How Transformers Get Rich: Training Dynamics Analysis

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2025

Abstract

Transformers have demonstrated exceptional in-context learning capabilities, yet the theoretical understanding of the underlying mechanisms remains limited. A recent work [15] identified a "rich" in-context mechanism known as induction head, contrasting with "lazy" n-gram models that overlook long-range dependencies. In this work, we provide *dynamics* analysis of how transformers learn from lazy to rich mechanism. Specifically, we study the training dynamics on a synthetic mixed target, composed of a 4-gram and an in-context 2-gram component. This controlled setting allows us to precisely characterize the entire training process and uncover an *abrupt transition* from lazy (4-gram) to rich (induction head) mechanisms as training progresses. The theoretical insights are validated experimentally in both synthetic and real-world settings.

1. Introduction

Transformer, introduced by Vaswani et al. [30], have achieved remarkable success across various domains, including natural language processing, computer vision, and scientific computing. An emergent observation is that transformers, trained on trillions of tokens, can perform (few-shot) in-context learning (ICL), which makes prediction based on the contextual information without needing model retraining [8]. This ICL ability is widely regarded as crucial for enabling large language models (LLMs) to solve reasoning tasks, representing a key step toward more advanced AI.

To understand how transformers implement ICL, Elhage et al. [15] and Olsson et al. [20] identified a simple yet powerful mechanism known as **induction head**. Specifically, given an input sequence $[\cdots \texttt{ab} \cdots \texttt{a}]$, an induction head predicts $\texttt{b}$ as the next token by leveraging the prior occurrence of the pattern $\texttt{ab}$ in the context, effectively modeling an in-context bi-gram. In contrast, traditional n-gram model [25] (with a small n) utilizes only a limited number of recent tokens to predict the next token, which is context-independent and inevitably overlooks long-range dependence. Based on the extent of context utilization, we categorize n-gram model as a *"lazy" mechanism*, whereas the induction head represents a more **"rich" mechanism**.

The pioneering works by Elhage et al. [15] and Olsson et al. [20] demonstrated that transformers undergo an abrupt phase transition to learning induction heads. A recent empirical study on synthetic datasets replicate this behavior, further showing that 2-gram is always learned prior to induction heads [7]. However, a rigorous theoretical analysis of this learning progression is still lacking. Closing this gap forms our research objective: *Understand how transformers transition from relying on n-gram patterns to employing the induction head mechanism as training progresses.*

(Our contributions). Dynamics analysis: how learning undergoes a sharp transition from n-gram to induction head. We study the learning dynamics of a two-layer transformer without FFNs for a mixed target, composed of a 4-gram and an in-context 2-gram component. This toy setting allows us to capture the entire training process precisely. Specifically, we show that learning progresses through four phases: partial learning of the 4-gram, plateau of induction head learning, emergence of the induction head, and final convergence, showcasing a sharp transition from 4-gram

to induction head. Our analysis identifies two key drivers of the transition: 1) *time-scale separation* due to low- and high-order parameter dependencies in self-attention, and 2) *speed differences* caused by the relative proportions of the two components in the mixed target. Additionally, in our analysis, we introduce a novel Lyapunov function that exploits the unique structure of self-attention, which may be of independent interest.

Finally, we conduct a series of experiments, ranging from simple toy models to real-world natural language training tasks, to validate our theoretical insights.

2. Preliminaries

Notations. We use $\mathrm{sm}(\cdot)$ to denote the softmax function. We use standard big-O notations $\mathcal{O}, \Omega, \Theta$ to hide absolute positive constants, and use $\tilde{\mathcal{O}}, \tilde{\Omega}, \tilde{\Theta}$ to further hide logarithmic constants.

Sequence modeling. Given a sequence of tokens $(x_1, x_2, x_3, \dots)$ with each token lying in $\mathbb{R}^d$, let $X_L = (x_1, x_2, \dots, x_L) \in \mathbb{R}^{d\times L}$. Given $A = (a_1, \cdots a_n) \in \mathbb{R}^{m\times n}$, we denote $(a_s)_{s=i}^{j} = (a_i, \cdots, a_j)$. We consider the next-token prediction task: predict x_{L+1} using $X_L = (x_1, x_2, \dots, x_L)$.

In a n**-gram model** [25], the conditional probability of predicting the next token is given by $p(x_{L+1}|X_L) = p(x_{L+1}|X_{L-n+2:L})$, meaning that the prediction depends only on the most recent $n-1$ tokens. In practice, the value of n is typically small (e.g., 2, 3, or 4), as the computational cost of n-gram models grows exponentially with n. However, n-gram models with small n cannot capture long-range interactions, leading to inferior performance in sequence modeling.

Self-attention in **Transformers** is designed to more efficiently capture long-range dependencies in sequence modeling [30]. Let SA represent the H-head self-attention operation. Specifically, when applied to a sequence $Z = (z_1, \cdots, z_L) \in \mathbb{R}^{D\times L}$, SA operates it as $\mathsf{SA}(Z) = W_O \sum_{h=1}^{H} \mathsf{SA}^{(h)}(Z)$, where $\mathsf{SA}^{(h)}(Z) = \Big(W_V^{(h)} Z\Big)\mathrm{softmax}\Big(\Big\langle W_Q^{(h)} Z, W_K^{(h)} Z\Big\rangle + R^{(h)}\Big)$, where $W_Q^{(h)}, W_K^{(h)}, W_V^{(h)}$ correspond to the query, key, value matrices of the h-th head, respectively. $\mathrm{softmax}$ represents taking softmax normalization across columns. $\left\langle W_Q^{(h)} X, W_K^{(h)} X \right\rangle$ is called the dot-product (DP) structure. Furthermore, $R^{(h)} = (R_{i,j}^{(h)}) \in \mathbb{R}^{L\times L}$ denotes the additive relative positional encoding matrix, which satisfies $R_{i,j}^{(h)} = -\infty$ if $i \leqslant j$ for the next-token prediction task.

Relative positional encoding (RPE). Throughout this paper, we focus on the Alibi RPE [21], where $R_{ij}^{(u,h)}$ exhibit a Toeplitz structure, i.e., $R_{ij}^{(u,h)} = \phi(i-j; p^{(u,h)})$ for $i, j \in [L]$. Here, $p^{(u,h)}$'s are learnable parameters and $\phi(\cdot; p)$ has the form: $\phi(z; p) = \begin{cases} -p\cdot(z-1), & \text{if } z \geqslant 1 \\ -\infty, & \text{otherwise} \end{cases}$. We adopt the Alibi only for simplicity and our results can be extended to other additive RPEs, such as [11, 22].

Vanilla Induction Heads. The original induction head [15, 20] is regarded as one of the key mechanisms to implement ICL and reasoning. This induction head suggests that *two-layer multi-head* transformers without FFNs can execute a simple in-context algorithm to predict the next token `b` from a context `[···ab···a]` through retrieval, copying, and pasting, based on in-context bi-gram pairs. We define the vanilla induction head $\mathsf{IH}_2 : \cup_{L\in\mathbb{N}^+} \mathbb{R}^{d\times L} \mapsto \mathbb{R}^d$ as follows:

$$\mathsf{IH}_2(X_L) = \sum_{s=2}^{L-1} x_s \,\mathrm{sm}\Big(\big(x_L^\top W^\star x_{\nu-1}\big)_{\nu=2}^{L-1}\Big)_{\nu=s}. \tag{1}$$

Specifically, IH_2 retrieves in-context information of arbitrary length. It retrieves previous tokens x_{s-1}'s that are similar to the current token x_L based on a dot-product similarity, and then copies and pastes x_{s-1}'s subsequent token x_s as the current prediction x_{L+1}. Note that the magnitude of matrix $W^\star$ controls the sparsity of retrieval, since increasing $\|W^\star\|$ causes the softmax output to concentrate as a delta measure over the preceding tokens.

3. The Transition from Lazy to Rich Mechanisms in Learning Dynamics

3.1. Setups

3.1.1. MIXED TARGET FUNCTION

Mixed target function. Let the input sequence be $X = (x_1, \cdots, x_L) \in \mathbb{R}^{1\times L}$. Our mixed target function $f^\star$ contains both a 4-gram component $f^\star_{\mathsf{G}_4}$ and an in-context 2-gram component $f^\star_{\mathsf{IH}_2}$:

$$f^\star(X) := \left(\frac{\alpha^\star}{1+\alpha^\star} f^\star_{\mathsf{G}_4}(X), \frac{1}{1+\alpha^\star} f^\star_{\mathsf{IH}_2}(X)\right)^\top \in \mathbb{R}^2, \tag{2}$$

where $\alpha^\star > 0$ represents the relative weight between the two components: $f^\star_{\mathsf{G}_4}(X)$ and $f^\star_{\mathsf{IH}_2}(X)$. Here, $f^\star_{\mathsf{G}_4}$ represents a 4-gram component and $f^\star_{\mathsf{IH}_2}$ is given by the vanilla induction head (1) to represent a type of in-context 2-gram information:

$$f^\star_{\mathsf{G}_4}(X) := x_{L-2}, \quad f^\star_{\mathsf{IH}_2}(X) := \sum_{s=2}^{L-1} x_s \,\mathrm{sm}\Big(\big(x_L w^{\star 2} x_{\nu-1}\big)_{\nu=2}^{L-1}\Big)_{\nu=s}.$$

Note that $f^\star_{\mathsf{G}_4}$ denotes a "simplest" 4-gram target, where the next token is predicted according to the conditional probability $p(z|X) = p(z|x_L, x_{L-1}, x_{L-2}) = \mathbb{I}\{z = x_{L-2}\}$.

Remark 1 *(*The reason for considering 4-gram.*) Our target includes a 4-gram component rather than simpler 2- or 3-gram components. Note that the induction head includes the 2-gram mechanism. Hence we focus on the more challenging 4-gram target to avoid trivializing the learning process.*

3.1.2. TWO-LAYER MULTI-HEAD TRANSFORMER WITH REPARAMETERIZATION

Two-layer multi-head transformer w/o FFNs. We consider a simple two-layer multi-head transformer TF, where the first layer contains a single head $\mathsf{SA}^{(1,1)}$, and the second layer contain two heads $\mathsf{SA}^{(2,1)}, \mathsf{SA}^{(2,2)}$. Given an input sequence $X = (x_1, \cdots, x_L) \in \mathbb{R}^{1\times L}$, it is first embedded as $X^{(0)} := (X^\top, 0^\top) \in \mathbb{R}^{2\times L}$. The model then processes the sequence as follows:

$$X^{(1)} = X^{(0)} + \mathsf{SA}^{(1,1)}(X^{(0)}), \quad \mathsf{TF}(X) = \mathsf{SA}^{(2,1)}(X^{(1)}) + \mathsf{SA}^{(2,2)}(X^{(1)}).$$

Reparameterization. Despite the simplification, the transformer above is still too complicated for dynamics analysis. To overcome this challenge, we adopt the reparametrization trick used in previous works [10, 16, 28]. We reparameterize the model as follows, (see Appendix C.1 for details):

- **First layer.** This layer consists of a single attention head without DP. The only trainable parameter is $p^{(1,1)}$, which governs the RPE component.
- **Second layer.** This layer contains two heads and five trainable parameters. The first head without DP is responsible to fit $f^\star_{\mathsf{G}_4}$ using parameters $p^{(2,1)}, w_V^{(2,1)}$, while the second head without RPE is responsible to fit $f^\star_{\mathsf{IH}_2}$ with parameters $w_V^{(2,2)}, w_K^{(2,2)}, w_Q^{(2,2)}$.

The set of all six trainable parameters across both layers is denoted by θ.

3.1.3. GRADIENT FLOW ON SQUARE LOSS

We consider the Gaussian input and square loss, both of which are commonly used in analyzing transformer dynamics and ICL [3, 16, 32]. The loss is defined as:

$$\mathcal{L}(\theta) = \frac{1}{2}\mathbb{E}_{X\sim\mathcal{N}(0, I_{L\times L})}\left[\|\mathsf{TF}_{-1}(X;\theta) - f^\star(X)\|_2^2\right], \tag{3}$$

To characterize the learning of G_4 and IH_2, we introduce two partial losses: $\mathcal{L}_{\mathsf{G}_4}(\theta) = \frac{1}{2}\mathbb{E}_X\left(\mathsf{TF}_{-1,1}(X;\theta) - f_1^\star(X)\right)^2$, $\mathcal{L}_{\mathsf{IH}_2}(\theta) = \frac{1}{2}\mathbb{E}_X\left(\mathsf{TF}_{-1,2}(X;\theta) - f_2^\star(X)\right)^2$, corresponding to the two dimensions in $\mathsf{TF}_{-1}(X;\theta) - f^\star(X) \in \mathbb{R}^2$, respectively. It follows that $\mathcal{L}(\theta) = \mathcal{L}_{\mathsf{G}_4}(\theta) + \mathcal{L}_{\mathsf{IH}_2}(\theta)$.

Gradient flow (GF). We analyze the GF for minimizing the objective (3):

$$\frac{\mathrm{d}\theta(t)}{\mathrm{d}t} = -\nabla\mathcal{L}(\theta(t)), \quad \text{starting with } \theta(0) = (\sigma_{\text{init}}, \cdots, \sigma_{\text{init}})^\top, \tag{4}$$

where $0 < \sigma_{\text{init}} \ll 1$ is sufficiently small. Note that $\sigma_{\text{init}} \neq 0$ prevents $\nabla\mathcal{L}(\theta(0)) = 0$.

Layerwise training paradigm. We consider a layerwise training paradigm in which, during each stage, only one layer is trained by GF. Specifically,

- **Training Stage I:** In this phase, only the parameter in the first layer, i.e., $p^{(1,1)}$, is trained.
- **Training Stage II:** In this phase,the first layer parameter $p^{(1,1)}$ keeps fixed and only parameters in the second layer are trained: $w_V^{(2,1)}, w_V^{(2,2)}, p^{(2,1)}, w_Q^{(2,2)}, w_K^{(2,2)}$.

This type of layerwise training has been widely used to study the training dynamics of neural networks, including FFN networks [6, 24, 31] and transformers [10, 19, 28].

Lemma 2 (Training Stage I) *For the Training Stage I,* $\lim_{t\to+\infty} p^{(1,1)}(t) = +\infty$.

