \frac{(\mu^\pi(z, Z_{-\mathcal{S}^\star} = E) - \widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star} = E))^2}{\mu^\pi(Z_{-\mathcal{S}^\star} = E)}\right] \cdot \sum_{E,z} \frac{p^\pi(X_{L+1-\mathcal{S}^\star} = E)}{\mu^\pi(Z_{-\mathcal{S}^\star} = E)}},
\end{aligned} \tag{H.16}$$

where the last two inequalities hold by the Cauchy-Schwarz inequality. We have an upper bound for the second term as

$$\sqrt{\sum_{E,z} \frac{p^\pi(X_{L+1-\mathcal{S}^\star} = E)}{\mu^\pi(Z_{-\mathcal{S}^\star} = E)}} \leq \sqrt{\frac{1}{\min_E \mu^\pi(Z_{-\mathcal{S}^\star} = E)}}. \tag{H.17}$$

We can also apply Lemma H.10 to the first term and conclude that

$$\begin{aligned} &\sqrt{\mathbb{E}_X \left[\sum_{E,z} \frac{\left(\mu^\pi(z, Z_{-\mathcal{S}^\star} = E) - \widehat{\mu}_X^\pi(z, Z_{-\mathcal{S}^\star} = E)\right)^2}{\mu^\pi(Z_{-\mathcal{S}^\star} = E)} \right]} \\ &\leq \sqrt{\frac{D_{\chi^2}(\mu_0(\cdot) \,\|\, \mu^\pi(\cdot)) + 1}{L(1-\lambda) \cdot \min_{z, Z_{-\mathcal{S}^\star}} \mu^\pi(z, Z_{-\mathcal{S}^\star})} + \frac{3M}{L \min_{z, Z_{-\mathcal{S}^\star}} \mu^\pi(z, Z_{-\mathcal{S}^\star})}}. \end{aligned} \tag{H.18}$$

In summary, we have

$$\begin{aligned} \mathbb{E}_X \left[\|\widetilde{\mu}_X^\pi(e_k) - y_X^\star(k)\|_1 \right] \leq\ & \frac{2}{\min_{z, Z_{-\mathcal{S}^\star}} \mu^\pi(z, Z_{-\mathcal{S}^\star})} \cdot \sqrt{\frac{D_{\chi^2}(\mu_0(\cdot) \,\|\, \mu^\pi(\cdot)) + 1}{L(1-\lambda)} + \frac{3M}{L}} \\ & + 2\sqrt{\frac{D_{\chi^2}(\mu_0(\cdot) \,\|\, \mu^\pi(\cdot)) + 1}{L(1-\lambda) \cdot \mu_{\min}^\pi} + \frac{r_n}{L \cdot \mu_{\min}^\pi}} \end{aligned}$$

□