

Transformers Provably Learn Chain-of-Thought Reasoning with Length Generalization

Yu Huang^{*†} Zixin Wen^{*‡}

Aarti Singh[‡] Yuejie Chi[§] Yuxin Chen[†]

November 11, 2025

Abstract

The ability to reason lies at the core of artificial intelligence (AI), and challenging problems usually call for deeper and longer reasoning to tackle. A crucial question about AI reasoning is whether models can extrapolate learned reasoning patterns to solve harder tasks with longer chain-of-thought (CoT). In this work, we present a theoretical analysis of transformers learning on synthetic state-tracking tasks with gradient descent. We mathematically prove how the *algebraic structure* of state-tracking problems governs the degree of extrapolation of the learned CoT. Specifically, our theory characterizes the length generalization of transformers through the mechanism of *attention concentration*, linking the retrieval robustness of the attention layer to the state-tracking task structure of long-context reasoning. Moreover, for transformers with limited reasoning length, we prove that a *recursive self-training* scheme can progressively extend the range of solvable problem lengths. To our knowledge, we provide the first *optimization guarantee* that constant-depth transformers provably learn NC^1 -complete problems with CoT, significantly going beyond prior art confined in TC^0 , unless the widely held conjecture $\text{TC}^0 \neq \text{NC}^1$ fails. Finally, we present a broad set of experiments supporting our theoretical results, confirming the length generalization behaviors and the mechanism of attention concentration.

1 Introduction

Reasoning is central to artificial intelligence [CND⁺23, BMR⁺20, CTJ⁺21, LAD⁺22, GZA⁺23, TLI⁺23]. Transformer-based [VSP⁺17] large language models (LLMs) achieve state-of-the-art results on complex reasoning tasks via chain-of-thought (CoT) reasoning [WWS⁺22b, MMA⁺22, WWS⁺22a, CKB⁺21, RM21, KGR⁺22, MY22, ZSH⁺22], where the model generates intermediate steps before delivering a final answer. Recent frontier models such as OpenAI-o1 [Ope24] and DeepSeek-R1 [DA25] typically produce long CoT traces at inference time, often elicited via reinforcement learning and/or supervised fine-tuning (SFT) that distills from longer chains [QLZ⁺24, MCJ⁺24]. These advances have enabled improved performance on more challenging problems [KZA⁺24, XGZ⁺24, ZLP⁺24, GJKM25, JB25], yet the mechanisms and limitations underlying CoT reasoning remain poorly understood, posing fundamental theoretical challenges.

Theoretical studies on transformers with CoT have recently advanced along two fronts: expressiveness [FZG⁺23, LLZM24, MS23a, CPW24, MS23b] and statistical learnability [ABL⁺24, WLS22, SFH⁺24,

^{*}Equal contribution, ordering determined by coin flip. This is the full version of a paper that was published at NeurIPS 2025.

[†]Department of Statistics and Data Science, Wharton School, University of Pennsylvania. Email: {yuh42, yuxinc}@wharton.upenn.edu

[‡]Machine Learning Department, Carnegie Mellon University. Email: {zixinw@andrew, aarti@cs}.cmu.edu

[§]Department of Statistics and Data Science, Yale University. Email: yuejie.chi@yale.edu

KWLS25, HZCY24, PLG23, LSG⁺23]. On the expressiveness front, seminal work [FZG⁺23, LLZM24, MS23a] showed that constant-depth transformers without CoT behave as shallow circuits and are restricted to express the circuit complexity class TC^0 of constant computation depth, whereas with $O(L)$ CoT steps on inputs of length L , they can express log-depth circuits in NC^1 , which is a problem class conjectured to require inherently serial computation.¹ These results reveal that CoT reasoning equips transformers with the expressive power to solve **inherently sequential** problems. In stark contrast, there remains limited understanding of *how* transformers acquire such reasoning abilities during training. Prior optimization analyses [KS24, WZLZ24, HWL25, HLY25] were limited to simple, fully parallelizable tasks in TC^0 that do not require sequential reasoning. This leaves a substantial gap between what transformers can *express* in principle and what they can *learn* through training.

Another key question, motivated by the success of reasoning models, is whether large models can extrapolate their reasoning beyond the sequence lengths of the training data: a feature known as **length generalization**. To fully exploit the potential of long CoT reasoning, the model must harness *long-context ability* [KBA⁺24, YSL⁺25, LLY⁺25, YLJ⁺25], which can be severely affected by a phenomenon called *context rot*: namely, a phenomenon in which model performance degrades as the number of tokens in the context increases [HSK⁺24, HTH25, tea25]. For reasoning tasks, empirical evidence on length generalization of transformers is mixed [ZBB⁺22, LAG⁺22, JdDE⁺23, ZBL⁺23, ZAC⁺24, HBK⁺24, XL25], and several prior studies reported limited extrapolation despite strong in-distribution performance. Architectural choices, including positional encoding and attention variants, can influence generalization considerably [KPNR⁺23, SPHG24, LCS⁺25, AWA⁺22, SUV18, RSR⁺20, DLS⁺23]. On the theoretical front, previous work established existence or statistical guarantees, showing that transformers can, in principle, represent length-generalizing algorithms or achieve favorable sample complexity independent of the length of the CoT [MDA⁺24, AM24, HYB⁺24, GJB⁺25, JVB⁺25]. Nonetheless, it remains unclear whether transformers can provably learn to length-generalize reliably when trained with gradient-based optimization algorithms.

In light of the aforementioned gaps in prior literature, the current paper seeks to make progress towards addressing the following two fundamental questions:

Research Questions

1. Can transformers, trained via gradient descent (GD), learn CoT reasoning to solve problems requiring *inherently* sequential reasoning beyond TC^0 ?
2. Can the learned reasoning ability *generalize* to problems that require longer CoTs beyond the lengths of training data?

To address these questions in a theoretically tractable manner, we analyze a minimally viable transformer: a one-layer transformer block with softmax attention and a feed-forward network (FFN), trained by GD with *no positional encoding* (NoPE). We study this model on synthetic *state-tracking* tasks, namely LEGO [ZBB⁺22], which distill core LLM skills such as entity tracking, game-state updates, and code evaluation [KS23, MPS24]. This setup is tractable for analysis, while capturing the mechanisms needed for step-by-step computation via CoT. We analyze the training dynamics with CoT on two LEGO task families with distinct *algebraic* action structures: simply transitive group actions, and symmetry group actions. By tracking attention patterns throughout training, we demonstrate how reasoning capabilities emerge and how length generalization is governed by the structural properties of these tasks. More concretely, our main contributions are summarized as follows.

1. **Provable guarantee of learning CoT with length generalization in state-tracking.** We prove that for the LEGO state-tracking task, a one-layer NoPE transformer trained with GD provably solves constant-length problems using CoT reasoning. Moreover, the learned transformer directly generalizes to problems of substantially longer length, when the group actions in LEGO is *simply transitive*. Conversely, for the canonical action of the symmetry group S_n on \mathbb{Z}_n , the learned transformer generalizes only up

¹ For background on circuit complexity and a detailed review of expressiveness results, see Section 2 and [Vol99, AB06].

to constant-factor length. We identify an **attention concentration** mechanism that dictates step-wise retrieval depending on the task structure, leading to distinct length generalization behaviors.

2. **Recursive self-training provably extends the solvable reasoning length.** When length generalization is limited (for example, under symmetry group actions), we introduce a *self-training* curriculum that recursively trains the model on its own CoT traces, motivated by the empirical work [LCS⁺25]. We prove that this scheme can bootstrap the solvable problem length up to maximal allowable length after sufficient rounds of self-training, thus offering a theoretical guarantee of recursive self-improvement.
3. **Constant-depth transformers provably learn to solve problems beyond TC^0 via CoT.** The first two results further establish that the model can learn the solution of state-tracking problems for non-solvable groups, which is NC^1 -complete and lies outside TC^0 unless the widely held conjecture $\text{TC}^0 \neq \text{NC}^1$ in circuit complexity theory fails. Therefore, we provide *the first optimization guarantee* showing that a one-layer transformer learns to solve reasoning tasks beyond TC^0 with CoT, matching the expressivity result of [FZG⁺23, MS23a, LLZM24] for linear-CoT transformers.
4. **Empirical evidence on synthetic LEGO tasks that supports our theory.** We present a wide range of experiments based on our theoretical setup. Our results corroborate our predictions on length generalization for different group actions, showing a clear separation between the two algebraic structures in the theoretical setting. Moreover, we empirically demonstrate how recursive self-training effectively improves reasoning length. The attention concentration mechanism identified in our theory is also supported by our experimental findings.

1.1 Overview of Theoretical Results

To facilitate discussion, let us first briefly introduce the *state-tracking problem*. Given a group \mathcal{G} acting on a state space \mathcal{Y} , the goal is to compute the final state y_L by applying a sequence of group actions $g_i \in \mathcal{G}$ starting from an initial state $y_0 \in \mathcal{Y}$:

$$y_0 \xrightarrow{g_1} y_1 \xrightarrow{g_2} \dots \xrightarrow{g_{L'}} y_{L'} \dots \xrightarrow{g_L} y_L.$$

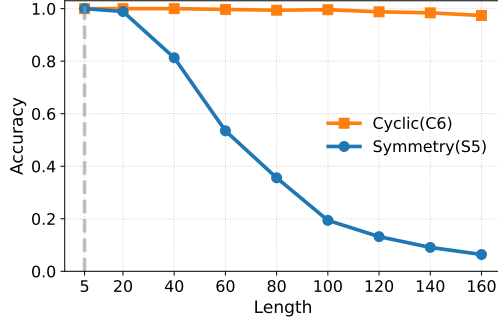
This task naturally lends itself to CoT reasoning, where intermediate steps explicitly track how the state evolves under successive actions.

We consider two types of group action: simply transitive actions and symmetry group actions [Fra23]. A simply transitive action is free and transitive; that is, there is *a unique* group element $g \in \mathcal{G}$ mapping any state to another. An example is a group acting on itself by the group composition. In contrast, in the symmetry case, the action is *transitive but not free*: for example, consider the canonical action of S_n on \mathbb{Z}_n by permutations, where *many* group elements send a given number i to j . In group theory, these two types of actions have different sizes of stabilizers, even when the group \mathcal{G} is the same.

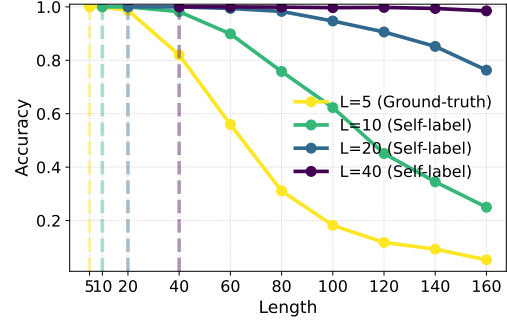
Result 1: Provable CoT learning for state tracking. Our first result shows that transformers can learn to solve these two state-tracking problems via CoT.

Theorem 1.1 (Learning CoT, informal). *One-layer transformers, trained via GD, can provably learn to solve state tracking problems for simply transitive and symmetry group actions via CoT reasoning.*

Since state-tracking problems for symmetry group are NC^1 -complete [Bar86], this result provides the first optimization-based training guarantee for solving problems beyond TC^0 , matching the linear-CoT expressiveness result of [LLZM24] with provable learnability under gradient-based training. In comparison, recent theoretical progress only showed that transformers can provably learn tasks in TC^0 such as parity [KS24, WZLZ24, HLY25] and linear regression [HWL25] with CoT, which, however, are solvable by log-precision, constant-depth transformers without CoT [LLZM24, MS23a].



(a) Length generalization results of cyclic (C_6) vs. symmetry (S_5) tasks.



(b) Self-improvement results on symmetry tasks (S_5).