According to (5), this lemma implies that, at the end of Training Stage I, the first layer captures the preceding token x_{s-1} for each token x_s, i.e., $y_s = x_{s-1}$. This property is crucial for transformers to implement induction heads. The proof of Lemma 2 is deferred to Appendix C.2.

3.2. Training Stage II: Transition from 4-gram to Induction Head

In this section, we analyze the dynamics in Training Stage II. We start from the following lemma:

Lemma 3 (Parameter balance) *In Training Stage II, it holds that* $|w_Q^{(2,2)}(t)|^2 \equiv |w_K^{(2,2)}(t)|^2$.

Lemma 3 is similar to the balance result for homogeneous networks [12], and its proof can be found at the start of Appendix C.3. By this lemma, we can define $w_{KQ}^{(2,2)} := w_Q \equiv w_K$. Additionally, Lemma 2 ensures that $p^{(1,1)} = +\infty$ holds during Stage II. For simplicity, we denote $w_{V_1} := w_V^{(2,1)}, w_{V_2} := w_V^{(2,2)}, p := p^{(2,1)}, w_{KQ} := w_{KQ}^{(2,2)}$. Consequently, the training dynamics are reduced to **four parameters** $\theta = (w_{V_1}, w_{V_2}, p, w_{KQ})$.

where we still denote the set of parameters as θ without introducing ambiguity. It is important to note that the problem remains **highly non-convex** due to the joint optimization of both inner parameters (p, w_{KQ}) and outer parameters (w_{V_1}, w_{V_2}) in the two heads. At this training stage, GF has a **unique fixed point** $w_{V_1} = \frac{\alpha^\star}{1+\alpha^\star}$, $w_{V_2} = \frac{1}{1+\alpha^\star}$, $p = +\infty$, $w_{KQ} = w^\star$, which corresponds to a global minimizer of the objective (3).

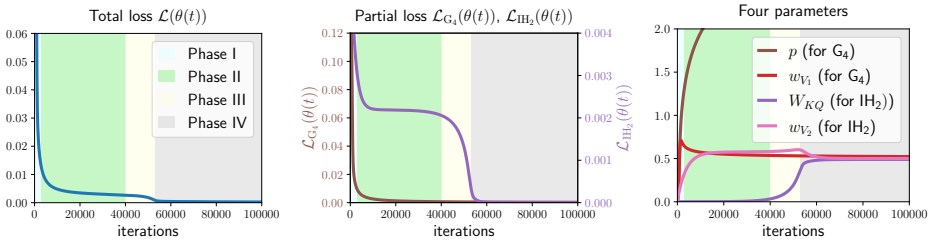

Figure 1: Visualization of the dynamical behavior of Training Stage II with total loss, partial loss, and the parameter evolution. Here, $\alpha^\star = 1, w^\star = 0.49, \sigma_{\text{init}} = 0.01, L = 40$. The is clearly shown that transformer learns the 4-gram component first and then, starts to learn the induction head mechanism. Notably, the entire dynamics unfold in four distinct phases, consistent with our theoretical results (Theorem 4). For more experimental details, we refer to Appendix B.1.

As shown in Figure 1, a learning transition from the 4-gram mechanism to the induction head mechanism does occur. Moreover, the learning process exhibits four-phase dynamics. The next theorem provides a precise characterization of the four phases, which proof is provided in Appendix C.3.

Theorem 4 (Learning transition and 4-phase dynamics) *Let* $\alpha^\star = \Omega(1)$ *and* $w^\star = \mathcal{O}(1)$*, and we consider the regime of small initialization* $(0 < \sigma_{\text{init}} \ll 1)$ *and long input sequences* $(L \gg 1)$*. Then we have the following results:*

- **Phase I (partial learning).** *In this phase, most of the 4-gram component in the mixed target is learned, while a considerable number of induction head component have not yet been learned. Specifically, let* $T_{\text{I}} = \mathcal{O}(1)$*, then we have the following estimates:*

$$\mathcal{L}_{\mathsf{G}_4}(\theta(T_{\text{I}})) \leqslant 0.01 \cdot \mathcal{L}_{\mathsf{G}_4}(\theta(0)),\ \mathcal{L}_{\mathsf{IH}_2}(\theta(T_{\text{I}})) \geqslant 0.99 \cdot \mathcal{L}_{\mathsf{IH}_2}(\theta(0)).$$

- **Phase II (plateau) + Phase III (emergence).** *In these two phases, the learning of the induction head first gets stuck in a plateau for T_{II} time, then is learned suddenly. Specifically, denoted by an observation time $T_o = \Theta(L)$, we have the following tight estimate of the duration:*

$$T_{\mathrm{II}} := \inf\left\{t > T_o : \mathcal{L}_{\mathsf{IH}_2}(\theta(t)) \leqslant 0.99 \cdot \mathcal{L}_{\mathsf{IH}_2}(\theta(T_o))\right\} = \Theta\left((\alpha^\star+1)^2 L \log(1/\sigma_{\mathrm{init}})/w^{\star 2}\right);$$
$$T_{\mathrm{III}} := \inf\left\{t > T_o : \mathcal{L}_{\mathsf{IH}_2}(\theta(t)) \leqslant 0.01 \cdot \mathcal{L}_{\mathsf{IH}_2}(\theta(T_o))\right\} = \Theta\left((\alpha^\star+1)^2 L \log(1/\sigma_{\mathrm{init}})/w^{\star 2}\right).$$

During these phases, the parameter w_{KQ} (for learning $w^\star$ in IH_2) increases exponentially:

$$w_{KQ}(t) = \sigma_{\mathrm{init}} \exp\left(\Theta\left(\frac{w^{\star 2} t}{(1+\alpha^\star)^2 L}\right)\right),\ t < T_{\mathrm{III}}.$$

- **Phase IV (convergence).** *In this phase, the loss converges toward zero. Specifically, the following convergence rates hold for all $t > T_{\mathrm{III}}$:*

$$\mathcal{L}_{\mathsf{G}_4}(\theta(t)) = \mathcal{O}\left(\frac{1}{t}\right),\ \mathcal{L}_{\mathsf{IH}_2}(\theta(t)) = \mathcal{O}\left(\exp\left(-\Omega\left(\frac{w^{\star 2} t}{(1+\alpha^\star)^2 L}\right)\right)\right),$$

and $\mathcal{L}(\theta(t)) = \mathcal{L}_{\mathsf{G}_4}(\theta(t)) + \mathcal{L}_{\mathsf{IH}_2}(\theta(t))$.

By this theorem, the 4-gram mechanism is first learned, taking time T_{I}. Then, the learning of the induction head mechanism enters a plateau, taking time T_{II}, followed by a sudden emergence of learning, taking time $T_{\mathrm{III}} - T_{\mathrm{II}}$. Finally, the loss for both components converges to zero.

The clear learning transition. When any one of $L, \alpha^\star, 1/\sigma_{\mathrm{init}}, 1/w^\star$ is sufficiently large, Phase II lasts for $T_{\mathrm{II}} \gg 1$. During this phase, the 4-gram component has been learned well but the induction head component remains underdeveloped, demonstrating a distinct learning transition. Moreover, Theorem 4 and its proof reveal two key factors that drive this transition:

- **Time-scale separation due to high- and low-order parameter dependence in self attention**. The learning of DP and RPE components differ in their parameter dependencies. DP component exhibits a quadratic dependence on the parameter w_{KQ}, while RPE component shows linear dependence on the parameter p. With small initialization $\sigma_{\mathrm{init}} \ll 1$, a clear time-scale separation emerges: $|\dot{w}_{KQ}| \sim w_{KQ} \ll 1$ (DP, slow dynamics) and $|\dot{p}| \sim 1$ (RPE, fast dynamics). Consequently, the induction head (fitted by DP) is learned much slower than the 4-gram component (fitted by RPE). This time-scale separation accounts for the term $\log(1/\epsilon_{\mathrm{init}})$ in the plateau T_{II}.
- **Speed difference due to component proportions in the mixed target.** The 4-gram target component and the induction-head component have differing proportions in the mixed target. A simple calculation shows: $\mathcal{L}_{\mathsf{G}_4}(0) \sim \alpha^{\star 2}/(1+\alpha^\star)^2$; If $w^\star = \mathcal{O}(1)$, then $\mathcal{L}_{\mathsf{IH}_2}(0) \sim 1/[(1+\alpha^\star)^2 L]$. Notably, $\mathcal{L}_{\mathsf{IH}_2}(0)$ is significantly smaller than $\mathcal{L}_{\mathsf{G}_4}(0)$. This proportion disparity accounts for the $(1+\alpha^\star)^2 L$ term in the plateau time T_{II}.

Proof idea. We highlight that our fine-grained analysis of entire learning process is guided by two key observations: 1) the dynamics of the two heads can be decoupled; 2) there exist a distinct transition point in the dynamics of each head, as shown in Figure 1 (right). These insights lead us to divide the analysis of each head into two phases: a monotonic phase and a convergence phase. Particularly, for the convergence phase, we introduce a novel Lyapunov function that leverages the unique dynamical structure of self-attention. This Lyapunov function may be of independent interest and offers potential for studying broader issues in self-attention dynamics.

Further experiments. We conduct additional experiments to validate our theoretical insights into the training dynamics and learning transition across a wider range of scenarios. These includes using data distribution (Figure 3) and optimization algorithms (Figure 4) in high-dimensional settings, as well as training real-world transformers on natural language datasets (Figure 2).

References

[1] Emmanuel Abbe, Samy Bengio, Enric Boix-Adsera, Etai Littwin, and Joshua Susskind. Transformers learn through gradual rank increase. *Advances in Neural Information Processing Systems*, 36, 2024.

[2] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Ekin Akyürek, Dale Schuurmans, Jacob Andreas, Tengyu Ma, and Denny Zhou. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

[4] Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. *Advances in Neural Information Processing Systems*, 36: 48314–48362, 2023.

[5] Yu Bai, Fan Chen, Huan Wang, Caiming Xiong, and Song Mei. Transformers as statisticians: Provable in-context learning with in-context algorithm selection. *arXiv preprint arXiv:2306.04637*, 2023.

[6] Alberto Bietti, Joan Bruna, and Loucas Pillaud-Vivien. On learning gaussian multi-index models with gradient flow. *arXiv preprint arXiv:2310.19793*, 2023.

[7] Alberto Bietti, Vivien Cabannes, Diane Bouchacourt, Herve Jegou, and Leon Bottou. Birth of a transformer: A memory viewpoint. *Advances in Neural Information Processing Systems*, 36, 2024.

[8] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[9] Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024.

[10] Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Unveiling induction heads: Provable training dynamics and feature learning in transformers. *arXiv preprint arXiv:2409.10559*, 2024.

[11] Ta-Chung Chi, Ting-Han Fan, Peter J Ramadge, and Alexander Rudnicky. Kerple: Kernelized relative positional embedding for length extrapolation. *Advances in Neural Information Processing Systems*, 35:8386–8399, 2022.

[12] Simon S Du, Wei Hu, and Jason D Lee. Algorithmic regularization in learning deep homogeneous models: Layers are automatically balanced. *Advances in neural information processing systems*, 31, 2018.

[13] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022.

[14] Benjamin L Edelman, Ezra Edelman, Surbhi Goel, Eran Malach, and Nikolaos Tsilivis. The evolution of statistical induction heads: In-context learning markov chains. *arXiv preprint arXiv:2402.11004*, 2024.

[15] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. https://transformer-circuits.pub/2021/framework/index.html.

[16] Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*, 2023.

[17] Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. Training nonlinear transformers for efficient in-context learning: A theoretical learning and generalization analysis. *arXiv preprint arXiv:2402.15607*, 2024.

[18] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

[19] Eshaan Nichani, Alex Damian, and Jason D Lee. How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*, 2024.

[20] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

[21] Ofir Press, Noah A Smith, and Mike Lewis. Train short, test long: Attention with linear biases enables input length extrapolation. *International Conference on Learning Representations*, 2022.

[22] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

[23] Gautam Reddy. The mechanistic basis of data dependence and abrupt learning in an in-context classification task. *International Conference on Learning Representations*, 2024.

[24] Itay Safran and Jason Lee. Optimization-based separations for neural networks. In *Conference on Learning Theory*, pages 3–64. PMLR, 2022.

[25] Claude Elwood Shannon. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423, 1948.

[26] Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023.

[27] Christos Thrampoulidis. Implicit bias of next-token prediction. *arXiv preprint arXiv:2402.18551*, 2024.

[28] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon S Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in Neural Information Processing Systems*, 36:71911–71947, 2023.

[29] Bhavya Vasudeva, Puneesh Deora, and Christos Thrampoulidis. Implicit bias and fast convergence rates for self-attention. *arXiv preprint arXiv:2402.05738*, 2024.

[30] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[31] Zihao Wang, Eshaan Nichani, and Jason D Lee. Learning hierarchical polynomials with three-layer neural networks. *arXiv preprint arXiv:2311.13774*, 2023.

[32] Zixuan Wang, Stanley Wei, Daniel Hsu, and Jason D Lee. Transformers provably learn sparse token selection while fully-connected nets cannot. *arXiv preprint arXiv:2406.06893*, 2024.

[33] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *arXiv preprint arXiv:2306.09927*, 2023.

Appendix A. Related Works

Empirical observations of induction head. The induction head mechanism was first identified by Elhage et al. [15] in studying how two-layer transformers perform language modeling. Subsequently, Olsson et al. [20] conducted a more systematic investigation, revealing two key findings: 1) induction head emerges abruptly during training, and 2) induction head plays a critical role in the development of in-context learning capabilities. To obtain a fine-grained understanding of how induction head emerges during training, recent studies have developed several synthetic settings [7, 14, 23]. Particularly, Bietti et al. [7] successfully reproduced the fast learning of (global) bigrams and the slower development of induction head. Despite these efforts, a comprehensive theoretical understanding of how the induction head operates in two-layer transformers and how it is learned during training remains elusive.

Training dynamics of transformers. To gain insights into the dynamics of training transformers, several studies have analyzed simplified transformers on toy tasks. These tasks include learning distinct/common tokens [28], leaning balance/inblanced features [16], linear regression task [2, 33], multi-task linear regression [9], binary classification [17], transformer with diagonal weights [1], learning causal structure [19], sparse token selection task [32], and learning n-gram Markov chain [10]. Additionally, studies such as those by Ataee Tarzanagh et al. [4], Tarzanagh et al. [26] and Vasudeva et al. [29] have analyzed scenarios where transformers converge to max-margin solutions. Furthermore, Thrampoulidis [27] has examined the implicit bias of next-token prediction. Among these works, the most closely related to ours are Nichani et al. [19] and Chen et al. [10], which proved that two-layer transformers can converge to induction head solutions. In this work, we explore a setting where the target is a mixture of 4-gram and induction head. We show that two-layer transformers can effectively converge to this mixed target and provide a precise description of the learning process associated with each component. Importantly, we are able to capture the *abrupt transition* from learning 4-gram patterns to mastering the induction head mechanism—a critical phase in the learning of induction heads, as highlighted in the seminal works [15, 20].

Now we discuss the relationship between our work and two closely related studies [7, 14].

Comparison with Bietti et al. [7].

Study objective: While Bietti et al. [7] examines the transition from 2-gram to induction head, our work focuses on the transition from 4-gram to induction head.

study methods: Bietti et al. [7] conducts extensive experiments supported by partial theoretical properties but does not fully characterize the training dynamics theoretically. In contrast, our study provides **a precise theoretical analysis of the entire training process** in a toy model, uncovering the sharp transition from 4-gram to induction head.

Main insights: Bietti et al. [7] emphasizes the the role of weight matrices as associative memories and the impact of data distributional properties. Our analysis, on the other hand, identifies two primary drivers of the transition: (1) the time-scale separation due to low- and high-order parameter dependencies in self-attention; (2) the speed differences caused by the relative proportions of the two components in the mixed target.

Comparison with Edelman et al. [14]. Edelman et al. [14] focuses on the transition from unigram to bi-gram mechanisms in Markov Chain data. In contrast, our study investigates the transition from 4-gram to in-context 2-gram mechanisms (induction head). Additionally, we theoretically identify two primary drivers of the transition: (1) the time-scale separation due to low- and high-order

parameter dependencies in self-attention; (2) the speed differences caused by the relative proportions of the two components in the mixed target.

Appendix B. Experiments

1. **Standard transformers on real-world natural language dataset.**

Setup. We train a two-layer two-head **standard transformer** with Alibi RPE (without any simplification) on the **wikitext-2** dataset, a natural language dataset [18]. The transformer has an embedding dimension $D = 128$ and FFN width $W = 512$. For this dataset, the input dimension is $d = 33278$. We use a context length $L = 200$ and batch size $B = 32$. The parameters are initialized with the scale 0.01. The model is trained for 1,500 epochs on 1 H100, using cross-entropy loss and SGD with learning rate 0.1, and the initialization scale is 0.01. It is important to note that **both layers are trained simultaneously**. The results are presented in Figure 2.

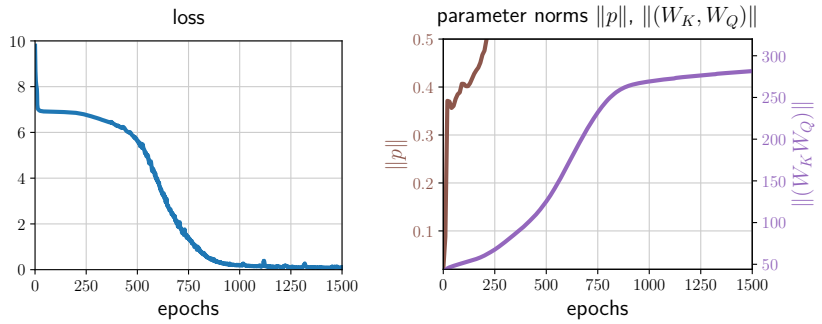

Figure 2: The loss and parameters for the experiment training a two-layer two-head **standard transformer** (without any simplification) on the **wikitext-2** dataset [18]. Here, $\|p\|$ and $\|(W_K, W_Q)\|$ denote the Frobenius norms of all positional encoding parameters and all W_K, W_Q parameters across layers and heads, respectively, The results show that: the loss exhibits a clear plateau; position encoding p's are learned first; and the dot-product structure W_K, W_Q are learned slowly at the beginning, resembling an exponential increase; additionally, as W_K, W_Q are learned, the loss escapes that plateau. These findings closely resemble the behavior observed in our toy model (Figure 1). This experiment provides further support for our theoretical insights regarding the **time-scale separation** between the learning of positional encoding and the dot-product structure.

2. **Discrete token distribution in toy setting.**

Setup. We modified the Gaussian input distribution used in the setup for Figure 1 to a boolean input distribution, where each input token, where each input token $x_i \overset{iid}{\sim} \mathrm{Unif}(\{\pm 1\})$ for $i \in [L]$, All other experimental setups remain the same as in the setup for Figure 1. The training dynamics of Stage (ii) are presented in Figure 3. We can see clearly that the dynamical behavior of the learning process is nearly the same as the one observed for Gaussian inputs in Figure 1.

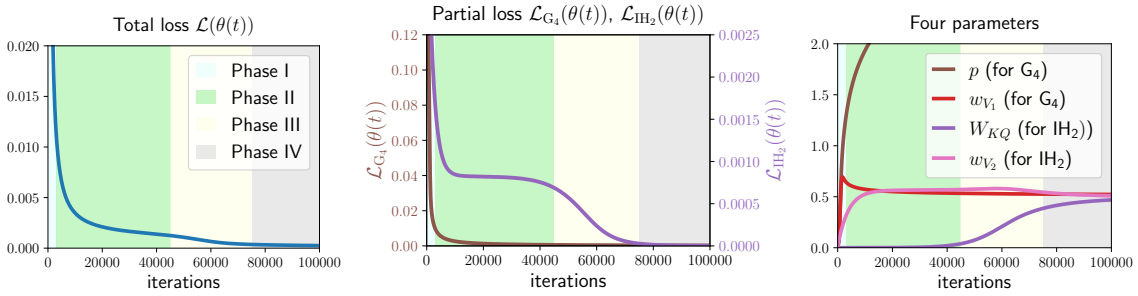

Figure 3: Visualization of the total loss, partial loss, and the parameter dynamics, for the experiment on **discrete token distribution** (Boolean, $X \sim \mathrm{Unif}(\{\pm 1\}^L)$) in our toy setting with $\alpha^\star = 1, w^\star = 0.49, \sigma_{\text{init}} = 0.01, L = 40$. The figure clearly shows that transformer learns the 4-gram component first and then, starts to learn the induction head mechanism. Notably, the entire dynamics exhibit four phases. These results are **extremely similar** to that observed with Gaussian inputs, as shown in Figure 1.

3. **Adam in high-dimensional toy setting.**

Setup. We modified the setup for Figure 1 to employ a high-dimensional model ($D = 100$). Specifically, the target is $w^\star = 0.49 I_D/D$, the dot-produce parameters are $W_K, W_Q \in \mathbb{R}^D$, initialized such that $\|W_K\|_F, \|W_Q\|_F = \sigma_{\text{init}}$. Additionally, for the Adam optimizer, we use learning rate `5e-4`. All other experimental setups remain the same as in the setup for Figure 1.

The training dynamics are depicted in Figure 4, where, for comparison, results using GD are also presented. In both scenarios, the learning process begins with the 4-gram pattern, followed by a gradual learning phase of the induction head mechanism. Notably, within the given number of iterations, GD remains stuck in the plateau, whereas Adam successfully escapes that plateau.

B.1. Experimental details for Figure 1

In line with our theoretical setting, we examine a simplified two-layer transformer, as described in Alibi RPE. Specifically, the first layer only contains RPE and the second layer consists of two heads: one uses only RPE and the other employs only dot-product structure. The target function is specified by (2) with $\alpha^\star = 1, w^\star = 0.49, \sigma_{\text{init}} = 0.01, L = 40$, and the distribution of each token is Gaussian, i.e., $x_i \overset{iid}{\sim} \mathcal{N}(0, 1)$ for $i \in [L]$. Training is conducted by minimizing the squared loss (3) using online SGD with learning rate 0.1 and batch size $B = 1,000$. Following our theoretical analysis, the two layers are trained sequentially:

- Training Stage I: only the first layer is trained for 100,000 iterations;
- Training Stage II: Subsequently, only the second layer undergoes training for another 100,000 iterations.

The dynamical behavior of the Training Stage II is visualized in Figure 1.

Compute resources. Real-world experiments on wikitext-2 are conducted on 1 A100 GPU, while other synthetic experiments are conducted on CPU.

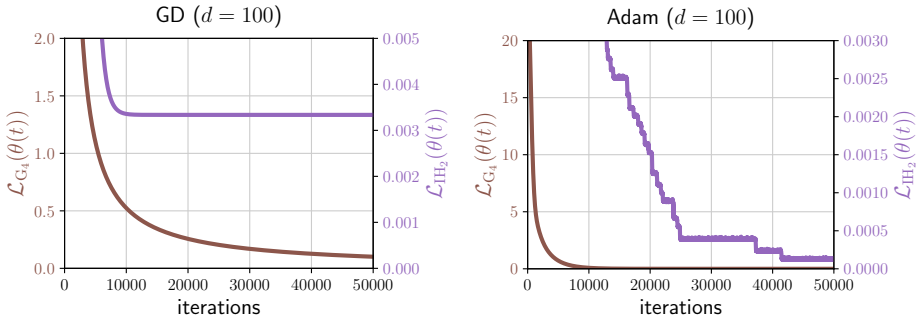

Figure 4: Partial loss for the experiment comparing **GD v.s. Adam optimizer** in high-dimensional settings ($D = 100$). In this setting, a larger D increases the difficulty of the transition from the lazy regime (learning 4-gram) to the rich regime (learning induction head). The results indicate that: (1) GD learns the 4-gram component first but becomes stuck in a plateau when learning induction head; (2) Adam, while eventually transitioning from the lazy regime (learning 4-gram) to the rich regime (learning induction head), experiences a **challenging** transition characterized by **multiple plateaus** during learning induction heads. This finding closely resembles the dynamics for GD.

Appendix C. Proofs in Section 3

C.1. Reparameterization

Despite the simplification, the transformer above is still too complicated for dynamics analysis. To overcome this challenge, we adopt the reparametrization trick used in previous works [10, 16, 28]. Specifically, to express vanilla induction head, *the first layer does not require DP, and the second layer does not require RPE*. Moreover, to express the 4-gram component $f^\star_{\mathsf{G}_4}$, we only need an additional head without DP in the second layer. Therefore, we can reparameterize the model as follows:

- **The first layer.** This layer has only one trainable parameter $p^{(1,1)}$. In the unique head $\mathsf{SA}^{(1,1)}$, DP is removed by setting $W_Q^{(1,1)} = W_K^{(1,1)} = 0$, and we let $W_V^{(1,1)} = \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix}$. The output sequence of this layer given by $X^{(1)} = X^{(0)} + \mathsf{SA}^{(1,1)}(X^{(0)}) = \begin{pmatrix} x_1, \cdots, x_L \\ y_1, \cdots, y_L \end{pmatrix}$, where

$$y_s = \sum_{\tau=1}^{s-1} x_\tau \ \mathrm{sm}\left(\left(-p^{(1,1)}(s-1-\nu) \right)_{\nu=1}^{s-1} \right)_{\nu=\tau} \tag{5}$$

  for $s \in [L]$, where $p^{(1,1)}$, used in RPE, is the unique trainable parameter in this layer.

- **The second layer.** This layer has 5 trainable parameters: $w_V^{(2,1)}, w_V^{(2,2)}, p^{(2,1)}, w_K^{(2,2)}, w_Q^{(2,2)}$ for parametrizing the two heads. The first head $\mathsf{SA}^{(2,1)}$ without DP is responsible to fit $f^\star_{\mathsf{G}_4}$, while the second head $\mathsf{SA}^{(2,2)}$ without RPE is responsible to fit $f^\star_{\mathsf{IH}_2}$. Specifically, $W_Q^{(2,1)} =$

$W_K^{(2,1)} = 0, W_V^{(2,1)} = \begin{pmatrix} 0 & w_V^{(2,1)} \\ 0 & 0 \end{pmatrix}, p^{(2,2)} = 0, W_V^{(2,2)} = \begin{pmatrix} w_V^{(2,2)} & 0 \\ 0 & 0 \end{pmatrix}$. Then the second layer processes $X^{(1)}$ and outputs the last token:

$$\mathsf{TF}_{-1}(X;\theta) = \Big(\sum_{s=2}^{L-2} w_V^{(2,1)} y_s \pi_s, \sum_{s=2}^{L-2} w_V^{(2,2)} x_s \rho_s \Big)^\top, \tag{6}$$

$$\pi_s = \mathrm{sm}\left(\big(-p^{(2,1)}(L-1-\nu) \big)_{\nu=2}^{L-2} \right)_{\nu=s}, \ \rho_s = \mathrm{sm}\left(\big(x_L w_Q^{(2,2)} w_K^{(2,2)} x_{\nu-1} \big)_{\nu=2}^{L-2} \right)_{\nu=s},$$

where y_s is given by Eq. (5). $p^{(2,1)}, w_V^{(2,1)}$ are trainable parameters in $\mathsf{SA}^{(2,1)}$, while $w_Q^{(2,2)}, w_K^{(2,2)}, w_V^{(2,2)}$ are trainable parameters in $\mathsf{SA}^{(2,2)}$.

The set of all six trainable parameters across both layers is denoted by θ.

C.2. Optimization Dynamics in Training Stage I

In this subsection we focus on training the first layer of Transformer model to capture the token ahead. For simplicity, we introduce some notations:

$$\tilde{p} := p^{(1,1)}, \quad p := p^{(2,1)}, \quad g := w_V^{(2,1)}, \quad h := w_V^{(2,2)}, \quad w_K := w_K^{(2,2)}, \quad w_Q := w_Q^{(2,2)},$$

and denote the initialization of each parameter as $\tilde{p}(0), p(0), g(0), w_Q(0), w_K(0), h(0)$ respectively.