Figure 1: Empirical results of length generalization on LEGO tasks with different group actions. (a). Transformers length-generalize to solve significantly longer CoT tasks for simply transitive (cyclic) group (Theorem 4.1), while generalizing poorly for symmetry group tasks. (b). When direct length generalization falls short for symmetry actions, a recursive self-training scheme that train on the model’s own longer CoT traces bootstraps the solvable problem length (Theorem 5.1). The dashed lines indicate the training length.

Result 2: Algebraic structure dictates length generalization. Going beyond training, our next result unveils the qualitative difference in length generalization induced by the difference in the group action structure. More importantly, we provide a mechanistic understanding of the length generalization property learned by gradient-based optimization.

Theorem 1.2 (Length generalization, informal). *The algebraic structure of the group actions in the state-tracking task dictates how far the reasoning length generalizes:*

- For simply transitive actions: *standard CoT training on constant-length problems already yields generalization to problems of length d^{c^*} , where $0 < c^* < 1$ is a constant.*
- For symmetry actions: *standard CoT training only generalizes to a constant factor of the training length.*

This provides a rigorous theory of length extrapolation for state-tracking: it characterizes when and how length generalization emerges for different reasoning problems. Formal statements are deferred to Section 4 (simply transitive actions) and Section 5 (symmetry group actions).

Our theory uncovers the **mechanism of attention concentration** that explains the varying degrees of length generalization. At a high level, the performance of long-context reasoning depends on the level at which the attention layer focuses on task-relevant contexts. The algebraic structure of the group action dictates the level of attention concentration at convergence: simply transitive actions enable sharp concentration and strong length generalization, whereas symmetry actions introduce distractors that dilute attention focus, limiting length generalization to constant factors.

Result 3: Self-training extends solvable problem length. Although length generalization is limited for symmetry actions, recent empirical studies [SCRA+23, GPS+23, LCS+25] have shown that neural networks can bootstrap their capability via *self-improvement*. In this work, we prove that a one-layer transformer obtained via a self-training scheme similar to [LCS+25], can extend its solvable problem length. To the best of our knowledge, this offers the first optimization guarantee of self-improvement for transformer networks.

Theorem 1.3 (Recursive self-improvement, informal). *A one-layer transformer, trained with a recursive self-training scheme, can self-improve: at each stage $k > 1$, the model learning on the self-labeled reasoning traces of length 2^k , length generalizes to solve the next doubled length 2^{k+1} . Consequently, after $\Theta(\log d)$ stages, the model can solve problems of length d , the maximal length in our setting.*

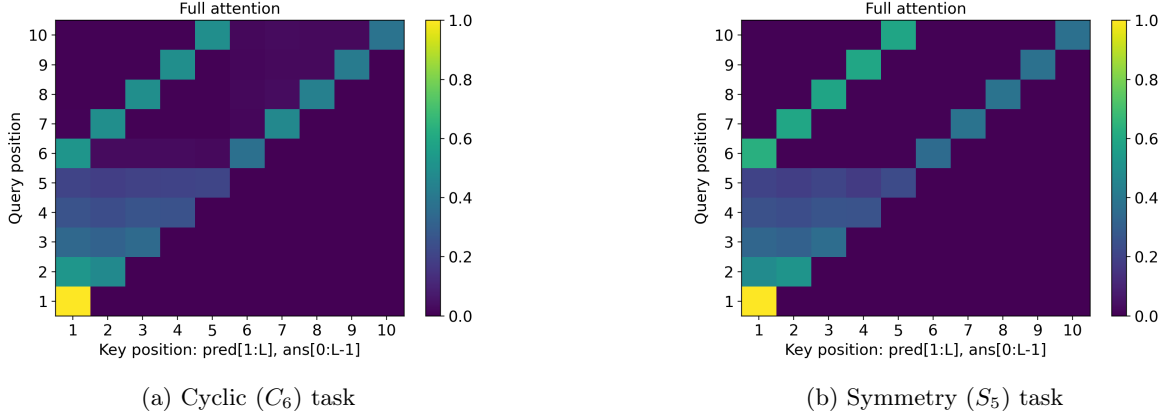


Figure 2: Attention concentration at convergence for LEGO task with length $L = 5$. The heatmap places the query clause index on the y -axis (keys on the x -axis). For a task of length L , the LEGO sequence prior to the final answer clause has length $2L$; we focus on query positions $L + 1$ to $2L$, corresponding to answer clauses $Z_{\text{ans},0}$ to $Z_{\text{ans},L-1}$. Two diagonal bands in the upper region indicate attention concentrating on the answer clause $Z_{\text{ans},\ell}$ and the predicate clause $Z_{\text{pred},\ell+1}$ when the query is $Z_{\text{ans},\ell}$.

Our work complements recent advances on length generalization in transformers. Prior studies [MDA⁺24, AM24, HYB⁺24, GJB⁺25, JVB⁺25] primarily established statistical guarantees and *non-gradient-based* learnability. Notably, [GJB⁺25] identifies conditions under which specialized positional encodings and sparse contextual dependencies enable length extrapolation, provided the model fits the source length. [JVB⁺25] analyzed time-invariant autoregressive models (a fixed next-token generator) and showed that the sample complexity can be independent of the CoT length but on the complexity of individual CoT steps. We provide a concrete setting that mirrors these works and, crucially, show that gradient-based optimization *actually finds* such solutions for transformer networks.

1.2 Our Empirical Results at a Glance

Empirically, we conduct a series of experiments aligned with our problem setup to test the theory’s predictions, and observe strong agreements with our predicted behavior. In particular:

- *Length generalization comparison.* Figure 1a shows that training at a short length yields nearly perfect accuracy at much longer lengths for tasks with simply transitive actions, whereas tasks with symmetry actions exhibit only constant-factor length generalization. These findings are consistent with Theorem 1.2.
- *Effectiveness of recursive self-training.* In Figure 1b, a double-and-self-labeled curriculum for the task with symmetry actions consistently shifts the length–accuracy curve rightward; after several stages, performance matches the simply transitive case, validating Theorem 1.3.
- *Attention concentration patterns.* At convergence, the attention heatmaps in Figure 2 exhibit two clear diagonals with concentrated mass, confirming the attention-concentration mechanism predicted by our theory.

Detailed descriptions and discussions, along with additional experiments, are postponed to Section 7.

2 Background

Computational complexity [AB06] has been employed to characterize the power of neural networks [GLL87, SS92]. Historically, *circuit complexity* has been used extensively to study the power of neural networks

[STAK92, Par94, Maa96, ŠO03, MS25a, MS23b, WNB⁺25], due to the structural resemblance of Boolean circuits and neural networks with threshold gates. In a nutshell, circuit complexity evaluates computation models by the size, depth, and gate types of Boolean circuits that implement them. Below we provide a brief introduction of the basics of circuit complexity.

2.1 Circuits and Expressiveness

A Boolean circuit is a finite acyclic network of logic gates that computes a Boolean function on $\{0, 1\}^n \rightarrow \{0, 1\}$ for some fixed n . The gates with fan-in 0 are the inputs, which are assigned one of the n Boolean variables. A *circuit family* $\{C_n\}_{n \geq 1}$ computes a language by using C_n on length- n inputs. We measure the *size* of a circuit by the number of gates, and *depth* by the length of the longest input-output path, respectively. For instance, “constant depth” means $\text{depth}(C_n) = O(1)$, while “polynomial size” means $\text{size}(C_n) = O(n^c)$ for some integer c . Standard circuit complexity classes are defined by restricting circuit depth, size, gate set and fan-in:

- TC^0 consists of *constant-depth*, polynomial-size circuits with unbounded fan-in $\{\text{AND}, \text{OR}, \text{NOT}\}$ gates augmented with threshold (e.g., MAJORITY) gates. The circuit class TC^0 captures exactly the complexity of integer multiplication and division, and sorting [CSV84].
- NC^1 consists of *log-depth*, polynomial-size circuits over $\{\text{AND}, \text{OR}, \text{NOT}\}$ with bounded fan-in. This class captures exactly the complexity of recognizing all regular languages.
- P/poly is the class of all languages computable by polynomial-size circuit families.

These classes satisfy the following relations:²

$$\text{TC}^0 \subseteq \text{NC}^1 \subseteq \text{P/poly}.$$

Expressiveness gap of transformers via circuit complexity. One can analyze the expressivity of transformer neural nets through the lens of circuit complexity. Prior work [MSS22, LAG⁺22, MS23b, FZG⁺23, LLZM24] showed that the expressive upper bound of a vanilla constant-depth transformer is limited to TC^0 , the class of problems solvable by extremely shallow, highly parallel circuits. In contrast, when equipped with CoT, recent work [FZG⁺23, MS23a, LLZM24] proved that $O(n)$ intermediate steps enable transformers to simulate NC^1 -complete language. Further, [LLZM24] showed that 2-layer transformers with polynomially many CoT steps suffice to express arbitrary polynomial-size circuits, i.e., P/poly . Hence, under the widely believed conjecture that $\text{TC}^0 \neq \text{NC}^1$ [Vol99], CoT strictly extends the expressive power of transformers beyond TC^0 , allowing them to solve problems that inherently require super-constant (here $O(\log n)$) computation depth. A standard example connecting algebraic structure with circuit complexity is the following *word problem*, define by Dehn [Deh11, Deh87].

Definition 2.1 (Word problem for a group G). Let G be a group and let e denote its identity. For a word $w = g_1 \cdots g_k \in G^*$ (each $g_i \in G$), the word problem asks to decide whether $g_1 \circ \cdots \circ g_k = e$.

Barrington [Bar86] proved that the above word problem is NC^1 -complete when the group is non-solvable.

Theorem 2.1 (Barrington [Bar86]). *The word problem of every finite non-solvable group is NC^1 -complete. The canonical example is S_n for $n \geq 5$, the symmetry group on n elements that encodes the permutations.*

The word problem captures a fundamental reasoning task: state tracking in a finite world [MPS24, MS25b, LGA25]. Given an initial state and a sequence of transformations, the goal of the task is to compute the resulting state. State tracking underlies practical LLM abilities such as narrative entity tracking, chess move analysis, and code execution [KS23, MPS24], while exhibiting a rich connection to the circuit complexity theory. Therefore, it has become a standard synthetic testbed for probing the reasoning abilities of language models, both theoretically [LAG⁺22, MPS24, MS25b] and empirically [ZBB⁺22, LAG⁺22, LGA25]. Motivated by this connection, we study the state-tracking task with varying algebraic structures.

² We intentionally omit uniformity conventions here as this section is only meant to give readers a sense of the classes we refer to and their standard relationships.

2.2 LEGO for State Tracking

We focus on a specific formulation of the state-tracking problem, LEGO (*Learning Equality and Group Operations*) [ZBB⁺22], which was originally proposed as a synthetic task to study the reasoning behavior of transformers empirically. A typical LEGO instance in [ZBB⁺22] takes the following form:

$$\mathbf{b} = + \mathbf{a}, \quad \mathbf{c} = - \mathbf{b}, \quad \dots, \quad \mathbf{t} = - \mathbf{s}, \quad \mathbf{a} = -1, \quad \dots$$

Here, $\mathbf{a}, \mathbf{b}, \mathbf{c}, \dots$ are **variables**, each taking a **value** (or state) in $\{-1, +1\}$ in this example. Short expressions such as $\mathbf{b} = + \mathbf{a}$ are **clauses**, where $= +$ and $= -$ denote **actions**: the action is applied to the *right-hand-side* variable's value to obtain the *left-hand-side* variable's value. For instance, from $\mathbf{b} = + \mathbf{a}$ and $\mathbf{a} = -1$, it follows that $\mathbf{b} = -1$. Formally, the LEGO language is defined as follows:

Definition 2.2 (LEGO language [ZBB⁺22]). Let $\mathcal{X}, \mathcal{G}, \mathcal{Y}$ be finite sets of variables, actions, and values, respectively, where each $g \in \mathcal{G}$ is a map $g : \mathcal{Y} \rightarrow \mathcal{Y}$. The formal language **LEGO**($\mathcal{X}, \mathcal{G}, \mathcal{Y}$) has alphabet $\mathcal{X} \cup \mathcal{G} \cup \mathcal{Y} \cup \{=, (,)\}$ and consists of two types of expressions (called **clauses**):

- (1) *Predicate clause* $x = g(x')$ specifies an action $g \in \mathcal{G}$ linking variables $x, x' \in \mathcal{X}$.
- (2) *Answer clause* $x = y$ assigns a value $y \in \mathcal{Y}$ to a variable $x \in \mathcal{X}$.