We initialize $p(0), w_k(0), w_Q(0) = 0$ while the other parameters are all initialized at σ_{init}. In this training stage, we only train $\tilde{p}$. And our goal, **the proof of Lemma 2** can be deduced from which, is to prove:

$$\lim_{t \to +\infty} \tilde{p}(t) = +\infty.$$

In this stage, the s-th output token of the first layer is represented as

$$\begin{pmatrix} x_s \\ (x_\tau)_{\tau=1}^{s-1} \ \mathrm{softmax}\left(\big(-\tilde{p}(s-1-\tau) \big)_{\tau=1}^{s-1} \right)^\top \end{pmatrix},$$

and the target function and output of transformer are as follows

$$f^*(X) = \begin{pmatrix} \frac{\alpha^\star}{1+\alpha^\star} x_{L-2} \\ \frac{1}{1+\alpha^\star} (x_s)_{s=2}^{L-1} \ \mathrm{softmax}\Big(\big(x_L w^{\star 2} x_{s-1} \big)_{s=2}^{L-1} \Big)^\top \end{pmatrix},$$

$$\begin{aligned} f_\theta(X) &= \begin{pmatrix} g(0) \left((x_\tau)_{\tau=1}^{s-1} \mathrm{softmax}\left((-\tilde{p}(s-1-\tau))_{\tau=1}^{s-1} \right)^\top \right)_{s=2}^{L-1} \mathrm{softmax}\Big((-p(0)(L-1-s))_{s=2}^{L-1} \Big)^\top \\ h(0) (x_s)_{s=2}^{L-2} \mathrm{softmax}\left(\Big(w_K(0) w_Q(0) x_L \cdot (x_\tau)_{\tau=1}^{s-1} \ \mathrm{softmax}\left((-\tilde{p}(s-1-\tau))_{\tau=1}^{s-1} \right)^\top \Big)_{s=2}^{L-2} \right)^\top \end{pmatrix} \\ &= \begin{pmatrix} g(0) \frac{1}{L-2} \sum_{\tau=1}^{L-2} \left(\sum_{s=\tau+1}^{L-1} \mathrm{softmax}\Big((-\tilde{p}(s-1-t))_{t=1}^{s-1} \Big) \right)_{t=\tau} x_\tau \\ h(0) \frac{1}{L-2} \sum_{s=2}^{L-2} x_s \end{pmatrix}. \end{aligned}$$

Since we only focus on $\tilde{p}$ and the other parameters remain the initialization value, the loss function can be simplified as

$$\mathcal{L}(\theta) = \mathop{\mathbb{E}}_{X\sim\mathbb{N}(0,1)^L}\left[\frac{\alpha^{\star 2}}{(1+\alpha^\star)^2}x_{L-2}^2 + \frac{g(0)^2}{(L-2)^2}\sum_{\tau=1}^{L-2}\left(\sum_{s=\tau+1}^{L-1}\mathrm{softmax}\Big((-\tilde{p}(s-1-t))_{t=1}^{s-1}\Big)_{t=\tau}\right)^2 x_\tau^2 \right.$$
$$\left. + \frac{2g(0)}{L-2}\frac{\alpha^\star}{1+\alpha^\star}\mathrm{softmax}\Big((-p(0)(L-1-s))_{s=2}^{L-1}\Big)_{s=L-1}x_{L-2}^2\right] + C(w^\star,\alpha^\star,w(0),h(0))$$

where the second term $C(w^\star,\alpha^\star,w(0),h(0))$ is a constant depends on $w^\star,\alpha^\star,w(0)$ and $h(0)$, produced by calculating the error of the second head, i.e., loss of induction head, while the first term is 4-gram loss.

We first define several functions that will be useful for calculation in this stage and the second one:

Function I. This function is purely defined for the calculation of $\frac{\mathrm{d}p}{\mathrm{d}t}$. Denoted by $q(\tilde{p}) := \sum_{\tau=1}^{L-2}\left(\sum_{s=\tau+1}^{L-1}\frac{e^{\tilde{p}(s-1-\tau)}}{\sum_{k=0}^{s-2}e^{-\tilde{p}k}}\right)^2$, we first prove $\frac{\mathrm{d}q}{\mathrm{d}\tilde{p}} \leqslant 0$.

$$\begin{aligned}
q(\tilde{p}) &:= \sum_{\tau=1}^{L-2}\left(\sum_{s=\tau+1}^{L-1}\frac{e^{\tilde{p}(s-1-\tau)}}{\sum_{k=0}^{s-2}e^{-\tilde{p}k}}\right)^2 \\
&= \sum_{\tau=1}^{L-2}\left(\sum_{s=\tau+1}^{L-1}\frac{e^{-\tilde{p}(s-1-\tau)}}{1-e^{-\tilde{p}(s-1)}}(1-e^{-\tilde{p}})\right)^2 \\
&= (1-e^{-\tilde{p}})^2\sum_{\tau=1}^{L-2}\left(\sum_{s=\tau+1}^{L-1}\frac{e^{-\tilde{p}(s-1-\tau)}}{1-e^{-\tilde{p}(s-1)}}\right)^2 \\
&= (1-e^{-\tilde{p}})^2\sum_{\tau=1}^{L-2}e^{2\tilde{p}\tau}\left(\sum_{s=\tau+1}^{L-1}\frac{e^{-\tilde{p}(s-1)}}{1-e^{-\tilde{p}(s-1)}}\right)^2 \\
&= (1-e^{-\tilde{p}})^2\sum_{\tau=1}^{L-2}e^{2\tilde{p}\tau}\left(\sum_{s=\tau+1}^{L-1}\frac{1}{e^{\tilde{p}(s-1)}-1}\right)^2
\end{aligned}$$

Then we take its derivative of $\tilde{p}$

$$\begin{aligned}
\frac{\mathrm{d}q}{\mathrm{d}\tilde{p}} &= 2(1-e^{-\tilde{p}})e^{-\tilde{p}}\sum_{\tau=1}^{L-2}e^{2\tilde{p}\tau}\left(\sum_{s=\tau+1}^{L-1}\frac{1}{e^{\tilde{p}(s-1)}-1}\right)^2 \\
&+ (1-e^{-\tilde{p}})^2\sum_{\tau=1}^{L-2}2\tau e^{2\tilde{p}\tau}\left(\sum_{s=\tau+1}^{L-1}\frac{1}{e^{\tilde{p}(s-1)}-1}\right)^2 \\
&+ (1-e^{-\tilde{p}})^2\sum_{\tau=1}^{L-2}2e^{2\tilde{p}\tau}\left(\sum_{s=\tau+1}^{L-1}\frac{1}{e^{\tilde{p}(s-1)}-1}\right)\left(\sum_{s=\tau+1}^{L-1}\frac{-(s-1)e^{\tilde{p}(s-1)}}{(e^{\tilde{p}(s-1)}-1)^2}\right) \\
&= 2(1-e^{-\tilde{p}})\sum_{\tau=1}^{L-2}e^{2\tilde{p}\tau}\left(\sum_{s=\tau+1}^{L-1}\frac{1}{e^{\tilde{p}(s-1)}-1}\right)\left(\sum_{s=\tau+1}^{L-1}\frac{e^{-\tilde{p}}+\tau(1-e^{-\tilde{p}})}{e^{\tilde{p}(s-1)}-1} - \frac{(s-1)e^{\tilde{p}(s-1)}}{(e^{\tilde{p}(s-1)}-1)^2}\right)
\end{aligned}$$

$\frac{\mathrm{d}q}{\mathrm{d}\tilde{p}}$'s last factor can be formed as

$$\begin{aligned}&\frac{\left(\tau-(\tau-1)e^{-\tilde{p}}\right)\left(e^{\tilde{p}(s-1)}-1\right)-(s-1)e^{\tilde{p}(s-1)}}{e^{\tilde{p}(s-1)}-1)^2}\\=&\frac{(\tau+1-s)t^{s-1}-(\tau-1)t^{s-2}-\tau+\frac{\tau-1}{t}}{e^{\tilde{p}(s-1)}-1)^2}\end{aligned}$$

where $t=e^{-\tilde{p}}\geqslant 1$. Since $s\geqslant\tau+1$, $\frac{\mathrm{d}q}{\mathrm{d}\tilde{p}}\leqslant 0$.

Function II. For simplicity, we define $M(p)$ and its derivative $m(p)$:

$$M(p):=\sum_{s=2}^{L-1}\exp(-p(L-1-s))=\sum_{s=0}^{L-3}\exp -ps=\frac{1-e^{-p(L-2)}}{1-e^{-p}},$$

$$m(p):=\sum_{s=1}^{l-3}s\exp(-ps)=\frac{e^{-p}-(L-2)e^{-p(L-2)}+(L-3)e^{-p(L-1)}}{(1-e^{-p})^2}.$$

Function III. The third function is derivative of softmax. By straightfoward calculation, we obtain:

$$\frac{\mathrm{d}}{\mathrm{d}p}\mathrm{softmax}\Big((-p(L-1-t))_{t=2}^{L-1}\Big)_{t=L-1-s}=\frac{\mathrm{d}}{\mathrm{d}p}\frac{\exp(-ps)}{\sum_{\tau=0}^{L-3}\exp(-p\tau)}=\frac{-s\exp(-ps)M(p)+\exp(-ps)m(p)}{M(p)^2}.$$

Through the quantities and their properties above, we obtain the dynamic of $\tilde{p}$

$$\begin{aligned}\frac{\mathrm{d}\tilde{p}}{\mathrm{d}t}&=-\frac{g(0)^2}{(L-2)^2}q'(\tilde{p})+\frac{2\alpha^\star g(0)}{(1+\alpha^\star)(L-2)}\frac{m(p)}{M(p)^2}\\&\geqslant\frac{2\alpha^\star g(0)}{(1+\alpha^\star)(L-2)}e^{-\tilde{p}},\end{aligned}$$

which implies:

$$\lim_{t\to+\infty}\tilde{p}(t)=+\infty.$$

C.3. Optimization Dynamics in Training Stage II

In this training stage, the first layer is already capable of capturing the token ahead i.e. $y_s=x_{s-1}$. And we train the parameters w_{V_1},w_{V_2},p,w_{KQ} in the second layer.

We start from proving the parameter balance lemma:

Lemma 5 (Restate of Lemma 3) *In Training Stage II, it holds that* $w_Q^{(2,2)^2}(t)\equiv w_K^{(2,2)^2}(t)$.

Proof Notice that

$$\frac{\mathrm{d}}{2\mathrm{d}t}\left({w_Q^{(2,2)}}^2(t)-{w_K^{(2,2)}}^2(t)\right)=-w_Q^{(2,2)}\frac{\partial\mathcal{L}}{\partial w_Q^{(2,2)}}+w_K^{(2,2)}\frac{\partial\mathcal{L}}{\partial w_K^{(2,2)}}$$
$$=-w_Q^{(2,2)}w_K^{(2,2)}\frac{\partial\mathcal{L}}{\partial\left(w_Q^{(2,2)}w_K^{(2,2)}\right)}+w_K^{(2,2)}w_Q^{(2,2)}\frac{\partial\mathcal{L}}{\partial\left(w_Q^{(2,2)}w_K^{(2,2)}\right)}\equiv 0.$$

Thus, we have:

$${w_Q^{(2,2)}}^2(t)-{w_K^{(2,2)}}^2(t)\equiv{w_Q^{(2,2)}}^2(0)-{w_K^{(2,2)}}^2(0)=0.$$

■

For simplicity, we still use the following notations:

$$p:=p_1,\quad g:=w_{V_1},\quad w:=w_{KQ},\quad h:=w_{V_2}.$$

and notations for initialization $p(0),g(0),w(0),h(0)$. Then the target function and output of Transformer can be formed as follows

$$f^\star(X)=\begin{pmatrix}\frac{\alpha^\star}{1+\alpha^\star}x_{L-2}\\ \frac{1}{1+\alpha^\star}(x_s)_{s=2}^{L-1}\mathrm{softmax}\Big(\left(w^{\star 2}x_Lx_{s-1}\right)_{s=2}^{L-1}\Big)^\top\end{pmatrix},$$
$$\mathsf{TF}(X;\theta)=\begin{pmatrix}g\cdot(x_{s-1})_{s=2}^{L-1}\mathrm{softmax}\Big((-p(L-1-s))_{s=2}^{L-1}\Big)^\top\\ h\cdot(x_s)_{s=2}^{L-1}\mathrm{softmax}\Big(\left(w^2x_Lx_{s-1}\right)_{s=2}^{L-1}\Big)^\top\end{pmatrix}.$$

And the loss function is expressed as:

$$\mathcal{L}(\theta)=\frac{1}{2}\mathop{\mathbb{E}}_{X\sim\mathbb{N}(0,1)^L}\left[\|f^\star(x)-\mathsf{TF}(x;\theta)\|^2\right]$$
$$=\frac{1}{2}\mathbb{E}_X\left[\left(\frac{\alpha^\star}{1+\alpha^\star}x_{L-2}-g\cdot(x_{s-1})_{s=2}^{L-1}\mathrm{softmax}\Big((-p(L-1-s))_{s=2}^{L-1}\Big)^\top\right)^2\right]$$
$$+\frac{1}{2}\mathbb{E}_X\left[\left(\frac{1}{1+\alpha^\star}(x_s)_{s=2}^{L-1}\mathrm{softmax}\Big(\left(w^{\star 2}x_Lx_{s-1}\right)_{s=2}^{L-1}\Big)^\top-h\cdot(x_s)_{s=2}^{L-1}\mathrm{softmax}\Big(\left(w^2x_Lx_{s-1}\right)_{s=2}^{L-1}\Big)^\top\right)^2\right].$$

The total loss can naturally be divided into two parts:

$$\mathcal{L}(\theta)=\mathcal{L}_{\mathsf{G}_4}(\theta)+\mathcal{L}_{\mathsf{IH}_2}(\theta),$$

where

$$\mathcal{L}_{\mathsf{G}_4}(\theta)=\mathcal{L}_{\mathsf{G}_4}(p,g)$$
$$=\frac{1}{2}\mathbb{E}_X\left[\left(\frac{\alpha^\star}{1+\alpha^\star}x_{L-2}-g\cdot(x_{s-1})_{s=2}^{L-1}\mathrm{softmax}\Big((-p(L-1-s))_{s=2}^{L-1}\Big)\right)^\top)^2\right],$$

$$\mathcal{L}_{\mathsf{IH}_2}(\theta) = \mathcal{L}_{\mathsf{IH}_2}(w,h)$$
$$= \frac{1}{2}\mathbb{E}_X\left[\left(\frac{1}{1+\alpha^\star}(x_s)_{s=2}^{L-1}\mathrm{softmax}\Big(\left(w^{\star 2}x_Lx_{s-1}\right)_{s=2}^{L-1}\Big)^\top - h\cdot(x_s)_{s=2}^{L-2}\mathrm{softmax}\Big(\left(w^2x_Lx_{s-1}\right)_{s=2}^{L-1}\Big)^\top\right)^2\right].$$

Notably, the dynamics of (p,g) and (w,h) are **decoupled**, which allows us to analyze them separately.

Additionally, we denote the optimal values of the parameters as:

$$p^\star = +\infty, \quad g^\star = \frac{\alpha^\star}{1+\alpha^\star}, \quad w^\star := w^\star, \quad h^\star = \frac{1}{1+\alpha^\star}.$$

For the initialization scale and the sequence length, we consider the case:

$$\sigma_{\text{init}} = \mathcal{O}(1) \ll 1, \quad L = \Omega(1/\sigma_{\text{init}}) \gg 1.$$

C.3.1. DYNAMICS OF THE PARAMETERS FOR 4-GRAM

First, we define two useful auxiliary functions:

$$M(p) := \frac{1-e^{-p(L-2)}}{1-e^{-p}},$$
$$m(p) := \frac{e^{-p} - (L-2)e^{-p(L-2)} + (L-3)e^{-p(L-1)}}{(1-e^{-p})^2}.$$

Then, a straightforward calculation, combined with Lemma 7 and Lemma 8, yields the explicit formulation of $\mathcal{L}_{\mathsf{G}_4}(\theta)$ and the GF dynamics of p and g:

$$\mathcal{L}_{\mathsf{G}_4}(\theta) = \frac{1}{2}\left(\frac{\alpha^\star}{1+\alpha^\star}\right)^2 + \frac{1}{2}g^2\frac{M(2p)}{M(p)^2} - \frac{\alpha^\star g}{1+\alpha^\star}\frac{1}{M(p)}. \tag{7}$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} = -\frac{\partial\mathcal{L}}{\partial p} = -\frac{\partial\mathcal{L}_{\mathsf{G}_4}}{\partial p} = \frac{m(p)}{M(p)^2}\left[g^2\frac{m(2p)}{m(p)} - g^2\frac{M(2p)}{M(p)} + \frac{\alpha^\star g}{1+\alpha^\star}\right],$$
$$\frac{\mathrm{d}g}{\mathrm{d}t} = -\frac{\partial\mathcal{L}}{\partial g} = -\frac{\partial\mathcal{L}_{\mathsf{G}_4}}{\partial g} = \frac{\alpha^\star}{1+\alpha^\star}\frac{1}{M(p)} - g\frac{M(2p)}{M(p)^2},$$

Equivalently, the dynamics can be written as:

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{m(p)g}{M(p)^2}\left(g^\star - g\frac{M(2p)}{M(p)} + g\frac{m(2p)}{m(p)}\right),$$
$$\frac{\mathrm{d}g}{\mathrm{d}t} = \frac{1}{M(p)}\left(g^\star - g\frac{M(2p)}{M(p)}\right).$$

Notice that at the initialization, it holds that $\frac{\mathrm{d}p}{\mathrm{d}t}|_{t=0} > 0$ and $\frac{\mathrm{d}g}{\mathrm{d}t}|_{t=0} > 0$. Then we first define a hitting time:

$$T_1^g := \inf\{t > 0 : g(t) > g^\star\}.$$

Noticing $g(0) = \sigma_{\text{init}} \ll g^\star$ and the continuity, $T_1^g > 0$.

Our subsequent proof can be divided into **two phases**: a monotonic phase $t < T_1^g$, and a stable convergence phase $t > T_1^g$.

Part I. Analysis for the monotonic phase $t < T_1^g$.

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{m(p)g}{M(p)^2}\left(g^\star - g\frac{M(2p)}{M(p)} + g\frac{m(2p)}{m(p)}\right) = \frac{m(p)g}{M(p)^2}\left(g^\star - g\frac{1+e^{-p(L-2)}}{1+e^{-p}} + g\frac{m(2p)}{m(p)}\right),$$
$$\frac{\mathrm{d}g}{\mathrm{d}t} = \frac{1}{M(p)}\left(g^\star - g\frac{M(2p)}{M(p)}\right) = \frac{1}{M(p)}\left(g^\star - g\frac{1+e^{-p(L-2)}}{1+e^{-p}}\right).$$

It is easy to see that p, g are monotonically increasing for $t < T_1^g$. We can choose sufficiently large

$$L = \Omega(1/p(0)) = \Omega(1/\sigma_{\text{init}})$$

such that:

$$(L-3)e^{-(L-3)p(t)},\ e^{-(L-5)p(t)} < 0.0001, \quad \forall p > \sigma_{\text{init}}.$$

Then we can calculate the following three terms in the dynamics:

$$\frac{m(p)}{M^2(p)} = \frac{e^{-p}\left(1-(L-2)e^{-p(L-3)}+(L-3)e^{-p(L-2)}\right)}{1-e^{-p(L-2)}} = \frac{e^{-p}(1+\xi_1(p))}{1+\xi_2(p)},$$
$$\frac{1}{M(p)} = \frac{1-e^{-p(L-2)}}{1-e^{-p}} = \frac{1+\xi_3(p)}{1-e^{-p}},$$

$$\frac{m(2p)}{m(p)} = \frac{e^{-p}\left(1-(L-2)e^{-2p(L-3)}+(L-3)e^{-2p(L-2)}\right)}{(1+e^{-p})^2\left(1-(L-2)e^{-p(L-3)}+(L-3)e^{-p(L-2)}\right)}$$
$$= \frac{e^{-p}(1+\xi_4(p))}{(1+e^{-p})^2(1+\xi_5(p))},$$

where the error functions satisfy:

$$|\xi_1(p)|, \cdots, |\xi_5(p)| \leqslant 0.0001,\ \forall t > T_1^g.$$

Then the dynamics satisfy:

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{e^{-p}g(1+\xi_1(p))}{1+\xi_2(p)}\left(g^\star - g\frac{1+e^{-p(L-2)}}{1+e^{-p}} + \frac{ge^{-p}(1+\xi_3(t))}{(1+e^{-p})^2(1+\xi_5(t))}\right),$$
$$\frac{\mathrm{d}g}{\mathrm{d}t} = \frac{1+\xi_3(p)}{1-e^{-p}}\left(g^\star - g\frac{1+e^{-p(L-2)}}{1+e^{-p}}\right).$$

When $g < \frac{1}{2}\frac{\alpha^\star}{1+\alpha^\star}$, we have

$$\frac{\mathrm{d}p}{\mathrm{d}g} \leqslant 2\left(e^{-p}-e^{-2p}\right)g.$$

By define $T^g_{1/2} := \inf\{t > 0 : g(t) > g^\star/2\}$ and $\tilde{p} := p(T^g_{1/2})$, we have

$$\ln(e^{\tilde{p}} - 1) \leqslant \frac{1}{4}g^{\star 2} - g(0)^2 + e^{p(0)} - 1 + \ln(e^{p(0)} - 1)$$

then $\tilde{p} \leqslant \mathcal{O}(\sqrt{p(0)})$, from which we infer that p barely increases when $t \leqslant T^g_{1/2}$.

For $0 \leqslant t \leqslant T^g_{1/2}$,

$$\frac{\mathrm{d}g}{\mathrm{d}t} \geqslant \frac{1}{1 - e^{-p(0)}}\left[g^\star - \frac{g}{1 + e^{-p(0)}}\right]$$

$$g \geqslant g^\star(1 + e^{-p(0))}) + \left[g(0) - g^\star(1 + e^{-p(0)})\right] \exp\left(\frac{-t}{1 - e^{-2p(0)}}\right)$$

so

$$T^g_{1/2} \leqslant (1 - e^{-2p(0)}) \ln\left(\frac{g^\star(1 + e^{-p(0)}) - g(0)}{g^\star\left((1 + e^{-p(0)}) - \frac{1}{2}\right)}\right) = \mathcal{O}\left(2p(0)\right)$$

For $T^g_{1/2} \leqslant t \leqslant T^g_1$, let $p_1 := p(T^g_1)$,

$$\begin{aligned}\frac{\mathrm{d}p}{\mathrm{d}g} &\leqslant 1.01e^{-p}(1 - e^{-p})g\left(1 + \frac{\frac{g}{1+e^{-p}} - \frac{g}{(1+e^{-p})^2}}{\frac{\alpha^\star}{1+\alpha^\star} - \frac{g}{1+e^{-p}}}\right) \\ &\leqslant \frac{1.01}{4}\frac{\alpha^\star}{1 + \alpha^\star}(1 + e^{-p_1})\end{aligned}$$

then

$$p_1 - p(0) \leqslant \frac{1.01}{4}\left(\frac{\alpha^\star}{1 + \alpha^\star}\right)^2 (1 + e^{p_1}),$$

$$p_1 \leqslant \frac{1}{2\left(\frac{\alpha^\star}{1+\alpha^\star}\right)^2 - 1},$$

and we take $\alpha^\star > 1$.

Since for $T^g_{1/2} \leqslant t \leqslant T^g_1$,

$$\frac{\mathrm{d}p}{\mathrm{d}t} \leqslant 2e^{-p}g^\star\left(g^\star - \frac{1}{8}g^\star\right),$$

$$\frac{\mathrm{d}p}{\mathrm{d}t} \geqslant \frac{1}{2}e^{-p}g^\star\left(g^\star - \frac{1}{1 + e^{-p_1}}g^\star\right),$$

we have

$$T^g_1 - t_1 \leqslant \mathcal{O}\left((e^{2p_1} - 1)\left(\frac{1 + \alpha^\star}{\alpha^\star}\right)^2\right).$$

Hence, putting the two part of time together we have

$$\begin{aligned}T^g_1 &\leqslant \mathcal{O}\left(p(0) + (e^{2p_1} - 1)\left(\frac{1 + \alpha^\star}{\alpha^\star}\right)^2\right) \\ &= \mathcal{O}\left(\sigma_{\text{init}} + (e^{2p_1} - 1)\left(\frac{1 + \alpha^\star}{\alpha^\star}\right)^2\right) = \mathcal{O}(1).\end{aligned} \tag{8}$$

Part II. Analysis for the convergence phase $t > T_1^g$.

We will prove that, in this phase, (p, g) keep in a stable region, and the convergence occurs. Recall the dynamics:

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{m(p)g}{M(p)^2}\left(g^\star - g\frac{1+e^{-p(L-2)}}{1+e^{-p}} + g\frac{m(2p)}{m(p)}\right),$$
$$\frac{\mathrm{d}g}{\mathrm{d}t} = \frac{1}{M(p)}\left(g^\star - g\frac{1+e^{-p(L-2)}}{1+e^{-p}}\right).$$

Using contradiction, it is easy to verify that for all $t > T_1^g$,

$$g^\star < g(t) < 2g^\star, \quad \frac{\mathrm{d}p(t)}{\mathrm{d}t} > 0,$$

which means g has entered a stable region (although it is possible that g is non-monotonic), while p keeps increase. In fact, if $T_{2g^\star}^g := \inf\{t > 0 : g(t) = 2g^\star\}$, then $\frac{\mathrm{d}g}{\mathrm{d}t}|_{T_{2g^\star}^g} < 0$, which leads to a contradiction. If $T_0^{\mathrm{d}p/\mathrm{d}t} := \inf\{t > 0 : \frac{\mathrm{d}p(t)}{\mathrm{d}t} = 0\}$, then

$$\left(g^\star - g\frac{1+e^{-p(L-2)}}{1+e^{-p}} + g\frac{m(2p)}{m(p)}\right)\Bigg|_{T_0^{\mathrm{d}p/\mathrm{d}t}} = 0, \quad \frac{\mathrm{d}g}{\mathrm{d}t} < 0,$$

$$\frac{\mathrm{d}}{\mathrm{d}t}\left(g^\star - g\frac{1+e^{-p(L-2)}}{1+e^{-p}} + g\frac{m(2p)}{m(p)}\right) = -g'\frac{1+e^{-p(L-2)}}{1+e^{-p}} + g'\frac{m(2p)}{m(p)} > 0,$$

where the last inequality leads to a contradiction.

Thus, $p(t) > p(T_1^g) > p(0) = \sigma_{\mathrm{init}}$ holds in this phase. Therefore, the dynamics

$$\frac{\mathrm{d}p}{\mathrm{d}t} = \frac{e^{-p}g(1+\xi_1(p))}{1+\xi_2(p)}\left(g^\star - g\frac{1+e^{-p(L-2)}}{1+e^{-p}} + \frac{ge^{-p}(1+\xi_3(t))}{(1+e^{-p})^2(1+\xi_5(t))}\right),$$
$$\frac{\mathrm{d}g}{\mathrm{d}t} = \frac{1+\xi_3(p)}{1-e^{-p}}\left(g^\star - g\frac{1+e^{-p(L-2)}}{1+e^{-p}}\right),$$

also satisfy

$$|\xi_1(p)|, \cdots, |\xi_5(p)| \leqslant 0.0001, \ \forall t > T_1^g.$$

For simplicity, we consider the transform:

$$u := e^{-p}.$$

Then the dynamics of u and g can be written as:

$$\frac{\mathrm{d}u}{\mathrm{d}t} = -\frac{(1+\xi_1(p))u^2 g}{1+\xi_2(p)}\left(g^\star - g\frac{1+u^{L-2}}{1+u} + \frac{gu(1+\xi_4(p))}{(1+u)^2(1+\xi_5(p))}\right),$$
$$\frac{\mathrm{d}g}{\mathrm{d}t} = \frac{1+\xi_3(p)}{1-u}\left(g^\star - g\frac{1+u^{L-2}}{1+u}\right).$$

Notice that this dynamics are controlled by high-order terms. Consequently, we construct a variable to reflect the dynamics of high-order term:

$$v := ug^\star + (g^\star - g).$$

Then the dynamics of u and v satisfy:

$$\begin{aligned}
\frac{\mathrm{d}u}{\mathrm{d}t} &= -\frac{(1+\xi_1(p))u^2 g}{1+\xi_2(p)}\left(\frac{v-u^{L-2}g}{1+u} + \frac{gu(1+\xi_4(p))}{(1+u)^2(1+\xi_5(p))}\right), \\
\frac{\mathrm{d}v}{\mathrm{d}t} &= -\frac{(1+\xi_1(p))u^2 g g^\star}{1+\xi_2(p)}\left(\frac{v-u^{L-2}g}{1+u} + \frac{gu(1+\xi_4(p))}{(1+u)^2(1+\xi_5(p))}\right) - \frac{1+\xi_3(p)}{1-u^2}\left(v-u^{L-2}g\right).
\end{aligned}$$