A canonical LEGO sentence of length L with answer up to L' concatenates predicate clauses $x_n = g_n(x_{n-1})$ for $n \in [L]$ and answer clauses $x_n = y_n$ for $n \in [L']$ with $L' \leq L$:

$$\underbrace{x_1 = g_1(x_0) \dots x_L = g_L(x_{L-1})}_{\text{predicates}} \underbrace{x_0 = y_0 \dots x_{L'} = y_{L'}}_{\text{answers}}, \quad (1)$$

which describes the chain of transitions:

$$x_0 \xrightarrow{g_1} x_1 \xrightarrow{g_2} x_2 \xrightarrow{g_3} \dots \xrightarrow{g_{L'}} x_{L'} \dots \xrightarrow{g_L} x_L, \quad \text{starting with } x_0 = y_0.$$

with answers $y_1, \dots, y_{L'}$ up to L'

For semantic validity, any sentence containing a path $x_n = g_n(x_{n-1}), \dots, x_{n-k+1} = g_{n-k+1}(x_{n-k})$ for $k \in [n]$ must satisfy: $y_n = g_n \circ g_{n-1} \circ \dots \circ g_{n-k+1}(y_{n-k})$.

In the LEGO language, predicate clauses encode transformations, while answer clauses encode observed states. Thus, solving a LEGO sentence is exactly state tracking: compose the listed actions along the path and propagate the observed states to predict the next answer consistent with the composition.

3 Problem Formulation

As introduced in Section 2, the current paper employs the LEGO framework to investigate the reasoning capabilities of transformers. To set the stage, this section presents precise mathematical formulations of the problems to be studied in this paper. Before proceeding, we introduce the following notation:

- *Vocabulary*. Define the vocabulary as $\mathcal{V} := \mathcal{X} \cup \mathcal{G} \cup \mathcal{Y} \cup \{\langle \text{blank} \rangle\}$, where the *blank* token $\langle \text{blank} \rangle$ is a null symbol indicating the absence of other tokens.
- *Vocabulary size*. Let $d := |\mathcal{V}|$ denote the (finite) vocabulary size.

To facilitate theoretical analysis, we concentrate on the asymptotic regime where $d \rightarrow \infty$. Note that $|\mathcal{X}|, |\mathcal{G}|, |\mathcal{Y}|$ may depend on d , and we shall specify any required scaling assumptions as needed.

Assumption 3.1 (Asymptotic regime). For a language **LEGO**($\mathcal{X}, \mathcal{G}, \mathcal{Y}$) defined in Definition 2.2, we consider the asymptotic regime where both the vocabulary size d and the number of variables $|\mathcal{X}|$ tend to infinity. Assume $|\mathcal{G}| \leq \log^{C_0} d$ for some constant $C_0 \in [1, 100)$, and hence \mathcal{Y} and \mathcal{G} are much smaller than \mathcal{X} in size.

3.1 Data Distribution

We begin by specifying how LEGO clauses are tokenized, followed by a definition of the LEGO distribution.

Definition 3.1 (LEGO encoding). Each LEGO clause from Definition 2.2 is encoded as a fixed-length, 5-token tuple $Z \in \mathcal{V}^5$. More specifically,

- For each predicate clause $x = g(x')$, set $Z_{\text{pred}} := (x, g, x', \langle \text{blank} \rangle, \langle \text{blank} \rangle) \in \mathcal{V}^5$;
- For each answer clause $x = y$, set $Z_{\text{ans}} := (\langle \text{blank} \rangle, \langle \text{blank} \rangle, \langle \text{blank} \rangle, x, y) \in \mathcal{V}^5$.

With Definition 3.1 in place, we can encode a LEGO sentence (1) to a sequence $Z^{L,L'}$:

$$Z^{L,L'} = (Z_{\text{pred},1}, \dots, Z_{\text{pred},L}, Z_{\text{ans},0}, \dots, Z_{\text{ans},L'}), \quad (2)$$

where $Z_{\text{pred},k}$ and $Z_{\text{ans},k}$ represent the k -th predicate and answer clauses, respectively. For convenience, we denote $\mathcal{I}^{L,L'} = \{(\text{pred}, \ell)\}_{\ell \in [L]} \cup \{(\text{ans}, \ell)\}_{\ell=0}^{L'}$ as the index set associated with the clauses in $Z^{L,L'}$. If $L = L'$, we write the sequence simply as Z^L and the corresponding index set as \mathcal{I}^L .

To feed LEGO tokens into a neural network, we first map each symbol to an integer index (*tokenization*), and then map indices to continuous vectors via a learned table (*embedding*). The following definitions formalize these two steps.

Definition 3.2 (Tokenization and token embedding). Each $v \in \mathcal{V}$ is assigned a unique index $\tau(v) \in \{1, \dots, d\}$. Denote by $e_i \in \mathbb{R}^d$ the embedding vector associated with index i , and write $e_v \equiv e_{\tau(v)}$ for convenience. The blank token is assigned the zero vector, $e_{\tau(\langle \text{blank} \rangle)} = \mathbf{0}_d \in \mathbb{R}^d$. For technical simplicity, we assume that $\{e_v : v \in \mathcal{V} \setminus \{\langle \text{blank} \rangle\}\}$ forms an orthonormal set in \mathbb{R}^d (note that this assumption can be relaxed to a well-conditioned embedding matrix without affecting our results).

Equipped with Definition 3.2, we can transform the LEGO sequence encoding into vector embeddings that can be used as inputs to neural networks.

Definition 3.3 (Embedding of LEGO sentences). Let $d_c := 5d$ be the clause embedding dimension. We define an operator $\text{Embed} : \mathcal{V}^5 \rightarrow \mathbb{R}^{d_c}$ that maps a clause to embedding by

$$\mathbf{Z} = \text{Embed}(Z) := (e_{v_1}, e_{v_2}, \dots, e_{v_5}) \in \mathbb{R}^{d_c}, \quad \text{for clause } Z = (v_1, v_2, \dots, v_5) \in \mathcal{V}^5.$$

Specifically, a LEGO sentence $Z^{L,L'}$ defined in (2) is embedded as

$$\mathbf{Z}^{L,L'} = (\mathbf{Z}_{\text{pred},1}, \dots, \mathbf{Z}_{\text{pred},L}, \mathbf{Z}_{\text{ans},0}, \dots, \mathbf{Z}_{\text{ans},L'}) \in \mathbb{R}^{d_c \times (L+L'+1)}$$

where each column $\mathbf{Z}_{\text{pred},\ell} \in \mathbb{R}^{d_c}$ (resp. $\mathbf{Z}_{\text{ans},\ell}$) is the embedding of clause $Z_{\text{pred},\ell}$ (resp. $Z_{\text{ans},\ell}$). When $L' = L$, we simply write $\mathbf{Z}^L \equiv \mathbf{Z}^{L,L'}$ for simplicity.

Next, we describe the distribution that governs the generation of a LEGO sentence.

Assumption 3.2 (LEGO distribution $\mathcal{D}^L, \mathcal{D}^{L,L'}$). Consider $\mathbf{LEGO}(\mathcal{X}, \mathcal{G}, \mathcal{Y})$ as defined in Definition 2.2, and let L denote the sequence length. We assume that the distribution \mathcal{D}^L of length- L LEGO sentences satisfies the following properties.

1. All LEGO sentences $Z^L \sim \mathcal{D}^L$ of the form (1) are encoded by Definition 3.1 into representation (2).
2. The variables $x_0, x_1, \dots, x_L \in \mathcal{X}$ are sampled uniformly at random from \mathcal{X} without replacement.
3. The first value $y_0 \in \mathcal{Y}$ is chosen uniformly at random from \mathcal{Y} .
4. The actions $g_1, g_2, \dots, g_L \in \mathcal{G}$ are sampled uniformly at random from \mathcal{G} with replacement.
5. The intermediate values y_1, y_2, \dots, y_L are computed recursively by $y_i = g_i(y_{i-1})$.

For any $L' < L$, we define the truncated distribution $\mathcal{D}^{L,L'}$ of sequences $Z^{L,L'}$ containing all the predicates and the first $L' + 1$ many answer clauses, where $Z^{L,L'}$ is obtained by first sampling $Z^L \sim \mathcal{D}^L$ and then removing the answer clauses $Z_{\text{ans},\ell}$ for all $\ell > L'$.

As can be easily seen, sequences sampled from \mathcal{D}^L or $\mathcal{D}^{L,L'}$ correspond to valid LEGO sentences as defined in Definition 2.2. With a slight abuse of notation, we write $\mathbf{Z}^{L,L'} \sim \mathcal{D}^{L,L'}$ to indicate that $\mathbf{Z}^{L,L'}$ is the embedding of a sentence $Z^{L,L'}$ drawn from $\mathcal{D}^{L,L'}$.

3.2 Transformer Architecture

In this subsection, we introduce the transformer architecture investigated in this paper. Towards this end, we first introduce a smoothed activation function that will be used in our network.

Definition 3.4 (Smooth ReLU). Define a continuously differentiable variant of the ReLU activation function [AZL20, HLZ⁺22] as follows:

$$\text{sReLU}(x) := \begin{cases} \frac{\varrho}{q}, & x \leq -\varrho, \\ \frac{x^q}{\varrho^{q-1}q}, & x \in (-\varrho, \varrho], \\ x - \varrho\left(1 - \frac{1}{q}\right), & x > \varrho, \end{cases}$$

where q and ϱ are some design parameters. Here and throughout, we choose $q = O(1)$ to be a large, even integer, and take $\varrho = \Theta(1/\text{polylog}(d))$.

Transformer architecture. In this work, we focus on an autoregressive transformer [VSP⁺17] whose block consists of a softmax attention layer followed by a position-wise feed-forward network, as described below.

- *Attention layer.* Given LEGO sentence embeddings $\mathbf{Z}^{L,L'}$ and indices $j, k \in \mathcal{I}^{L,L'}$, the attention from clause \mathbf{Z}_j to clause \mathbf{Z}_k is defined through the softmax operator as

$$\text{Attn}_{j \rightarrow k}(\mathbf{Q}, \mathbf{Z}^{L,L'}) := \frac{\exp(\mathbf{Z}_j^\top \mathbf{Q} \mathbf{Z}_k)}{\sum_{r \in \mathcal{I}^{L,L'}} \exp(\mathbf{Z}_j^\top \mathbf{Q} \mathbf{Z}_r)}.$$

Since the model is autoregressive, a standard causal mask is applied to ensure that the latest (answer) token attends only to *preceding* tokens (including itself). The attention output is

$$\text{Attention}(\mathbf{Q}, \mathbf{Z}^{L,L'}) := \sum_{k \in \mathcal{I}^{L,L'}} \text{Attn}_{\text{ans}, L' \rightarrow k}(\mathbf{Q}, \mathbf{Z}^{L,L'}) \cdot \mathbf{Z}_k.$$

Note that in the standard formulation of transformers, the score takes the form $\mathbf{Z}_j^\top \mathbf{W}^Q \mathbf{W}^K \mathbf{Z}_k$ instead for a “query” parameter matrix \mathbf{W}^Q and a “key” parameter matrix \mathbf{W}^K . Here, we fold $\mathbf{W}^Q \mathbf{W}^K$ into a single parameter matrix \mathbf{Q} , resulting in a score $\mathbf{Z}_j^\top \mathbf{Q} \mathbf{Z}_k$, which can be viewed as an equivalent reparameterization that simplifies analysis without changing expressivity [HCL24, HWCL25, YHLC25].