Now we consider the Lyapunov function about u, v:

$$G(u,v) := \frac{1}{2}\left(u^2+v^2\right).$$

Then it is straightforward:

$$\begin{aligned}
&\frac{\mathrm{d}G}{2\mathrm{d}t} = u\frac{\mathrm{d}u}{\mathrm{d}t} + v\frac{\mathrm{d}v}{\mathrm{d}t} \\
=&-\frac{u^3 g(1+\xi_1(p))}{1+\xi_2(p)}\left(\frac{v-u^{L-2}g}{1+u} + \frac{gu(1+\xi_4(p))}{(1+u)^2(1+\xi_5(p))}\right) \\
&-\frac{(1+\xi_1(p))u^2 v g g^\star}{1+\xi_2(p)}\left(\frac{v-u^{L-2}g}{1+u} + \frac{gu(1+\xi_4(p))}{(1+u)^2(1+\xi_5(p))}\right) \\
&-\frac{1+\xi_3(p)}{1-u^2}\left(v-u^{L-2}g\right)v.
\end{aligned}$$

By $|\xi_1|, \cdots, |\xi_5| \leqslant 0.0001$, we have the following estimate for the Lyapunov dynamics:

$$\begin{aligned}
\frac{\mathrm{d}G}{2\mathrm{d}t} \leqslant& \frac{1.001g}{1+u}|u^3 v| + \frac{1.0001g^2}{1+u}u^{L+1} - \frac{0.999g^2}{(1+u^2)}u^4 \\
&-\frac{0.999gg^\star}{1+u}u^2v^2 + \frac{1.001g^2g^\star}{1+u}|u^L v| + \frac{1.001g^2g^\star}{(1+u^2)}|u^3 v| \\
&-\frac{0.999}{1-u^2}v^2 + \frac{1.001g}{1-u^2}|u^{L-2}v|
\end{aligned}$$

By $u^{L-5} = e^{-p(L-5)} < 0.0001$ and $0 < u < e^{-p(T_1^g)}$, we further have:

$$\begin{aligned}
&\frac{\mathrm{d}G}{2\mathrm{d}t} \leqslant \frac{1.002g}{1+u}|u^3 v| - \frac{0.99g^2}{(1+u)^2}u^4 - \frac{0.999gg^\star}{1+u}u^2v^2 + \frac{1.005g^2g^\star}{(1+u)^2}|u^3v| - \frac{0.999}{1-u^2}v^2 \\
\leqslant& -\frac{0.99g^2}{(1+u)^2}u^4 - \frac{0.99gg^\star}{1+u}u^2v^2 - \frac{0.99}{1-u^2}v^2 + 1.01\left(\frac{g}{1+u} + \frac{g^2g^\star}{(1+u)^2}\right)|u^3v|.
\end{aligned}$$

By using the following inequalities:

$$\frac{g^2g^\star}{(1+u)^2}|u^3v| \leqslant \frac{1}{2}\left(\frac{1.98}{1.01}\frac{gg^\star}{1+u}u^2v^2 + \frac{1.01}{1.98}\frac{g^3g^\star}{(1+u)^3}u^4\right)$$

$$\frac{g}{1+u}|u^3v| \leqslant \frac{1}{2}\left(\frac{0.99}{1.01}(1+u)v^2 + \frac{1.01}{0.99}\frac{g^2}{(1+u)^3}u^6\right)$$
$$-\frac{1}{1-u^2} + \frac{1}{2}(1+u) < -\frac{2}{5}$$

we have

$$\frac{\mathrm{d}G}{\mathrm{d}t} \leqslant -0.99\frac{g^2}{(1+u)^2}u^4 + \frac{1.01}{3.96}\frac{g^3g^\star}{(1+u)^3}u^4 + \frac{1.01}{1.98}\frac{g^2}{(1+u)^3}u^6 - \frac{1.98}{5}v^2.$$

Since $g^\star < g < 2g^\star$, $u > 0$ for $t > T_1^g$, and $\frac{u^2}{1+u} \leqslant \frac{1}{2}$ for $0 \leqslant u \leqslant 1$, we have:

$$\frac{1}{4}\frac{g^3g^\star}{(1+u)^3} + \frac{1}{2}\frac{g^2u^2}{(1+u)^3} \leqslant \frac{g^2}{(1+u)^2}\left(\frac{g^{\star 2}}{2(1+u)} + \frac{u^2}{2(1+u)}\right)$$
$$\leqslant \frac{g^2}{(1+u)^2}\left(\frac{1}{2}+\frac{1}{4}\right) = \frac{3}{4}\frac{g^2}{(1+u)^2},$$

then

$$\frac{\mathrm{d}G(u,v)}{\mathrm{d}t} \leqslant -0.22\frac{g^2}{(1+u)^2}u^4 - \frac{2}{5}v^2$$
$$\leqslant -\frac{0.99}{16}g^{\star 2}u^4 - \frac{1.98}{5}v^2 \leqslant -\frac{g^{\star 2}}{65}G(u,v)^2,$$

which implies:

$$G(u(t),v(t)) \leqslant \frac{1}{G(u(t_1),v(t_1)) + \frac{g^{\star 2}}{64}(t-t_1)}, \quad \forall t > T_1^g.$$

Hence,

$$u^2(t), \quad v^2(t) = \mathcal{O}\left(\frac{1}{g^{\star 2}t}\right) = \mathcal{O}\left(\frac{1}{t}\right), \quad \forall t > T_1^g = \mathcal{O}(1)$$

which implies:

$$\begin{aligned} e^{-p(t)} = u(t) &= \mathcal{O}\left(\frac{1}{\sqrt{t}}\right), \quad \forall t > T_1^g = \mathcal{O}(1); \\ g(t) - g^\star = g^\star u(t) - v(t) &\leqslant \mathcal{O}\left(\frac{g^\star}{\sqrt{t}}\right) + \mathcal{O}\left(\frac{1}{\sqrt{t}}\right) = \mathcal{O}\left(\frac{1}{\sqrt{t}}\right), \quad \forall t > T_1^g = \mathcal{O}(1). \end{aligned} \tag{9}$$

Notably, these proofs capture the **entire** training dynamics of p, g, from $t = 0$ to $t = T_1^g$, and finally to $t \to +\infty$, providing a fine-gained analysis for each phase.

C.3.2. DYNAMICS OF THE PARAMETERS FOR INDUCTION HEAD

Recall the partial loss about the induction head:

$$\mathcal{L}_{\mathsf{IH}_2}(\theta)=\frac{1}{2}\mathbb{E}_X\left[\left(\frac{1}{1+\alpha^\star}(x_s)_{s=2}^{L-1}\mathrm{softmax}\Big(\left(w^{\star 2}x_Lx_{s-1}\right)_{s=2}^{L-1}\Big)^\top-h\cdot(x_s)_{s=2}^{L-2}\mathrm{softmax}\Big(\left(w^2x_Lx_{s-1}\right)_{s=2}^{L-2}\Big)^\top\right)^2\right].$$

Technical simplification. Unlike $\mathcal{L}_{\mathsf{G}_4}(\theta)$, the denominators of the softmax terms $\mathrm{softmax}\Big(\left(w^{\star 2}x_Lx_{s-1}\right)_{s=2}^{L-1}\Big)$ and $\mathrm{softmax}\Big(\left(w^2x_Lx_{s-1}\right)_{s=2}^{L-2}\Big)$ in $\mathcal{L}_{\mathsf{IH}_2}(\theta)$ depend on the input tokens X, making it hard to derive a closed-form expression for $\mathcal{L}_{\mathsf{IH}_2}(\theta)$. In [5], the authors consider a simplified transformer model, which replaces $\mathrm{softmax}(z_1,\cdots,z_L)$ with $\frac{1}{L}\exp(z_1,\cdots,z_L)$. This approximation is nearly tight when $z_1,\cdots,z_L\approx 0$. Notice that 1) $w^2x_Lx_{s-1}\approx 0$ holds near the small initialization, i.e., for $w\approx\sigma_{\text{init}}\ll 1$. In fact, our analysis shows that $w\approx\sigma_{\text{init}}$ is maintained over a long period. 2) $w^\star=\mathcal{O}(1)$, which implies that $w^2x_Lx_{s-1}\approx 0$ for most input sequence. Thus, we adopt the simplification used in [5], resulting in the following approximation of the loss function:

$$\mathcal{L}_{\mathsf{IH}_2}(\theta):=\frac{1}{2}\mathbb{E}_X\left[\left(\frac{1}{1+\alpha^\star}\frac{1}{L-2}\sum_{s=2}^{L-1}\exp(w^{\star 2}x_Lx_{s-1})x_s-h\frac{1}{L-2}\sum_{s=2}^{L-2}\exp(w^2x_Lx_{s-1})x_s\right)^2\right].$$

Then by a straightforward calculation with Lemma 7, we can derive its explicit formulation:

$$\mathcal{L}_{\mathsf{IH}_2}(\theta)=\frac{(1-4w^{\star 4})^{-\frac{1}{2}}}{2(1+\alpha^\star)^2(L-2)}+\frac{1}{2}\frac{h^2}{L-2}(1-4w^4)^{-\frac{1}{2}}-\frac{h(1-(w^2+w^{\star 2})^2)^{-\frac{1}{2}}}{(1+\alpha^\star)(L-2)}. \tag{10}$$

Furthermore, we can calculate GF dynamics as follows:

$$\begin{aligned}\frac{\mathrm{d}w}{\mathrm{d}t}&=\frac{h}{(1+\alpha^\star)(L-2)}(1-(w^2+w^{\star 2})^2)^{-\frac{3}{2}}\cdot(w^2+w^{\star 2})\cdot 2w-\frac{h^2}{L-2}(1-4w^4)^{-\frac{3}{2}}\cdot 4w^3,\\ \frac{\mathrm{d}h}{\mathrm{d}t}&=\frac{1}{(1+\alpha^\star)(L-2)}(1-(w^2+w^{\star 2})^2)^{-\frac{1}{2}}-\frac{h}{L-2}(1-4w^4)^{-\frac{1}{2}}.\end{aligned}$$

For simplicity, we denote:

$$w^\star:=w^\star,\quad h^\star:=\frac{1}{1+\alpha^\star}.$$

Part I. The trend and monotonicity of w,h.

For simplicity, we denote the tuning time point of h:

$$T_2^h:=\inf\left\{t>0:\frac{\mathrm{d}h(t)}{\mathrm{d}t}=0\right\}.$$

In this step, we will prove the following three claims regarding the trend and monotonicity of w,h, which are essential for our subsequent analysis:

- **(P1.1)** h initially increases beyond $h^\star$, and then remains above this value.
- **(P1.2)** w keeps increasing but always stays below $w^\star$.

- **(P1.3)** h increases before T_2^h, but decreases after T_2^h.

(P1.1) *h initially increases beyond $h^\star$, and then remains above this value.*
We will prove that initially, h increases beyond $h^\star$, and keeps growing beyond $h^\star$. Define

$$T_1^h := \inf\{t > 0 : h(t) > h^\star\},$$

we will prove that h remains above $h^\star$ thereafter.
For simplicity, we denote

$$\psi(x) = (1-x^2)^{-\frac{1}{2}}, \quad \phi(x) = (1-x^2)^{-\frac{3}{2}} \cdot x,$$

then the dynamics holds:

$$\begin{aligned}
\frac{\mathrm{d}h}{\mathrm{d}t} &= \frac{h}{L-2}\psi(w^2+w^{\star 2})\left[\frac{h^\star}{h} - \frac{\psi(2w^2)}{\psi(w^2+w^{\star 2})}\right], \\
\frac{\mathrm{d}w}{\mathrm{d}t} &= \frac{2h^2 w}{L-2} \cdot \phi(w^2+w^{\star 2}) \cdot \left[\frac{h^\star}{h} - \frac{\phi(2w^2)}{\phi(w^2+w^{\star 2})}\right].
\end{aligned}$$

Notice that $\frac{\phi(2w^2)}{\phi(w^2+w^{\star 2})} < \frac{\psi(2w^2)}{\psi(w^2+w^{\star 2})}$, $w < w^\star$, while $\frac{\phi(2w^2)}{\phi(w^2+w^{\star 2})} > \frac{\psi(2w^2)}{\psi(w^2+w^{\star 2})}$, $w > w^\star$.
We denote the first hitting time of h decreasing to $h^\star$ as $T_{h^\star}^h$:

$$T_{h^\star}^h := \inf\left\{t > T_2^h : h(t) < h^\star\right\}.$$

If $w(T_{h^\star}^h) \geqslant w^\star$, then at the first hitting time of w increasing to $w^\star$, $\frac{\mathrm{d}w}{\mathrm{d}t} < 0$, which leads to a contradiction. If $w(T_{h^\star}^h) < w^\star$, then $\frac{\mathrm{d}h}{\mathrm{d}t}|_{T_{h^\star}^h} > 0$, which also leads to a contradiction. Hence, $T_{h^\star}^h = +\infty$, which means that h always remains above $h^\star$ for $t > T_2^h$.

(P1.2) *w keeps increasing but always below $w^\star$.*
We first prove that w always remains below $w^\star$. We denote the first hitting time of w increasing to $w^\star$ as t', then it is not difficult to see $\frac{\mathrm{d}w}{\mathrm{d}t}|_{t'} < 0$, which leads to a contradiction.

Next we prove that w keeps increasing throughout. We define the following functions

$$H := \frac{1}{1+\alpha^\star}\left(1-(w^2+w^{\star 2})^2\right)^{-\frac{3}{2}}(w^2+w^{\star 2}) - h(1-4w^4)^{-\frac{3}{2}} \cdot 2w^2$$

$$Q := \frac{1}{1+\alpha^\star}\left(1-(w^2+w^{\star 2})^2\right)^{-\frac{1}{2}} - h\left(1-4w^4\right)^{-\frac{1}{2}}$$

If at some $\bar{t}$, $\frac{\mathrm{d}w}{\mathrm{d}t}$ reaches its zero point at the first time, then

$$\left.\frac{\mathrm{d}H}{\mathrm{d}t}\right|_{\bar{t}} = -h'(\bar{t})(1-4w^{\star 4})^{-\frac{3}{2}} \cdot 2w(\bar{t}) > 0,$$

which leads to a contradiction. Hence $\bar{t}$ does not exist and w keeps increasing.