- *Feed-forward network (FFN).* The FFN with parameter $\mathbf{W} \in \mathbb{R}^{5 \times d \times m \times d_c}$ is defined by

$$\text{FFN}_{i,j}(\mathbf{W}, \mathbf{X}) := \sum_{r \in [m]} \text{sReLU}(\langle \mathbf{W}_{i,j,r}, \mathbf{X} \rangle + b_{i,j,r}), \quad \forall i \in [5], j \in [d],$$

where \mathbf{X} denotes the input, $\mathbf{W}_{i,j,r} \in \mathbb{R}^{d_c}$ are neuron weights, m indicates the number of neurons, and $b_{i,j,r} \in \mathbb{R}$ is some *fixed* bias.

With the above layers defined, we are ready to introduce the transformer model used in this work, which is summarized as follows.

Definition 3.5 (Transformer language model). Assume that our learner neural network F is a one-layer decoder transformer block composed of an attention layer with no positional encoding (NoPE) as well as an FFN layer: $F = \text{FFN} \circ \text{Attention}$, with parameters $\mathbf{W} \in \mathbb{R}^{5 \times d \times m \times d_c}$ and $\mathbf{Q} \in \mathbb{R}^{d_c \times d_c}$. Formally, for the i -th token position and the j -th vocabulary index, we have

$$\left[F_i(\mathbf{Z}^{L,L'}) \right]_j := \text{FFN}_{i,j}(\mathbf{W}, \text{Attention}(\mathbf{Q}, \mathbf{Z}^{L,L'})) \in \mathbb{R}, \quad \forall i \in [5], j \in [d]. \quad (3)$$

We interpret $F(\mathbf{Z}^\ell)$ as five logit vectors $\{F_i(\mathbf{Z}^\ell)\}_{i=1}^5 \subset \mathbb{R}^d$, each parameterizing the distribution of the i -th token of the next clause. Recall that \mathcal{V} is the vocabulary with $|\mathcal{V}| = d$ and $\tau : \mathcal{V} \rightarrow [d]$ denotes the index map. Given an encoded LEGO sequence $Z^\ell = (Z_1, \dots, Z_\ell)$ with embedding \mathbf{Z}^ℓ , the model's predictive distribution for the i -th token of the $(\ell + 1)$ -th clause is the following softmax:

$$p_{F_i}(Z_{\ell+1,i} = v \mid Z_1, \dots, Z_\ell) = \frac{e^{[F_i(\mathbf{Z}^\ell)]_{\tau(v)}}}{\sum_{j \in [d]} e^{[F_i(\mathbf{Z}^\ell)]_j}}, \quad \forall v \in \mathcal{V}. \quad (4)$$

Now we can sample the next clause $Z_{\ell+1}$ by sampling from the product distribution

$$Z_{\ell+1} = (Z_{\ell+1,1}, \dots, Z_{\ell+1,5}) \sim \bigotimes_{i=1}^5 p_{F_i} =: p_F,$$

in an autoregressive manner.

3.3 LEGO Task via CoT Reasoning and Training Objective

With the LEGO distribution and the transformer model in place, we now formalize the task within the LEGO framework. We view solving a length- L LEGO problem as generating a sequence of CoT steps, where each step ℓ produces an intermediate result used to predict $x_\ell = y_\ell$, ultimately leading to the final solution $x_L = y_L$. To be precise, we define the following reasoning task.

Definition 3.6 (Reasoning tasks \mathcal{T}^L). We define a family $\{\mathcal{T}^L\}_{L \in \mathbb{N}^+}$ that captures the ability to solve sequential reasoning problems. For each $L \geq 1$, task \mathcal{T}^L measures the model's *accuracy along the chain* from step 1 to step L :

$$\text{Acc}_L(F) = \frac{1}{L} \sum_{0 \leq L' < L} \mathbb{E}_{Z^L \sim \mathcal{D}^L} \left[\mathbb{E}_{\hat{Z}_{\text{ans}, L'+1} \sim p_F(\cdot \mid Z^{L,L'})} [\mathbb{1}\{\hat{Z}_{\text{ans}, L'+1} = Z_{\text{ans}, L'+1}\}] \right], \quad (5)$$

where p_F is induced by the model F from (4).

Clearly, $\text{Acc}_L(F) \in [0, 1]$. We say that task \mathcal{T}^L is solved if $\text{Acc}_L(F) \approx 1$. As L grows, $\{\mathcal{T}^L\}_{L \in \mathbb{N}^+}$ poses increasingly difficult state tracking challenges. At step L' , the model conditions on the partial transcript $Z^{L,L'}$ and predicts the next answer $Z_{\text{ans}, L'+1}$. To enforce step-by-step generation of the CoT trace, we define the following training objective.

Definition 3.7 (Next clause loss). The training objective for \mathcal{T}^L with $L \geq 1$ is the *next clause loss*

$$\text{Loss}^L(F) := \sum_{1 \leq L' \leq L} \text{Loss}^{L,L'}(F), \quad (6a)$$

$$\text{where } \text{Loss}^{L,L'}(F) := \mathbb{E}_{Z^{L,L'} \sim \mathcal{D}^{L,L'}} [-\log p_F(Z_{\text{ans}, L'} \mid Z^{L,L'-1})]. \quad (6b)$$

Here, p_F is induced by the model F from (4). We also define the per-token loss

$$\text{Loss}_i^{L,L'}(F) := \mathbb{E}_{Z^{L,L'} \sim \mathcal{D}^{L,L'}} [-\log p_{F_i}(Z_{\text{ans}, L', i} \mid Z^{L,L'-1})].$$

This is a teacher forcing style CoT objective: the model is given the ground truth answers for all previous steps and is trained to match the next step’s answer [HSY22, KS24].

Further, we adopt the following initialization for training.

Assumption 3.3 (Initialization). Let F be the transformer network in Definition 3.5 with parameters \mathbf{W} and \mathbf{Q} . The attention parameter is initialized to zero: $\mathbf{Q}^{(0)} = \mathbf{0}_{d_c \times d_c}$. The FFN weights are initialized randomly and independently as $\mathbf{W}_{i,j,r}^{(0)} \sim \mathcal{N}(\mathbf{0}, \sigma_0^2 \mathbf{I}_d)$ with $\sigma_0 = d^{-1/2}$. The biases are not trained and instead fixed at $b_{i,j,r} = \sigma_0 \log d$ for all i, j, r , chosen to keep most sReLU units active at initialization. All random draws are independent across indices.

Additional Technical Assumptions. Additionally, we introduce a couple of technical assumptions to be used throughout the proof. To control rare large deviations in the logits during training-time analysis, we adopt a bounded-output assumption as stated below.

Assumption 3.4 (Logit clipping). There exists $B = C_B \log d$, for some sufficiently large constant $C_B > 0$, such that each coordinate of the raw model output F_i is clipped from above:

$$[F_i]_j \leftarrow \min\{[F_i]_j, B\} \quad \text{for all } i, j.$$

This coordinate-wise clipping is a technical device to control large-deviation tails and simplify the analysis of the dynamics; B can be chosen large enough to avoid interfering with the regimes we study.

Moreover, to simplify the analysis of attention dynamics, we impose a fixed block-sparsity pattern on the attention parameter \mathbf{Q} .

Assumption 3.5 (Block-sparse attention matrix). Let $\mathbf{Q} = [\mathbf{Q}_{p,q}]_{p,q \in [5]} \in \mathbb{R}^{5d \times 5d}$ be partitioned into 5×5 blocks with $\mathbf{Q}_{p,q} \in \mathbb{R}^{d \times d}$. We assume that

$$\mathbf{Q}_{p,q} \equiv \mathbf{0}_{d \times d} \quad \text{for all } (p, q) \notin \{(4, 3), (4, 4)\},$$

i.e., only the blocks $(4, 3)$ and $(4, 4)$ are trainable.

This block-sparsity pattern, which zeros out most inter-token attention, is standard in recent theoretical analyses of transformer training dynamics [HCL24, HWCL25, YHLC25, YHLC24, CHX⁺25]. Importantly, although sparse at the 5×5 token level, the two retained blocks $(4, 3)$ and $(4, 4)$ are fully *dense* $d \times d$ matrices trained without constraints, leaving $2d^2$ free parameters and thus a substantive, non-trivial learning problem.

4 Learning CoT on Simply Transitive Actions

In this section, we present our first main results for the case where the group action \mathcal{G} on \mathcal{Y} is simply transitive, which is isomorphic to the action of the cyclic group C_n on \mathbb{Z}_n .

Assumption 4.1 (Simply transitive group action). Let $\mathcal{Y} = \{0, 1, \dots, n_y - 1\}$ with $n_y \in [\Omega(\log \log d), \log d]$. Assume the group \mathcal{G} acts simply transitively on \mathcal{Y} : the action is transitive and for any $y_1, y_2 \in \mathcal{Y}$ there exists a unique $g \in \mathcal{G}$ such that $g \cdot y_1 = y_2$. Equivalently, for any fixed $y \in \mathcal{Y}$ the map $g \mapsto g \cdot y$ is a bijection from \mathcal{G} to \mathcal{Y} , so $|\mathcal{G}| = |\mathcal{Y}| = n_y$.

Our first main result demonstrates that, for such a simple task the model obtained via Algorithm 1 for short-chain tasks \mathcal{T}^1 and \mathcal{T}^2 , successfully generalizes to significantly longer tasks.

Theorem 4.1. *Under Assumptions 3.1, 3.2, 3.3, 3.4, 3.5 and 4.1, for some constant $0 < c^* < 1$, $n_y < m \ll \log^2 d$, the transformer model $F^{(T_1+T_2)}$ obtained by Algorithm 1 with learning rate $\eta = \frac{1}{\text{poly}(d)}$, and stage 1 and 2 iteration $T_1 = \tilde{O}\left(\frac{d}{\eta(\sigma_0)^{q-2}}\right)$, $T_2 = \tilde{O}\left(\frac{\text{poly}(d)}{\eta\sigma_0}\right)$ satisfies*

Algorithm 1: Curriculum training for simply transitive actions

Input: Model $F^{(0)}$ with parameters $(\mathbf{W}^{(0)}, \mathbf{Q}^{(0)})$; Learning rate η ; Stage snapshots T_1, T_2 .

Stage 1: Learning one-step reasoning (\mathcal{T}^1);

for $t = 1$ **to** T_1 **do** // Update the FFN parameter \mathbf{W}

$\mathbf{W}^{(t)} \leftarrow \mathbf{W}^{(t-1)} - \eta \nabla_{\mathbf{W}} \text{Loss}^1(F^{(t-1)})$;

$\mathbf{Q}^{(t)} \equiv \mathbf{Q}^{(t-1)}$;

Stage 2: Learning two-step reasoning for length extension (\mathcal{T}^2);

for $t = T_1 + 1$ **to** T_2 **do** // Update the attention parameter \mathbf{Q}

$\mathbf{Q}^{(t)} \leftarrow \mathbf{Q}^{(t-1)} - \eta \nabla_{\mathbf{Q}} \sum_{\ell=1}^2 \text{Loss}_5^{2,\ell}(F^{(t-1)})$;

$\mathbf{W}^{(t)} \equiv \mathbf{W}^{(t-1)}$;

Output: Model $F^{(T_1+T_2)}$.

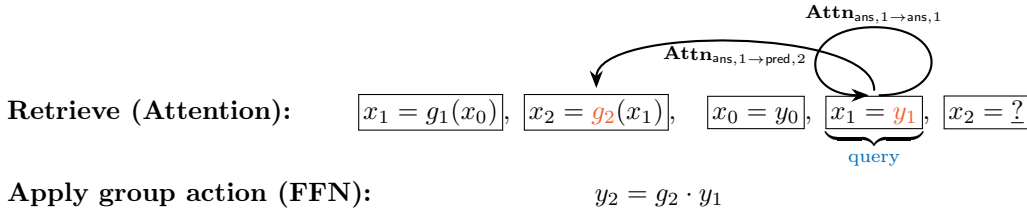


Figure 3: Illustration of how the model solves the LEGO task: given $Z^{2,1}$, the goal is to predict y_2 .