(P1.3) *After the tuning point $t > T_2^h$, h will be monotonically decreasing.*

The first sign-changing zero point of $\frac{\mathrm{d}h}{\mathrm{d}t}$ is T_2^h, then $Q(T_2^h)=0$. $H(T_2^h)>0$,

$$\left.\frac{\mathrm{d}Q}{\mathrm{d}t}\right|_{T_2^h} = \frac{1}{1+\alpha^\star}(1-(w(T_2^h)^2+w^{\star 2})^2)^{-\frac{1}{2}}\cdot 2w(T_2^h)\cdot w'(T_2^h)$$
$$\cdot\left[(1-(w(T_2^h)^2+w^{\star 2})^2)^{-1}\cdot(w(T_2^h)^2+w^{\star 2})-(1-4w(T_2^h)^4)^{-1}\cdot 4w(T_2^h)^2\right].$$

We can see that T_2^h is a sign-changing zero point only if

$$\frac{(1-4w(T_2^h)^4)\cdot(w(T_2^h)^2+w^{\star 2})}{(1-(w(T_2^h)^2+w^{\star 2})^2)\cdot 4w(T_2^h)^2}<1,$$

i.e. we have:

$$w(T_2^h)>w^\circ:=\sqrt{\frac{3-4w^{\star 4}-\sqrt{(4w^{\star 4}-3)^2-16w^{\star 4}}}{8w^{\star 2}}}\geqslant\frac{w^\star}{2}, \tag{11}$$

when $w^\star=\mathcal{O}(1)$.

Next we show that h keeps decreasing after T_2^h. We denote the first zero point of $\frac{\mathrm{d}h}{\mathrm{d}t}$ as t°, then $Q(t^\circ)=0$. Since $\frac{\mathrm{d}w}{\mathrm{d}t}|_{t^\circ}>0$, we have $\frac{\mathrm{d}Q}{\mathrm{d}t}|_{t^\circ}>0$ which leads to a contradiction. Hence t° does not exist and h keeps decreasing after T_2^h.

Part II. Estimation of T_1^h, T_2^h, and the tight estimate of $w(t)$ before T_2^h.

At the first stage, we prove that h grows first and w barely increases. If $w\leqslant 0.01w^\star$ and $h\leqslant\frac{1}{1+\alpha^\star}\frac{(1-w^{\star 4})^{-\frac{1}{2}}}{(1-0.01^4w^{\star 4})^{-\frac{1}{2}}}$,

$$\frac{\mathrm{d}h}{\mathrm{d}t}\geqslant\frac{-1}{L-2}\left[h(1-0.01^4w^{\star 4})^{-\frac{1}{2}}-\frac{1}{1+\alpha^\star}(1-w^{\star 4})^{-\frac{1}{2}}\right],$$

$$h\geqslant\frac{1}{1+\alpha^\star}\frac{(1-w^{\star 4})^{-\frac{1}{2}}}{(1-0.01^4w^{\star 4})^{-\frac{1}{2}}}-\left[\frac{1}{1+\alpha^\star}\frac{(1-w^{\star 4})^{-\frac{1}{2}}}{(1-0.01^4w^{\star 4})^{-\frac{1}{2}}}-h(0)\right]\exp\left(\frac{-t}{(L-2)(1-0.01w^{\star 4})^{\frac{1}{2}}}\right). \tag{12}$$

For h increasing from $h(0)$ to $\frac{1}{1+\alpha^\star}$, it takes

$$T_1^h\leqslant(1-0.01w^{\star 4})^{\frac{1}{2}}(L-2)\ln\left(\frac{1}{1-\frac{(1-w^{\star 4})^{\frac{1}{2}}}{(1-0.01^4w^{\star 4})^{\frac{1}{2}}}}\right)$$
$$\leqslant 2(L-2)(1-\frac{1}{2}w^{\star 4})=\mathcal{O}(L). \tag{13}$$

For $0\leqslant t\leqslant T_1^h$,

$$\frac{\mathrm{d}w}{\mathrm{d}t}\leqslant\frac{1}{L-2}(1-4w^{\star 4})^{-\frac{3}{2}}\cdot w^{\star 2}\cdot 4w.$$

Hence, it take $\mathcal{O}(L\log(1/\sigma_{\text{init}}))$ for w to reach $0.01w^\star$, which allows sufficient time for h to reach $\frac{1}{1+\alpha^\star}$ beforehand.

Therefore, there exists a small constant $\varepsilon(w(0), w^\star)$ only depends on $w(0)$ and $w^\star$ such that h is dominated by $1+\varepsilon(w(0),w^\star)$ times right hand side of (12), from which we deduce that (13) is a tight estimation of T_1^h instead of an upper bound, i.e. $T_1^h = \Theta(L)$.

We then give a bound for $h(T_2^h)$. By $\frac{\mathrm{d}h}{\mathrm{d}t} = 0$,

$$h(T_2^h)/h^\star \leqslant \frac{(1-4w^4)^{\frac{1}{2}}}{(1-(w^2+w^{\star 2})^2)^{\frac{1}{2}}} := r(w).$$

Moreover, $r(w)$ is an decreasing function of w for $w > w^\circ$, and w° is a function of $w^\star$, we have

$$h(T_2^h)/h^\star \leqslant r(w^\circ) := R(w^\star),$$

where w° is a function about $w^\star$, defined in Eq. (11). It is clear that

$$R(w^\star = 0) = 1, \quad R'(w^\star = 0) = 0.$$

Then using the continuity of $R'(\cdot)$ (in $[0, 0.4]$), there exists $c > 0$ such that $|R'(w^\star)| < 0.04$ holds for all $0 < w^\star < c$, which implies:

$$R(w^\star) = R(0) + \int_0^{w^\star} R'(v)\mathrm{d}v < 1 + 0.04w^\star, \quad 0 < w^\star < c.$$

i.e., if $w^\star = O(1)$, then $R(w^\star) < 1 + 0.04w^\star$. This implies:

$$h^\star \leqslant h(t) \leqslant (1 + 0.04375w^\star)h^\star, \quad \forall t \geqslant T_1^h. \tag{14}$$

By some computation, we can prove that $w^\circ(w^\star)$ is an increasing function of $w^\star$, and is always above $\frac{1}{2}w^\star$. Thus we obtain a lower bound of w° for the estimation of lower bound of T_2^h:

For the second stage, h barely changes and w starts to grow exponentially fast, and we use the tight estimation of $T_{1/2}^w := \inf\left\{t > 0 : w(t) > \frac{1}{2}w^\star\right\}$ to give a lower bound of T_2^h. During this stage,

$$\begin{aligned}\frac{\mathrm{d}w}{\mathrm{d}t} &\leqslant \frac{2w}{(1+\alpha^\star)^2(L-2)}\left[(1-(w^2+w^{\star 2})^2)^{-\frac{3}{2}}\cdot(w^2+w^{\star 2}) - (1-4w^4)^{\frac{3}{2}}\right] \\ &\leqslant \frac{2w}{(1+\alpha^\star)^2(L-2)}(1-4w^{\star 4})^{\frac{3}{2}}\cdot 2w^{\star 2},\end{aligned}$$

and w has upper bound

$$w \leqslant w(0)\exp\left(\frac{4w^{\star 2}(1-4w^{\star 4})^{\frac{3}{2}}}{(1+\alpha^\star)(L-2)}t\right). \tag{15}$$

Hence, the lower bound of time for w to reach $\frac{1}{2}w^\star$ is

$$T_{1/2}^w - T_1^h = \frac{(1+\alpha^\star)^2(L-2)}{4w^{\star 2}(1-4w^{\star 4})^{\frac{3}{2}}}\ln(\frac{w^\star}{2w(0)}),$$

and lower bound for $T_{1/2}^w$ is

$$T_{1/2}^w \geqslant (L-2)\left[\frac{(1+\alpha^\star)^2\ln(\frac{w^\star}{2w(0)})}{4w^{\star 2}(1-4w^{\star 4})^{\frac{3}{2}}} - \ln\left(1-(1-w^{\star 4})^{\frac{1}{2}}\right)\right]$$
$$\geqslant \frac{(L-2)(1+\alpha^\star)^2}{16w^{\star 2}}\ln\left(\frac{1}{w(0)}\right) = \Omega\left(\frac{(1+\alpha^\star)^2 L}{w^{\star 2}}\log\left(\frac{1}{\sigma_{\text{init}}}\right)\right). \tag{16}$$

On the other hand, we estimate the lower bound of w. Let

$$C(x) = (1-x^2)^{-\frac{3}{2}}\cdot x,$$

then

$$C'(x) = 3(1-x^2)^{-\frac{5}{2}}x^2 + (1-x^2)^{-\frac{3}{2}} > 1, \quad 0<x<1,$$
$$C''(x) = 15x^3(1-x^2)^{-\frac{7}{2}} + 6x(1-x^2)^{-\frac{5}{2}} + 3x(1-x^2)^{-\frac{5}{2}} > 0, \quad 0<x<1.$$

$C(x)$ is a monotonically increasing convex function on $(0,1)$ and $C(x)\geqslant x$.

Using conclusions above, before w^2 increases to $\frac{1}{2\gamma(w^\star)+\beta-1}w^{\star 2}$ for some $\beta>0$,

$$\begin{aligned}
&C(w^2+w^{\star 2})\\
&\geqslant C((2\gamma(w^\star)+\beta)w^2)\\
&\geqslant C(2\gamma(w^\star)\cdot w^2) + C(\beta w^2) \quad \text{(Lemma 10)}\\
&\geqslant \gamma(w^\star)\cdot C(2w^2) + \beta w^2 \quad (C(ax)\geqslant aC(x), \text{for } a>1)
\end{aligned}$$

then we have

$$\begin{aligned}
\frac{\mathrm{d}w}{\mathrm{d}t} &\geqslant \frac{2w}{(1+\alpha^\star)^2(L-2)}(C(w^2+w^{\star 2}) - \gamma(w^\star)\cdot C(2w^2))\\
&\geqslant \frac{2w}{(1+\alpha^\star)^2(L-2)}\frac{\beta}{\gamma(w^\star)+\beta}w^{\star 2}
\end{aligned}$$

and

$$w \geqslant w(0)\exp\left(\frac{2\beta}{\gamma(w^\star)+\beta}\frac{1}{(1+\alpha^\star)^2(L-2)}w^{\star 2}t\right).$$

Take $\beta=2$, then

$$w \geqslant w(0)\exp\left(\frac{w^{\star 2}t}{(1+\alpha^\star)^2(L-2)}\right),\ \forall t\in[0,T_{1/2}^w]. \tag{17}$$

From the above inequality, (16) is not only an upper bound, but a tight estimation of $T_{1/2}^w$, i.e.

$$T_{1/2}^w = \Theta\left(\frac{(1+\alpha^\star)^2 L}{w^{\star 2}}\log\left(\frac{1}{\sigma_{\text{init}}}\right)\right).$$

Part II. Dynamics after the critical point $T_{1/2}^w$.

For simplicity, we consider:

$$v := w^2,$$

and denote $v^\star := w^{\star 2}, h^\star := \frac{1}{1+\alpha^\star}$. Then we focus on the dynamics of v and h.

Additionally, we introduce a few notations used in this part:

$$\phi(x) := \frac{x}{(1-x^2)^{3/2}}, \quad \psi(x) := \frac{1}{(1-x^2)^{1/2}}.$$

Then the dynamics of v and g are:

$$\begin{aligned}\frac{\mathrm{d}v}{\mathrm{d}t} &= \frac{4vh}{L-2}\big(h^\star\phi(v+v^\star) - h\phi(2v)\big),\\ \frac{\mathrm{d}h}{\mathrm{d}t} &= \frac{1}{L-2}\big(h^\star\psi(v+v^\star) - h\psi(2v)\big).\end{aligned}$$

Step II.1. *A coarse estimate of the relationship between v and h.*

It is easy to verify the monotonicity that $\frac{\mathrm{d}v}{\mathrm{d}t} > 0$ and $\frac{\mathrm{d}h}{\mathrm{d}t} < 0$ for $t > t_2$. Additionally, we have

$$\frac{\psi(v+v^\star)}{\psi(2v)} < \frac{h}{h^\star} < \frac{\phi(v+v^\star)}{\phi(2v)}.$$

Then by Monotone convergence theorem, we obtain:

$$\lim_{t\to+\infty} v = v^\star, \quad \lim_{t\to+\infty} h = h^\star.$$

Step II.2. *Convergence analysis by Lyapunov function.*

This step aims to establish the convergence rate of v and h.