1. Direct short-to-long length generalization:

$$\text{Acc}_L\left(F^{(T_1+T_2)}\right) \geq 1 - \frac{1}{\text{poly}(d)}, \text{ for every } L \leq O(d^{c^*}), \quad (7)$$

i.e., $F^{(T_1+T_2)}$, which is trained for task \mathcal{T}^1 and \mathcal{T}^2 , generalizes to solve the tasks $\mathcal{T}^\ell, \ell \leq L$.

2. Attention concentration:³ given $Z^{2,\ell}$ with $\ell \in \{0, 1\}$, we have

$$\text{Attn}_{\text{ans}, \ell \rightarrow \text{pred}, \ell+1}^{(T_1+T_2)} + \text{Attn}_{\text{ans}, \ell \rightarrow \text{ans}, \ell}^{(T_1+T_2)} \geq 1 - O\left(\frac{1}{d^{c^*}}\right). \quad (8)$$

Mechanism of CoT for state-tracking. Given the current sequence $Z^{L,\ell}$ with intermediate steps up ℓ , predicting the next state $y_{\ell+1} = g_{\ell+1}(y_\ell)$ requires two steps:

- (i) **Retrieve** the correct action $g_{\ell+1}$ from the context clause $Z_{\text{pred}, \ell+1}$ and the current state y_ℓ from the answer clause $Z_{\text{ans}, \ell}$;
- (ii) **Apply the group action**, that is, compute the next state $y_{\ell+1} = g_{\ell+1}(y_\ell)$.

It is well established that attention can implement content-based retrieval [ENO⁺21], and that FFN can represent the group operation [LPW⁺17] (see Figure 3 for an example). Algorithm 1 decouples learning the attention (retrieval) and FFN (action) components to simplify the analysis. For task \mathcal{T}^1 , the transcript $Z^{1,0}$ contains only the two relevant clauses, $Z_{\text{pred}, 1}$ and $Z_{\text{ans}, 0}$, without useless contents. Fixed uniform attention (\mathbf{Q} initialized to be zero in Assumption 3.3) therefore suffices to expose both clauses to the FFN, and we optimize the FFN to learn the group operation. The high accuracy for \mathcal{T}^1 in Theorem 4.1 indicates that the

³ For readability, we abbreviate $\text{Attn}_{\mathbf{k} \rightarrow \mathbf{k}'}(\mathbf{Q}^{(t)}, \mathbf{Z})$ by $\text{Attn}_{\mathbf{k} \rightarrow \mathbf{k}'}^{(t)}$, omitting explicit dependence when immaterial.

FFN has indeed *learned* to apply the operation correctly. For task \mathcal{T}^2 , with the FFN already trained, the attention layer only needs to learn to route the correct context to the FFN input. The attention concentration result in (8) confirms that the learned routing pattern is correct. Figures 2 and 5 also demonstrate the attention concentration patterns empirically.

How does attention concentration induce strong length generalization? As we increase the chain length in \mathcal{T}^ℓ for $\ell > 2$, the FFN layer remains largely insensitive to input length since the learned group action is location-invariant. By contrast, the attention layer is affected: more *irrelevant* clauses appear, so retrieval must scan over longer contexts, which risks diluting attention on the relevant clause. Theorem 4.1 guarantees that training on short chains already yields attention concentration with error $O(d^{-c^*})$. This “purity” allows the model to tolerate dilution and maintain high attention on the relevant clauses for chain lengths up to $O(d^{c^*})$. Technically, this concentration arises because the query matrix \mathbf{Q} learns to locate the same variable x_ℓ that appears simultaneously: the third token of the context clause $Z_{\text{pred},\ell+1}$ and the fourth token of the answer clause $Z_{\text{ans},\ell}$. This co-occurrence furnishes a strong, consistent signal that enables robust retrieval across longer chains, which will be elaborated in the proof overview in Section 6.2.

5 Learning CoT on Symmetry Group Actions on \mathbb{Z}_n

We now turn to the case where the action group \mathcal{G} is isomorphic to the symmetry group, under Assumption 5.1. In this case, the problem is NC^1 -complete for $n_y \geq 5$.

Assumption 5.1 (Symmetry group actions). Let $\mathcal{Y} = \{0, 1, \dots, n_y - 1\}$. We set $\mathcal{G} = \mathbf{Sym}(\mathcal{Y})$, the symmetry group of all permutations of \mathcal{Y} , so that $|\mathcal{G}| = n_y!$. We let \mathcal{G} act on \mathcal{Y} in the natural way and write $g \cdot y$ (or $g(y)$) for the image of $y \in \mathcal{Y}$ under $g \in \mathcal{G}$. For example, when $\mathcal{Y} = \{0, 1, 2\}$ and g swaps 0 and 1, then $g(0) = 1$, $g(1) = 0$, and $g(2) = 2$. We assume $n_y = \Theta(\frac{\log \log d}{\log \log \log d})$, and hence $|\mathcal{G}| = n_y! = \text{polylog} d$.

Challenges of learning CoT for the symmetry task. When the model tries to retrieve the group element $g_{\ell+1}$ from the correct context clause $Z_{\text{pred},\ell}$, there can be other context clauses whose group elements also send y_ℓ to $y_{\ell+1}$; we call these *distractor* clauses. In the symmetry case on \mathcal{Y} , each pair (i, j) admits $(n_y - 1)!$ elements mapping i to j , so the fraction of distractors is substantial. By contrast, in the simply transitive setting each pair has a unique element, so distractors are unlikely and can be ignored. Attending to distractors still produces the correct next answer, so training may converge with **insufficient attention concentration**. This weaker concentration makes the attention layer less robust to dilution in longer contexts. Hence, for this harder setting, directly proving $d^{\Omega(1)}$ length CoT generalization from constant-length training is difficult.

Self-improvement for reasoning length extension. Recent empirical studies [SCRA+23, GPS+23, LCS+25] show that *length generalization* can be bootstrapped via model *self-improvement*: models training on their own output can bootstrap their capability to solve longer problems. In particular, the work [LCS+25] motivates a **recursive self-training** scheme for the symmetry task. To perform recursive self-training, we adopt the greedy language model as data annotator: the greedy language model \hat{p}_F induced by the network F is defined by

$$\hat{p}_F(Z_{\text{ans},L'+1} | Z^{L,L'}) = \begin{cases} 1, & \text{if } Z_{\text{ans},L'+1} = \arg\max_Z p_F(Z | Z^{L,L'}), \\ 0, & \text{otherwise.} \end{cases} \quad (9)$$

Now we can define the self-annotated LEGO data distribution:

Definition 5.1 (Bootstrapped LEGO distribution). We define $\mathcal{D}_F^{L,L'}$ as the LEGO distribution in Assumption 3.2 except that the answers $Z_{\text{ans},\ell}$, $1 \leq \ell \leq L'$ is given recursively by sampling the prediction $Z_{\text{ans},\ell} \sim \hat{p}_F(\cdot | Z^{L,\ell-1})$, $1 \leq \ell \leq L'$ from the greedy language model \hat{p}_F .

Definition 5.2 (Self-training loss). Given a (fixed) model \tilde{F} and length L , The self-training next-clause-

Algorithm 2: Recursive self-training for symmetry actions

Input: Model $F^{(0)}$ with parameters $(\mathbf{W}^{(0)}, \mathbf{Q}^{(0)})$; Learning rate η ; Error degree $E_1 > 0$ (constant) ;
 $\tau_1, \tau_2 = \tilde{O}(\frac{\text{poly}d}{\eta})$; Total Stage K .

Stage 1.1: Train FFN for one-step reasoning (\mathcal{T}^1) ;

```
for  $t = 1$  to  $\tau_1$  do                                     // Update the FFN parameter  $\mathbf{W}$ 
     $\mathbf{W}^{(t)} \leftarrow \mathbf{W}^{(t-1)} - \eta \nabla_{\mathbf{W}} \text{Loss}^1(F^{(t-1)})$  ;
     $\mathbf{Q}^{(t)} \equiv \mathbf{Q}^{(t-1)}$ ;
```

Stage 1.2: Train attention for length extension (\mathcal{T}^2) ;

```
for  $t = \tau_1 + 1$  to  $\tau_1 + \tau_2$  do                         // Update the attention parameter  $\mathbf{Q}$ 
     $\mathbf{Q}^{(t)} \leftarrow \mathbf{Q}^{(t-1)} - \eta \nabla_{\mathbf{Q}} \text{Loss}_5^{2,2}(F^{(t-1)})$  ;
     $\mathbf{W}^{(t)} \equiv \mathbf{W}^{(t-1)}$ ;
```

$T_1 \leftarrow t$;

Till Stage K : Recursive self-train for length extension ;

```
for  $k = 2$  to  $K$  do                                         // Stage  $k$  to solve  $\mathcal{T}^{2^k}$ 
     $L \leftarrow 2^k, \tilde{F}^{(k)} \leftarrow F^{(T_{k-1})}$ ;
    while  $\text{Loss}_{\tilde{F}^{(k)},5}^{L,2}(F^{(t-1)}) > \frac{1}{d^{E_1}}$  do      // Update the attention parameter  $\mathbf{Q}$ 
         $t \leftarrow t + 1$ ;
         $\mathbf{Q}^{(t)} \leftarrow \mathbf{Q}^{(t-1)} - \eta \nabla_{\mathbf{Q}} \text{Loss}_{\tilde{F}^{(k)},5}^{L,2}(F^{(t-1)})$ ;
         $\mathbf{W}^{(t)} \equiv \mathbf{W}^{(t-1)}$ ;
```

$T_k \leftarrow t$;

Output: Models $\{F^{(T_k)}\}_{k=1}^K$.

prediction loss is defined by replacing $\mathcal{D}^{L,L'}$ with $\mathcal{D}_F^{L,L'}$ (Definition 5.1) in (6):

$$\text{Loss}_{\tilde{F}}^{L,L'}(F) \triangleq \mathbb{E}_{Z^{L,L'} \sim \mathcal{D}_{\tilde{F}}^{L,L'}} \left[-\log p_F(Z_{\text{ans},L',i} \mid Z^{L,L'-1}) \right], \quad (10a)$$

$$\text{Loss}_{\tilde{F},i}^{L,L'} = \mathbb{E}_{Z^{L,L'} \sim \mathcal{D}_{\tilde{F}}^{L,L'}} [-\log p_{F_i}(Z_{\text{ans},L',i} \mid Z^{L,L'-1})] \quad \text{for } i \in [5]. \quad (10b)$$

We now present our main results, establishing that a recursive self-training scheme can provably bootstrap the reasoning length for the symmetry LEGO task.

Theorem 5.1. *Assume the distribution \mathcal{D}^L induced from $\text{LEGO}(\mathcal{X}, \mathcal{G}, \mathcal{Y})$ satisfies Assumption 3.1, 3.2 and 5.1, and assume the transformer network satisfies Assumption 3.3, 3.4, 3.5, and $n_y < m \ll \log^2 d$. Then for any $1 \leq k < \log_2 |\mathcal{X}|$, the transformer $F^{(T_k)}$ trained via Algorithm 2 up to length $L_k = 2^k$ and $T_k = O(\frac{\text{poly}(d)}{\eta})$ satisfies:*

1. **Constant-factor length generalization:** $F^{(T_k)}$ is able to solve $\mathcal{T}^{L_{k+1}}$ with $L_{k+1} = 2^{k+1}$

$$\text{Acc}_{L_{k+1}}(F^{(T_k)}) = 1 - \frac{1}{\text{poly}(d)}. \quad (11)$$

2. **Attention concentration:** given $Z^{L_k, \ell}$ with $\ell \in \{0, \dots, L_k - 1\}$, we have

$$\text{Attn}_{\text{ans}, \ell \rightarrow \text{pred}, \ell+1}^{(T_k)} + \text{Attn}_{\text{ans}, \ell \rightarrow \text{ans}, \ell}^{(T_k)} \geq 1 - \tilde{c}, \quad (12)$$

where \tilde{c} is some sufficiently small constant (smaller than 0.01).

At convergence for the current task \mathcal{T}^{L_k} (at time T_k , the loss has fallen below $1/d^{E_1}$ in Algorithm 2), (12)

confirms that attention concentration is still insufficient. Nevertheless, Theorem 5.1 shows that while this level of concentration cannot withstand the dilution from much longer contexts, it is sufficient for doubling the length. Consequently, a model trained progressively on \mathcal{T}^L for $L = 1, 2, \dots, 2^k$ generalizes to the more challenging task of length 2^{k+1} , yielding the following corollary.

Corollary 5.1 (Self-improvement for $|\mathcal{X}|$ -length reasoning). *Under the same assumptions as Theorem 5.1, letting $K = \Theta(\log d)$, for any length $L \leq |\mathcal{X}|$, the model $F^{(T_K)}$ trained via Algorithm 2 achieves*

$$\text{Acc}_L\left(F^{(T_K)}\right) \geq 1 - \frac{1}{\text{poly}(d)}.$$

Significance of the result. Note that $\{x_\ell\}_{\ell=0}^L$ is sampled from \mathcal{X} *without replacement* (Assumption 3.2), the longest feasible chain scales with the variable size: $L+1 \leq |\mathcal{X}| = \Theta(d)$. Thus our guarantee attains the best possible length in this setting. Corollary 5.1 also demonstrates that the transformer can be trained to solve a task beyond TC^0 with linear-step CoT, matching the expressivity result of [LLZM24].⁴ While prior empirical work reports self-improvement in practice, theoretical guarantees, especially for transformers and length generalization, have been scarce [HBF⁺24, SLZT25, SIZE⁺24]. Theorem 5.1 provides, to our knowledge, the first rigorous evidence that transformers can *bootstrap* their reasoning via self-training without additional supervision.

6 Proof Overview

In this section, we outline the proof ideas for the main theorem. Our training schemes in Algorithms 1 and 2 alternate between two phases. First, we train the FFN parameters \mathbf{W} to solve the one-step task \mathcal{T}^1 . Then, holding \mathbf{W} fixed, we train the attention parameters \mathbf{Q} to solve \mathcal{T}^2 and, for symmetry group actions, recursively \mathcal{T}^{2^k} . This mirrors the division of labor in our setting (Figure 3): the FFN learns the *local update rule*, while the attention layer learns to *route and compose* these updates by locating relevant context over long sequences.

Guided by this picture, our proof overview proceeds in two parts: (1) learning the one-step mechanism for LEGO (\mathcal{T}^1): in-context variable retrieval (Section 6.1.1) and group operations (Section 6.1.2). (2) learning the attention layer: direct short-to-long generalization on \mathcal{T}^2 under simply transitive actions (Section 6.2.1), and recursive length generalization via self-training on \mathcal{T}^{2^k} under symmetry group actions (Section 6.2.2).

Notations Let us first define a few notations to facilitate the presentation of the proofs. For each $i \in [5]$, $j \in [d]$, $r \in [m]$, define

$$\Lambda_{i,j,r}(\mathbf{Z}^{L,\ell-1}) \triangleq \sum_{\mathbf{k} \in \mathcal{T}^{L,\ell-1}} \text{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}} \cdot \langle \mathbf{W}_{i,j,r}, \mathbf{Z}_{\mathbf{k}} \rangle + b_{i,j,r}. \quad (13)$$

The quantity $\Lambda_{i,j,r}$ is the FFN pre-activation, i.e., the input to **sReLU**, for token position i , vocabulary index j , and hidden unit r . According to (3), given $\mathbf{Z}^{L,\ell-1}$, the model’s output at token position i and vocabulary index j is

$$\left[F_i(\mathbf{Z}^{L,\ell-1}) \right]_j = \sum_{r \in [m]} \text{sReLU}(\Lambda_{i,j,r}(\mathbf{Z}^{L,\ell-1})). \quad (14)$$

We denote $\mathbf{W}_i \triangleq \{\mathbf{W}_{i,j,r}\}_{j \in [d], r \in [m]}$ as the FFN parameters associated with token position i . Each $\mathbf{W}_{i,j,r} \in \mathbb{R}^{5d}$ is written as a vertical concatenation $\mathbf{W}_{i,j,r} = [\mathbf{W}_{i,j,r,1}; \dots; \mathbf{W}_{i,j,r,5}]$, where $\mathbf{W}_{i,j,r,i'} \in \mathbb{R}^d$ for $i' \in [5]$; these five blocks align with the five token-type inputs.

⁴ There are a few caveats. For example, we do not analyze an embedding dimension logarithmic in the problem length. We believe our techniques can be extended to cover this setting.

6.1 Learning One-Step Reasoning

For solving task \mathcal{T}^1 , in stage 1 (stage 1.1 for symmetry group task), we train \mathbf{W} via the full loss Loss^1 , which contains the prediction loss across five tokens in the answer clause and \mathbf{W}_i is responsible for predicting the i -th token in the answer clause. Thus, there are three different types of prediction tasks here: the $\langle \text{blank} \rangle$ tokens (tokens 1, 2, 3), the correct variable x_1 (token 4) and the action update $y_1 = g_1(y_0)$ (token 5). Notice that the $\langle \text{blank} \rangle$ tokens are deterministic, the learning task is straightforward. We therefore focus on the learning dynamics for the 4th and 5th tokens, which involve in-context retrieval (\mathbf{W}_4) and one-step group action (\mathbf{W}_5). Across these stage, the attention layer is fixed and keeps as uniform attention due to the zero initialization of the attention matrix, i.e., the input for the FFN layer is $\frac{1}{2}\mathbf{Z}_{\text{pred},1} + \frac{1}{2}\mathbf{Z}_{\text{ans},0}$.

6.1.1 Learning In-Context Retrieval of Variables

Given the input $\frac{1}{2}\mathbf{Z}_{\text{pred},1} + \frac{1}{2}\mathbf{Z}_{\text{ans},0}$ for the FFN layer, the goal is to predict the fourth token in the answer clause $\mathbf{Z}_{\text{ans},1}$, which is the variable x_1 . Intuitively, the network should retrieve the occurrence of x_1 in the first predicate token $\mathbf{Z}_{\text{pred},1}$ and copy it to the target position. Specifically, the FFN pre-activation for predicting the fourth token to be j is

$$\Lambda_{4,j,r}(\mathbf{Z}^{1,0}) = \frac{1}{2} \langle \mathbf{W}_{4,j,r}, \mathbf{Z}_{\text{pred},1} \rangle + \frac{1}{2} \langle \mathbf{W}_{4,j,r}, \mathbf{Z}_{\text{ans},0} \rangle + b_{4,j,r}.$$

Using the 5-vector decomposition of $\mathbf{W}_{4,j,r}$, and the fact that the embedding of $\langle \text{blank} \rangle$ tokens are zero vectors (recall $e_{\tau(\cdot)}$ denote the token embedding vectors), we further obtain:

$$\begin{aligned} \Lambda_{4,j,r}(\mathbf{Z}^{1,0}) = & \frac{1}{2} \left(\langle \mathbf{W}_{4,j,r,1}, e_{\tau(x_1)} \rangle + \langle \mathbf{W}_{4,j,r,2}, e_{\tau(g_1)} \rangle + \langle \mathbf{W}_{4,j,r,3}, e_{\tau(x_0)} \rangle \right) \\ & + \frac{1}{2} \left(\langle \mathbf{W}_{4,j,r,4}, e_{\tau(x_0)} \rangle + \langle \mathbf{W}_{4,j,r,5}, e_{\tau(y_0)} \rangle \right) + b_{4,j,r}. \end{aligned} \quad (15)$$

Therefore, the main idea of our analysis is to track the training dynamics of $\langle \mathbf{W}_{4,j,r,p}, e_{s'} \rangle$ for $j, s' \in [d]$, $p \in [5]$, and $r \in [m]$. Letting $s \in \tau(\mathcal{X})$ be an embedding index for a variable, our analysis shows the diagonal correlations $\langle \mathbf{W}_{4,s,r,1}, e_s \rangle$ receive strictly larger updates than all other $\langle \mathbf{W}_{4,j,r,p}, e_{s'} \rangle$. This occurs because x_1 co-occurs simultaneously as the first input token at $\mathbf{Z}_{\text{pred},1}$ and as the supervised target at $\mathbf{Z}_{\text{ans},1}$, which amplifies the gradient on $\langle \mathbf{W}_{4,s,r,1}, e_s \rangle$ when $s = \tau(x_1)$. In contrast, non-target coordinates (off-diagonals, wrong variables, value tokens, and group-action tokens) incur negligible gradients and remain $o(1)$ throughout training. Hence the correct variable's diagonal signal becomes order-wise larger, and the *active diagonal mass* $\sum_r \langle \mathbf{W}_{4,s,r,1}, e_s \rangle$ dominantly grows until the end of training.

Given the dominance above, the learned weights align so that $\mathbf{W}_{4,s,r,1}$ points toward e_s for $s \in \tau(\mathcal{X})$. Substituting this alignment into (15) shows that, when $j = \tau(x_1)$, the red term $\langle \mathbf{W}_{4,\tau(x_1),r,1}, e_{\tau(x_1)} \rangle$ contributes dominantly to $\Lambda_{4,\tau(x_1),r}(\mathbf{Z}^{1,0})$, thereby realizing the intended in-context retrieval: the model copies x_1 from the first position of $\mathbf{Z}_{\text{pred},1}$ to the fourth position of $\mathbf{Z}_{\text{ans},1}$.

6.1.2 Learning the Group Actions

For task \mathcal{T}^1 , the FFN input $\frac{1}{2}\mathbf{Z}_{\text{pred},1} + \frac{1}{2}\mathbf{Z}_{\text{ans},0}$ already contains the current value y_0 and the action g_1 , with no distracting information. Accordingly, when predicting the next value y_1 , the role of \mathbf{W}_5 is to correctly apply the action g_1 to the current value y_0 . In this section, we sketch how the model learns to implement simply transitive group actions, and we briefly discuss the symmetry case.

We first introduce notation to explain what the model should learn.

Definition 6.1 (Combinations, simply transitive actions). Assuming the group \mathcal{G} follows Assumption 4.1, for each class index $j \in \tau(\mathcal{Y})$, define the *combinations*

$$\Phi := \bigcup_{j \in \tau(\mathcal{Y})} \Phi_j^*, \quad \text{where } \Phi_j^* := \{(g', y') \in \mathcal{G} \times \mathcal{Y} : \tau(g'(y')) = j\}.$$

$|\Phi| = n_y^2$ for the simply transitive case. We call $\phi = (g, y) \in \Phi_j^*$ a *combination* for predicting $j = \tau(g(y))$. Hence, the goal of the model is to correctly identify all of these $\phi = (g, y) \in \Phi$.

Analogously to (15), the pre-activation at the fifth token for class j and neuron r is

$$\Lambda_{5,j,r}(\mathbf{Z}^{1,0}) = \frac{1}{2} \langle \mathbf{W}_{5,j,r,2}, e_{\tau(g_1)} \rangle + \frac{1}{2} \langle \mathbf{W}_{5,j,r,5}, e_{\tau(y_0)} \rangle + \text{other terms},$$

where slot “2” reads the action token g_1 and slot “5” reads the current value token y_0 . We then define the following feature-magnitude notation:

Definition 6.2 (*V-Notations*). Given $\phi = (g, y) \in \Phi$, for the r -th neuron in the j -th class with $j \in \tau(\mathcal{Y})$ and $r \in [m]$, i.e., $\mathbf{W}_{5,j,r} \in \mathbb{R}^{5d}$, we define

$$V_{j,r}(g) := \langle \mathbf{W}_{5,j,r,2}, e_g \rangle, \quad V_{j,r}(y) := \langle \mathbf{W}_{5,j,r,5}, e_y \rangle,$$

and the composite feature magnitude

$$V_{j,r}(\phi) := \frac{1}{2} (V_{j,r}(g) + V_{j,r}(y)).$$

Then $V_{j,r}(\phi)$ is exactly the contribution of the input pair (g, y) to $\Lambda_{5,j,r}(\mathbf{Z}^{1,0})$ for predicting j . Notice that for each class j , there are m associated neurons $\{\mathbf{W}_{5,j,r}\}_{r=1}^m$, and we index neurons by the pair (j, r) to emphasize that neuron r is specific to class j ; an index r alone has no cross-class meaning in this context.