In fact, the dynamics of v, h can be approximately characterized by their linearized dynamics. In contrast, the dynamics of p, g are controlled by high-order terms. Therefore, the proof for v and h is significantly simpler than the corresponding proof for p and g. We only need to consider the simplest Lyapunov function:

$$G(v,h) := \frac{1}{2}\Big((v-v^\star)^2 + (h-h^\star)^2\Big).$$

It is easy to verify that

$$\begin{aligned}&(L-2)\frac{\mathrm{d}G(v,h)}{\mathrm{d}t} = (v-v^\star)\frac{\mathrm{d}v}{\mathrm{d}t} + (h-h^\star)\frac{\mathrm{d}h}{\mathrm{d}t}\\ =&4vh(v-v^\star)\big(h^\star\phi(v+v^\star) - h\phi(2v)\big) + (h-h^\star)\big(h^\star\psi(v+v^\star) - h\psi(2v)\big)\\ =&4vh(v-v^\star)\Big(\phi(v+v^\star)(h^\star-h) - h(\phi(v+v^\star)-\phi(2v))\Big)\\ &+ (h-h^\star)\Big((h^\star-h)\psi(v+v^\star) + h(\psi(v+v^\star)-\psi(2v))\Big)\\ =&-4vh^2(v^\star-v)(\phi(v+v^\star)-\phi(2v)) - \psi(v+v^\star)(h-h^\star)^2\\ &+ 4vh\phi(v+v^\star)(v-v^\star)(h^\star-h) + h(h-h^\star)(\psi(v+v^\star)-\psi(2v)).\end{aligned}$$

Let $v^\star \leqslant 0.3 = \mathcal{O}(1)$. Recalling (11) and (14), as well as the monotonicity about p and w, we have:

$$\frac{v^\star}{4} < v(t) < v^\star; \quad h^\star < h(t) < 1.02h^\star, \quad \forall t > T_2^h.$$

Combining these estimates with the properties of ϕ and ψ, we have the following straight-forward estimates:

$$\begin{aligned}
\phi(v+v^\star)-\phi(2v)=\phi'(\xi)(v^\star-v)=\frac{1+2\xi^2}{(1-\xi^2)^{5/2}}(v^\star-v)\geqslant v^\star-v;\\
\phi(v+v^\star)\leqslant\phi(2v^\star)\leqslant 1;\\
\psi(v+v^\star)=\frac{1}{(1-(v+v^\star)^2)^{1/2}}\geqslant 1;\\
\psi(v+v^\star)-\psi(2v)=\psi'(\xi)(v^\star-v)=\frac{\xi}{(1-\xi^2)^{3/2}}(v^\star-v)\leqslant 1.3v^\star(v^\star-v).
\end{aligned}$$

Thus, we have the following estimate for the Lyapunov function:

$$\begin{aligned}
&(L-2)\frac{\mathrm{d}G(v,h)}{\mathrm{d}t}\\
\leqslant&-\frac{4}{1.02}v^\star h^{\star 2}(v-v^\star)^2-(h-h^\star)^2\\
&+4.08v^\star h^\star(v-v^\star)(h^\star-h)+1.3\cdot 1.02v^\star h^\star(v^\star-v)(h-h^\star)\\
=&-\frac{4}{1.02}v^\star h^{\star 2}(v-v^\star)^2-(h-h^\star)^2+5.41v^\star h^\star(v^\star-v)(h-h^\star)\\
\leqslant&-3.92v^\star h^{\star 2}(v-v^\star)^2-(h-h^\star)^2+\left(9.6v^{\star 2}h^{\star 2}(v-v^\star)^2+\frac{3}{4}(h-h^\star)^2\right)\\
\leqslant&-(3.92-9.6\cdot 0.3)v^\star h^{\star 2}(v-v^\star)^2-0.25(h-h^\star)^2\leqslant-\frac{1}{4}v^\star h^{\star 2}G(v,h).
\end{aligned}$$

Consequently, we have the exponential bound for all $t>T_2^h$:

$$G(v(t),h(t))\leqslant G\left(v(T_2^h),h(T_2^h)\right)\exp\left(-\frac{v^\star h^{\star 2}}{4(L-2)}(t-T_2^h)\right),\quad\forall t>T_2^h,$$

This can imply:

$$\begin{aligned}
(h(t)-h^\star)^2&=(h(T_2^h)-h^\star)^2\exp\left(-\Omega\left(\frac{w^{\star 2}(t-T_2^h)}{L(1+\alpha^\star)^2}\right)\right)\\
&=\mathcal{O}\left(h^{\star 2}\exp\left(-\Omega\left(\frac{w^{\star 2}(t-T_2^h)}{L(1+\alpha^\star)^2}\right)\right)\right),\quad\forall t>T_2^h;\\
(w(t)-w^\star)^2&=(w(T_2^h)-w^\star)^2\exp\left(-\Omega\left(\frac{w^{\star 2}(t-T_2^h)}{L(1+\alpha^\star)^2}\right)\right)\\
&=\mathcal{O}\left(w^{\star 2}\exp\left(-\Omega\left(\frac{w^{\star 2}(t-T_2^h)}{L(1+\alpha^\star)^2}\right)\right)\right),\quad\forall t>T_2^h.
\end{aligned}\tag{18}$$

Notably, these proofs capture the **entire** training dynamics of w,h, from $t=0$ to $t=T_1^h$, to $t=T_{1/2}^w\leqslant T_2^h$, and finally to $t\to+\infty$, providing a fine-gained analysis for each phase.

C.4. Proof of Theorem 4

This theorem is a direct corollary of our analysis of the entire training dynamics in Appendix C.3.1 and C.3.2, leveraging the relationship between the parameters and the loss.

Proof of Phase I (partial learning).

By combining (7) and (9), it follows that: $\mathcal{L}_{\mathsf{G}_4}(\theta(0)) = \Theta(1)$. Moreover,

$$\mathcal{L}_{\mathsf{G}_4}(\theta(t)) = \mathcal{O}\left(\frac{1}{t}\right), \quad t > T_1^g = O(1).$$

Thus, there exists a sufficiently large $T_{\mathrm{I}} = \Theta(1)$, such that:

$$\mathcal{L}_{\mathsf{G}_4}(\theta(T_{\mathrm{I}})) \leqslant 0.01\mathcal{L}_{\mathsf{G}_4}(\theta(0)).$$

Recalling our proof in Appendix C.3.2, for $t < T_{1/2}^h = \mathcal{O}(L)$, it holds that $h(t) < \sigma_{\text{init}} + \mathcal{O}(t/((1+\alpha^\star)L))$, $w(t) < \sigma_{\text{init}} + o(t/((1+\alpha^\star)L))$. Additionally, since $T_{\mathrm{I}} = \Theta(1) \ll \Theta(L)$, it follows that

$$w(T_{\mathrm{I}}) = \mathcal{O}(\sigma_{\text{init}} + 1/L) < 2\sigma_{\text{init}} \ll w^\star, \quad h(T_{\mathrm{I}}) = \mathcal{O}(\sigma_{\text{init}} + 1/L) < 2\sigma_{\text{init}} \ll h^\star.$$

Substituting these estimates into (10), we obtain by Lipschitz continuity of $\mathcal{L}_{\mathsf{lH}_2}$:

$$\begin{aligned}
|\mathcal{L}_{\mathsf{lH}_2}(\theta(T_{\mathrm{I}})) - \mathcal{L}_{\mathsf{lH}_2}(\theta(0))| &\leqslant 2\sigma_{\text{init}}\left(\left|\frac{\partial \mathcal{L}_{\mathsf{lH}_2}}{\partial w}\right| + \left|\frac{\partial \mathcal{L}_{\mathsf{lH}_2}}{\partial h}\right|\right) \\
&\leqslant 2\sigma_{\text{init}}\left(\mathcal{O}\left(\frac{1}{(1+\alpha^\star)L}\right) + o\left(\frac{1}{(1+\alpha^\star)L}\right)\right) \\
&\leqslant 0.01\mathcal{L}_{\mathsf{lH}_2}(\theta(0)).
\end{aligned}$$

Thus,

$$\mathcal{L}_{\mathsf{lH}_2}(\theta(T_{\mathrm{I}})) \geqslant 0.99\mathcal{L}_{\mathsf{lH}_2}(\theta(0)).$$

Proof of Phase II (plateau) + Phase III (emergence).

First, (15) and (17) ensures that w grows exponentially before $t < T_{1/2}^w$:

$$\sigma_{\text{init}} \exp\left(\frac{w^{\star 2}}{(1+\alpha^\star)^2(L-2)}t\right) \leqslant w \leqslant \sigma_{\text{init}} \exp\left(\frac{4w^{\star 2}(1-4w^{\star 4})^{\frac{3}{2}}}{(1+\alpha^\star)(L-2)}t\right).$$

Thus, we have:

$$w(t) = \sigma_{\text{init}} \exp\left(\Theta\left(\frac{w^{\star 2}t}{(1+\alpha^\star)^2 L}\right)\right), \quad t < \Theta\left(\frac{(1+\alpha^\star)^2 L}{w^{\star 2}} \log\left(\frac{1}{\sigma_{\text{init}}}\right)\right).$$

Now we define the observation time $T_o := T_1^h = \Theta(L)$. Notably,

$$h(T_o) = h^\star, \quad w(T_o) < 0.01w^\star.$$

The exponential growth of w further implies:

$$T_{0.01}^{w} := \{t > 0 : w(t) > 0.01w^\star\} = \Theta\left(\frac{(1+\alpha^\star)^2 L}{w^{\star 2}} \log\left(\frac{1}{\sigma_{\text{init}}}\right)\right).$$

Regarding the dynamics of h, by (14), we have $|h(t) - h(T_o)| < 0.02|h(T_o)|,\ \forall t \geqslant T_o$.

Now we incorporate these facts ($0 < w(T_o) < 0.01w^\star$, $0 < w(T_{0.01}^w) \leqslant 0.01w^\star$, $|h(T_{0.01}^w) - h(T_o)| < 0.02|h(T_o)|$, $h(T_o) = h^\star$) into the loss (10). By the Lipschitz continuity of $\mathcal{L}_{\mathsf{IH}_2}$, it is straightforward that

$$\mathcal{L}_{\mathsf{IH}_2}(\theta(T_{0.01}^w)) \geqslant 0.99\mathcal{L}(\theta(T_o)).$$

Thus, we have established the lower bound for T_{II}:

$$\begin{aligned} T_{\text{II}} &:= \inf\left\{t > T_o : \mathcal{L}_{\mathsf{IH}_2}(\theta(t)) \leqslant 0.99 \cdot \mathcal{L}_{\mathsf{IH}_2}(\theta(T_o))\right\} \\ &\geqslant T_{0.01}^w = \Omega\left(\frac{(1+\alpha^\star)^2 L}{w^{\star 2}} \log\left(\frac{1}{\sigma_{\text{init}}}\right)\right). \end{aligned}$$

Combining the loss (10) and our parameter estimates (18), we obtain:

$$\mathcal{L}_{\mathsf{IH}_2}(\theta(t)) = \mathcal{O}\left(\exp\left(-\Omega\left(\frac{w^{\star 2} t}{L(1+\alpha^\star)^2}\right)\right)\right),\ t > T_2^h = \Theta\left(\frac{(1+\alpha^\star)^2 L}{w^{\star 2}} \log\left(\frac{1}{\sigma_{\text{init}}}\right)\right).$$

This implies the upper bound for T_{III}:

$$\begin{aligned} &T_{\text{III}} := \inf\left\{t > T_o : \mathcal{L}_{\mathsf{IH}_2}(\theta(t)) \leqslant 0.01 \cdot \mathcal{L}_{\mathsf{IH}_2}(\theta(T_o))\right\} \\ &= T_{1/2}^w + \mathcal{O}\left((\alpha^\star+1)^2 L \log(1/\sigma_{\text{init}})/w^{\star 2}\right) = \mathcal{O}\left((\alpha^\star+1)^2 L \log(1/\sigma_{\text{init}})/w^{\star 2}\right). \end{aligned}$$

Combining the fact $T_{\text{II}} < T_{\text{III}}$, the lower bound for T_{II}, and the uppper bound for T_{III}, we obtain the two-sided bounds for both T_{II} and T_{III}:

$$T_{\text{II}},\ T_{\text{III}} = \Theta\left((\alpha^\star+1)^2 L \log(1/\sigma_{\text{init}})/w^{\star 2}\right).$$

Proof of Phase IV (convergence).

By combining the loss (7), (10), and our parameter estimates (9), (18), it follows that:

$$\mathcal{L}_{\mathsf{G}_4}(\theta(t)) = \mathcal{O}\left(\frac{1}{t}\right), \quad \mathcal{L}_{\mathsf{IH}_2}(\theta(t)) = \mathcal{O}\left(\exp\left(-\Omega\left(\frac{w^{\star 2} t}{L(1+\alpha^\star)^2}\right)\right)\right), \quad t > T_{\text{III}}.$$

Appendix D. Useful Inequalities

Lemma 6 (Corollary A.7 in Edelman et al. [13]) *For any $\theta, \theta' \in \mathbb{R}^d$, we have*

$$\|\text{softmax}(\theta) - \text{softmax}(\theta')\|_1 \leqslant 2\|\theta - \theta'\|_\infty$$

Lemma 7 $\mathop{\mathbb{E}}_{X,Y,Z} \exp(aXY)Z^2 = (1-a^2)^{-1/2}$, $a < 1$.

Proof [Proof of Lemma 7]

$$\begin{aligned}
&\int \exp(aXY)Z^2 \left(\frac{1}{2\pi}\right)^{-3/2} \exp(-\frac{1}{2}X^2 - \frac{1}{2}Y^2 - \frac{1}{2}Z^2)\, dXdYdZ \\
&= \int \frac{1}{2\pi} \exp(-\frac{1}{2}(X - aY)^2 - \frac{1}{2}Y^2 + \frac{1}{2}a^2Y^2)\, d(X - aY)dY \\
&= \int \frac{1}{\sqrt{2\pi}} \exp(-\frac{1}{2}W^2)\, dW \ (W = (1 - a^2)^{1/2}Y) \\
&= (1 - a^2)^{-1/2}
\end{aligned}$$

■

Lemma 8 *Let* $M(p) := \frac{1-e^{-p(L-2)}}{1-e^{-p}}$*, then it holds that*

$$\left\|\mathrm{softmax}\big((-p(L-1-s))_{s=1}^{L-1}\big)\right\|_2^2 = \frac{M(2p)}{M(p)^2}.$$

Definition 9 (weakly majorizes) *A vector* $\mathbf{x} \in \mathbb{R}^n$ *is said to weakly majorize another vector* $\mathbf{y} \in \mathbb{R}^n$*, denoted by* $\mathbf{x} \prec_w \mathbf{y}$*, if the following conditions hold:*

1. $\sum_{i=1}^{k} x_{[i]} \leqslant \sum_{i=1}^{k} y_{[i]}$ *for all* $k = 1, 2, \ldots, n-1$*,*
2. $\sum_{i=1}^{n} x_{[i]} = \sum_{i=1}^{n} y_{[i]}$*,*

where $x_{[i]}$ *and* $y_{[i]}$ *are the components of* $\mathbf{x}$ *and* $\mathbf{y}$*, respectively, arranged in decreasing order.*

Lemma 10 (Weighted Karamata Inequality) *Let* $f : \mathbb{R} \to \mathbb{R}$ *be a convex function, and let* $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ *and* $\mathbf{y} = (y_1, y_2, \ldots, y_n)$ *be two vectors in* $\mathbb{R}^n$*. If* $\mathbf{x}$ *weakly majorizes* $\mathbf{y}$ *(i.e.,* $\mathbf{x} \prec_w \mathbf{y}$*), and* $w_1, w_2, \ldots, w_n$ *are non-negative weights such that*

$$\sum_{i=1}^{n} w_i = 1,$$

then the following inequality holds:

$$\sum_{i=1}^{n} w_i f(x_i) \leqslant \sum_{i=1}^{n} w_i f(y_i).$$