The main proof idea is to track $V_{j,r}(\phi)$ throughout training and show that it amplifies the correct correlations across the combinations Φ while suppressing the incorrect ones. To make this concrete and to clarify the roles of different neurons, we introduce the following neuron–feature index set:

$$\Psi := \{(j, r, \phi) \mid j \in \tau(\mathcal{Y}), r \in [m], \phi \in \Phi\}.$$

Here, (j, r) again refers to neuron r for class j , as in Definition 6.2. With this notation we can write $V_{j,r}(\phi)$ as V_ψ for $\psi = (j, r, \phi) \in \Psi$.

Our proof shows that, given j , for each $\phi \in \Phi_j^*$, there is exactly one neuron r in $\mathbf{W}_{5,j,r}$ that is activated, denoted $r_{g,y}$, to learn ϕ ; that is, $V_{j,r_{g,y}}(\phi)$ will grow to a large value. Therefore, in total n_y^2 distinct neurons will be activated to learn all combinations, i.e., $\{(j, r_{g,y}, (g, y)), \forall j \in \tau(\mathcal{Y}), (g, y) \in \Phi_j^*\}$. The magnitude of remaining V_ψ with non-activated neurons will stay close to the initialization. Our analysis shows that learning follows an *implicit curriculum* induced by the magnitude of features V_ψ at initialization:

$$\psi \prec \psi' \iff V_\psi^{(0)} \geq V_{\psi'}^{(0)}, \quad \forall \psi, \psi' \in \Psi.$$

Items on the left under this ordering are learned first, and those on the right are learned later. We denote by Σ^* the *learning curriculum*: the ordered set of neuron–feature indices $\{(j, r_{g,y}, (g, y))\}$ identified above, equipped with this order.

Then, for each $\psi \in \Sigma^*$, the learning process follows its associated ordering, and when it is the turn of $\psi = (j, r_{g,y}, (g, y))$, the learning process is mainly characterized by the following two phases:

- **Phase I:** Emergence of the feature V_ψ among other features. During this phase, V_ψ grows faster than any $\psi' \neq \psi$ with $\psi \prec \psi'$, while not affecting the already learned predecessors $\psi' \in \Sigma^*$ with $\psi' \prec \psi$, with growth rate

$$V_\psi^{(t+1)} \geq V_\psi^{(t)} + \tilde{\Omega}(\eta) \cdot (b_{j,r} + V_\psi^{(t)} \pm o(\mu))^{q-1}. \quad (16)$$

This form permits the application of the tensor power method (TPM) [AZL20] and explains the ordering induced by the magnitude of V_ψ at initialization, since TPM implies that a slightly larger initial value leads to dramatically faster growth. By the end of Phase I, for any other feature $\psi' \neq \psi$ with $\psi \prec \psi'$, the growth of that feature is capped at $\tilde{O}(\sigma_0)$, its initial magnitude.

- **Phase II:** Growth of V_ψ and cancellation of incorrect features. After Phase I, the target feature V_ψ already has a relatively large magnitude; however, features in the set $\{\psi' = (j, r_{g \cdot y}, \phi') \in \Psi \mid \phi' = (g', y) \text{ or } \phi' = (g, y')\}$, namely, the features in the same neuron $(j, r_{g \cdot y})$ that share exactly one component of $\phi = (g, y)$, may also grow as the shared component increases. Note that j is not the correct label for such ϕ' , and we call them **confounding features**. The key characterization in this phase is that, due to a stationarity property of the gradients, although the dynamics are coupled, the wrong half of the confounding features grows in the negative direction while the correct half continues to grow, and ultimately the confounding feature cancels out.

We show that for each $\psi \in \Sigma^*$ that should be learned, it retains its structure after its own learning process (i.e., at the end of Phase II) and persists through the final convergence while the other features in Σ^* are learned. Specifically, we have the following properties at the end of training:

$$V_{j, r_{g \cdot y}}(g), V_{j, r_{g \cdot y}}(y) \approx B, \quad (17a)$$

$$V_{j, r_{g \cdot y}}(g'), V_{j, r_{g \cdot y}}(y') \approx -B, \quad (17b)$$

$$V_{j, r_{g \cdot y}}(g) + V_{j, r_{g \cdot y}}(y') \leq o(1) \text{ and } V_{j, r_{g \cdot y}}(g') + V_{j, r_{g \cdot y}}(y) \leq o(1) \quad \forall g' \neq g, y' \neq y. \quad (17c)$$

How does this structure perform group actions? For an input pair (g, y) , let the correct answer be $j = \tau(g(y))$. Then the pre-activation for predicting j is around B at neuron $r_{g \cdot y}$, with all other neurons for j near the initial value $\tilde{O}(\sigma_0)$. Thus the model output at j is around B . For all other predictions j' , there exist $\phi_1 = (g', y)$ and $\phi_2 = (g, y')$ such that $j' = \tau(g'(y)) = \tau(g(y'))$. By cancellation of the incorrect features, we have $V_{j', r_{g' \cdot y}}(\phi_1), V_{j', r_{g \cdot y'}}(\phi_2) \leq o(1)$. Since $B = \Theta(\log d)$, this implies that the logit on the correct prediction j is very close to 1.

Symmetry Group Actions. The proof strategy for symmetry group actions mirrors the simply transitive case through the emergence, refinement, and convergence phases. However, symmetry actions create richer interactions because multiple group elements can map the same y to the same j , which requires more nuanced control of the training dynamics and leads to different learned feature structures. To illustrate this pattern, we slightly modify the definition of **combinations** used in the simply transitive case by introducing the notion of the **fiber** of a value.

Definition 6.3 (Combinations, Symmetry Actions). Assuming the group \mathcal{G} follows Assumption 5.1, define $\text{Fiber}_{j,y} := \{g \in \mathcal{G} \mid \tau(g(y)) = j\}$.⁵ This allows us to define the combinations as

$$\Phi = \{\varphi_{j,y} \mid j \in \tau(\mathcal{Y}), y \in \mathcal{Y}\}, \quad \text{where } \varphi_{j,y} = \text{Fiber}_{j,y} \times \{y\}.$$

Moreover, $|\varphi_{j,y}| = (n_y - 1)!$. We continue to call $\phi = (g, y) \in \mathcal{G} \times \mathcal{Y}$ a combination, whereas an element of Φ is now a subset of combinations $\varphi_{j,y}$ (there are n_y^2 such $\varphi_{j,y}$ in total), which includes all pairs (g, y) such that g sends y to j , and which reduces to the single pair (g, y) in the simply transitive case.

Based on these notions, the main difference from the simply transitive case is that previously the basic learning unit is each pair $(g, y) \in \Phi$, whereas now it is the subset $\varphi_{j,y}$ of combinations. All combinations ϕ within $\varphi_{j,y}$ are captured by the same and unique neuron, denoted $r_{j,y}$, in the sense that the feature magnitude $V_{j, r_{j,y}}(\phi)$ is large for any $\phi \in \varphi_{j,y}$. Accordingly, the learning curriculum Σ^* is now based on the initial value of the ensemble of feature magnitudes: $\frac{1}{|\varphi_{j,y}|} \sum_{\phi \in \varphi_{j,y}} V_{j, r_{j,y}}(\phi)$. Finally, at convergence, we have the following imbalance of feature magnitudes:

$$V_{j, r_{j,y}}(g) \approx 2B, \quad V_{j, r_{j,y}}(y) \approx \frac{2B}{n_y} \quad \forall g \in \text{Fiber}_{j,y}; \quad (18a)$$

$$V_{j, r_{j,y}}(g') \approx \frac{2B}{n_y}, \quad V_{j, r_{j,y}}(y') \approx -2B; \quad (18b)$$

⁵ We use the notion of fiber to denote the left cosets $gG_y \subset \mathcal{G}$, $g \in \mathcal{G}$, where $y \in \mathcal{Y}$ and $G_y = \{g \in \mathcal{G} \mid g(y) = y\}$ is the stabilizer. Since the fiber of the orbit map $f_y(g) = g(y)$ is the preimage of f_y , i.e., $f_y^{-1}(y')$ with $y' = \tau^{-1}(j)$, $j \in \tau(\mathcal{Y})$, this is exactly the set $\text{Fiber}_{j,y}$.

$$V_{j,r_{j,y}}(g) + V_{j,r_{j,y}}(y') \leq o(1) \text{ and } V_{j,r_{j,y}}(g') + V_{j,r_{j,y}}(y) \leq o(1) \quad \forall g' \notin \text{Fiber}_{j,y}, y' \neq y. \quad (18c)$$

6.1.3 What Changes for Longer Tasks?

FFN is length invariant. The feed-forward network (with weights \mathbf{W}) only acts on the *attended linear combination* at the current output clause, rather than scanning the sequence to retrieve information. Retrieval from the context is delegated to the attention layer. Hence, as sequence length grows and positions shift, the FFN computation remains the same mapping on its local input, making it length- and position-invariant. On the other hand, uniform attention becomes increasingly diluted as irrelevant clauses accumulate, so selective attention and learning attention layer are required.

Desired attention patterns for predicting each tokens. To predict the fourth token, the model only needs the variable $x_{\ell+1}$ from the predicate clause $\mathbf{Z}_{\text{pred},\ell+1}$. If the attention mass on $\mathbf{Z}_{\text{pred},\ell+1}$ does not vanish with L , the same pattern learned in \mathcal{T}^1 with \mathbf{W}_4 applies directly. Predicting the fifth token is harder because it requires *two* retrievals: (i) the group action from $\mathbf{Z}_{\text{pred},\ell+1}$, and (ii) the current value from $\mathbf{Z}_{\text{ans},\ell}$, followed by applying the update via \mathbf{W}_5 . Thus robustness over length hinges on maintaining both attention links. If the attention layer is robust enough to support the fifth-token prediction, then the fourth-token prediction follows as a special case, since it needs only one of the two attention links to persist. Accordingly, in what follows we focus on optimizing the loss for the fifth token; achieving high accuracy there is effectively equivalent to solving the entire task.

6.2 Learning the Attention Layer

Successful training on \mathcal{T}^1 shows that the model has learned the one-step update $g \circ y$. Turning to task \mathcal{T}^2 , the remaining difficulty is *routing*: directing attention to the appropriate locations. Thus we train the attention matrix \mathbf{Q} to learn the routing pattern and keep \mathbf{W} fixed. In this part, we will show how training induces the **attention concentration** pattern highlighted in Theorems 4.1 and 5.1: given an input $\mathbf{Z}^{L,\ell-1}$, the attention mass concentrates on $\text{Attn}_{\text{ans},\ell-1 \rightarrow \text{pred},\ell}$ and $\text{Attn}_{\text{ans},\ell-1 \rightarrow \text{ans},\ell-1}$.

We quantify routing quality via the *attention concentration degree*

$$\epsilon_{\text{attn}}^{L,\ell}(\mathbf{Z}^{L,\ell-1}) = 1 - \text{Attn}_{\text{ans},\ell-1 \rightarrow \text{pred},\ell}(\mathbf{Z}^{L,\ell-1}) - \text{Attn}_{\text{ans},\ell-1 \rightarrow \text{ans},\ell-1}(\mathbf{Z}^{L,\ell-1}), \quad (19)$$

which measures the fraction of attention mass *not* placed on the two key clauses, and the *attention gap*

$$\Delta^{L,\ell}(\mathbf{Z}^{L,\ell-1}) = \left| \text{Attn}_{\text{ans},\ell-1 \rightarrow \text{pred},\ell}(\mathbf{Z}^{L,\ell-1}) - \text{Attn}_{\text{ans},\ell-1 \rightarrow \text{ans},\ell-1}(\mathbf{Z}^{L,\ell-1}) \right|, \quad (20)$$

which captures how balanced the two target attentions are. Thus, effective routing corresponds to small $\epsilon_{\text{attn}}^{L,\ell}$ (high concentration) and small $\Delta^{L,\ell}$ (good balance).

Under Assumption 3.5, for an input $\mathbf{Z}^{L,\ell-1}$ the (unnormalized) attention score from $\mathbf{Z}_{\text{ans},\ell-1}$ to a clause $\mathbf{Z}_{\mathbf{k}}$ decomposes as

$$\mathbf{Z}_{\text{ans},\ell-1}^\top \mathbf{Q} \mathbf{Z}_{\mathbf{k}} = \mathbf{Z}_{\text{ans},\ell-1,4}^\top \mathbf{Q}_{4,3} \mathbf{Z}_{\mathbf{k},3} + \mathbf{Z}_{\text{ans},\ell-1,4}^\top \mathbf{Q}_{4,4} \mathbf{Z}_{\mathbf{k},4},$$

where $\mathbf{Z}_{\mathbf{k}} = [\mathbf{Z}_{\mathbf{k},1}, \dots, \mathbf{Z}_{\mathbf{k},5}]$ with $\mathbf{Z}_{\mathbf{k},i} \in \mathbb{R}^d$. By design, in the clause embeddings the *fourth* token of a *predicate* clause and the *third* token of an *answer* clause are $\langle \text{blank} \rangle$. Consequently, we further have

$$\mathbf{Z}_{\text{ans},\ell-1}^\top \mathbf{Q} \mathbf{Z}_{\text{pred},\ell'} = \mathbf{Z}_{\text{ans},\ell-1,4}^\top \mathbf{Q}_{4,3} \mathbf{Z}_{\text{pred},\ell',3}, \quad \mathbf{Z}_{\text{ans},\ell-1}^\top \mathbf{Q} \mathbf{Z}_{\text{ans},\ell'} = \mathbf{Z}_{\text{ans},\ell-1,4}^\top \mathbf{Q}_{4,4} \mathbf{Z}_{\text{ans},\ell',4}, \quad (21)$$

which means $\mathbf{Q}_{4,3}$ governs attention to predicate clauses, while $\mathbf{Q}_{4,4}$ governs attention to answer clauses. An immediate observation from (21) is that the desired allocation can be realized by growing the diagonal entries $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$ ⁶ for $s \in \tau(\mathcal{X})$ since $x_{\ell-1}$ appears as the third token in $\mathbf{Z}_{\text{pred},\ell}$ and as the fourth token in $\mathbf{Z}_{\text{ans},\ell-1}$ simultaneously. Our analysis of tracking the attention dynamics will show that these diagonal coordinates indeed receive asymptotically larger gradient magnitudes than all other entries due to

⁶ $[\mathbf{A}]_{s,s'}$ denotes the entry of the matrix \mathbf{A} at the s -th row and s' -th column

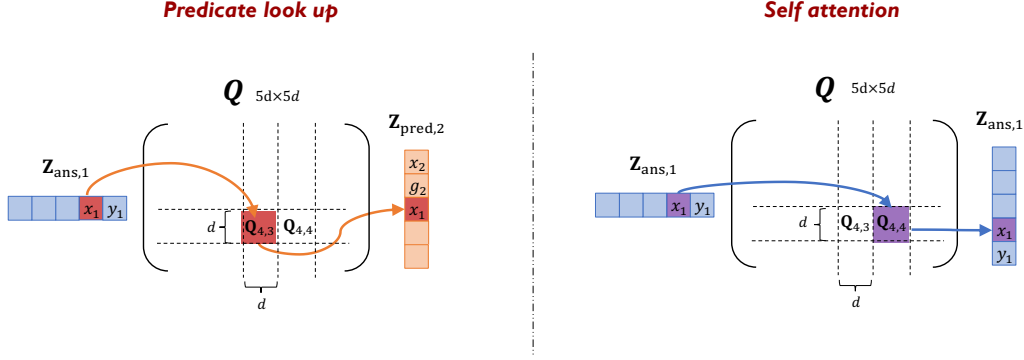


Figure 4: The illustration of how different components of the attention matrix \mathbf{Q} are used to route the attention to the appropriate locations. The query clause is $\mathbf{Z}_{\text{ans},1}$ and the goal is to retrieve the correct action g_2 from $\mathbf{Z}_{\text{pred},2}$ and value y_1 from the current answer clause $\mathbf{Z}_{\text{ans},1}$. $[\mathbf{Q}_{4,p}]_{s,s}$ will grow and dominate the learning dynamics for $p \in \{3, 4\}$ and $s \in \tau(\mathcal{X})$. Thus, in this example, large $[\mathbf{Q}_{4,3}]_{\tau(x_1), \tau(x_1)}$ indicates the large attention to the predicate clause $\mathbf{Z}_{\text{pred},2}$ and large $[\mathbf{Q}_{4,4}]_{\tau(x_1), \tau(x_1)}$ indicates the large self-attention to the answer clause $\mathbf{Z}_{\text{ans},1}$.

strong co-occurrence signal, and the training dynamics are dominated by their growth. See Figure 4 for an illustration.

For notational simplicity, we will refer to the relevant diagonal entries $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$ simply as $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ below. The remainder of the proof quantifies how the growth of $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ simultaneously drives the concentration degree $\epsilon_{\text{attn}}^{L,\ell}$ toward zero and controls the gap $\Delta^{L,\ell}$, ensuring that the FFN consistently receives $(g_\ell, y_{\ell-1})$ and thus outputs the correct y_ℓ .

6.2.1 Simply Transitive Actions

For the simply transitive case, we analyze the gradient contribution at position $i = 5$ on task \mathcal{T}^2 , i.e., the loss $\sum_{\ell=1}^2 \text{Loss}_5^{2,\ell}$. We show that for \mathcal{T}^2 the *attention concentration degree* $\epsilon_{\text{attn}}^{2,\ell}$ (for $\ell \in [2]$) can be reduced below $O(1/d^{c^*})$ for some constant $0 < c^* < 1$, indicating highly focused mass on the relevant clauses. When irrelevant entries of \mathbf{Q} are small, we also have $\epsilon_{\text{attn}}^{2,1} \leq \epsilon_{\text{attn}}^{2,2}$, since the number of irrelevant clauses doubles from $\ell = 1$ to $\ell = 2$; hence we focus on controlling $\epsilon_{\text{attn}}^{2,2}$.

- **Stage 2.1: Growth of an initial gap.** Early in training, attention is close to uniform, so given $\mathbf{Z}^{2,\ell-1}$ we have the approximations

$$\Lambda_{5,j,r}(\mathbf{Z}^{2,0}) \approx \frac{1}{3} V_{j,r}(g_1) + \frac{1}{3} V_{j,r}(g_2) + \frac{1}{3} V_{j,r}(y_0),$$

$$\Lambda_{5,j,r}(\mathbf{Z}^{2,1}) \approx \frac{1}{4} V_{j,r}(g_1) + \frac{1}{4} V_{j,r}(g_2) + \frac{1}{4} V_{j,r}(y_0) + \frac{1}{4} V_{j,r}(y_1).$$

By the cancellation at the convergence in (17), for $\ell = 2$ all Λ lie in the small smoothed regime, whereas for $\ell = 1$ we obtain a correct logit for $y_1 = g_1(y_0)$ and a spurious logit for $g_2(y_0)$ of magnitude about $B/3$. Consequently, $-\nabla_{\mathbf{Q}} \text{Loss}_5^{2,2}$ is negligible, while $-\nabla_{\mathbf{Q}} \text{Loss}_5^{2,1}$ is comparatively large and drives $\mathbf{Q}_{4,3}$ to grow faster than $\mathbf{Q}_{4,4}$ (increasing $\mathbf{Q}_{4,4}$ would also amplify the wrong prediction $\tau(g_2(y_0))$ and thus not reduce $\text{Loss}_5^{2,1}$). An $\Omega(1/\log d)$ gap emerges between the diagonals $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$, yielding an early routing advantage toward right predicate clause.

- **Stage 2.2: Joint growth with a controlled gap.** As $\mathbf{Q}_{4,3}$ increases, the weight $\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}$ becomes large, moving $\Lambda_{5,\tau(g_2(y_1)),r_{g_2 \cdot y_1}}$ and $\Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}$ for $\ell = 2$ into the linear regime. Gradients from $\ell = 2$ then dominate, and $\mathbf{Q}_{4,4}$ starts to grow to separate the correct $y_2 = g_2(y_1)$ from the incorrect $\tau(g_2(y_0))$. Throughout, the gap between $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ stays within $[\Omega(1/\log d), O(1)]$, so the attention gap satisfies $\Delta^{2,2} = \Omega(1/\log d)$.

- **Stage 2.3: Convergence and gap reduction.** Continued joint growth of $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ concentrates attention near its ideal limit, making $\epsilon_{\text{attn}}^{2,2}$ small. We show $\Delta^{2,2}$ cannot remain above $o(1)$ for long; otherwise an incorrect logit $\text{logit}_{5,\tau(g_2(y_0))}$ would acquire a stronger gradient and force $\mathbf{Q}_{4,4}$ to outpace $\mathbf{Q}_{4,3}$, which contradicts stability. Therefore, at convergence we have (i) $\epsilon_{\text{attn}}^{2,2} = O(d^{-c^*})$ for some $c^* \in (0, 1)$ and (ii) $\Delta^{2,2} = o(1)$; hence both $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ equal $C \log d \pm o(1)$, while all other entries of the attention matrix remain close to their initial values, where $C > 0$ depends on c^* .

Strong length generalization. The key is that attention concentrates cleanly while remaining stably balanced. For \mathcal{T}^L , we obtain

$$\epsilon^{L,\ell} \leq \frac{O(1) \cdot L}{O(1) \cdot L + 1/\epsilon_{\text{attn}}^{2,2}} \leq \frac{1}{1 + \Omega(d^{c^*}/L)}.$$

Hence the model tolerates $O(d^{c^*})$ irrelevant clauses: in particular, if $L = o(d^{c^*})$ then $\epsilon^{L,\ell} = o(1)$; and if $L = \Theta(d^{c^*})$ then $\epsilon^{L,\ell} \leq \frac{1}{1+\Omega(1)}$, which can be made small by choosing the proportionality constant in $L = \Theta(d^{c^*})$ appropriately. Moreover, $\Delta^{L,\ell} \leq \Delta^{2,2} \leq o(1)$, so the two target attention masses remain balanced as L grows, preventing errors from large imbalance (e.g., predicting $\tau(g_2(y_0))$ in Stage 2.3). Consequently, the correct logit satisfies $1 - \text{logit}_{5,\tau(g_{\ell+1}(y_\ell))} \leq \frac{1}{\text{poly}(d)}$, so \mathcal{T}^L is solved with accuracy $1 - 1/\text{poly}(d)$.

6.2.2 Symmetry Group Actions

We now turn to symmetry group tasks \mathcal{T}^L and analyze GD updates with respect to the per-token loss $\text{Loss}_5^{L,2}$ (i.e., predicting the value token in $\mathbf{Z}_{\text{ans},2}$ from $\mathbf{Z}^{L,1}$).

The case $L = 2$. The high-level picture mirrors the simply transitive case, but because multiple group elements can map a given y to the same j , the learned $V_{j,r}(g)$ and $V_{j,r}(y)$ structures are unbalanced. Moreover, as we discussed in the hardness part for the symmetry group in Section 5, there will be a non-negligible proportions of distractor clauses. These make it harder to keep a tight balance between $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ as in the simply transitive case. Nevertheless, we prove that both $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ grow, and the attention gap $\Delta^{2,2}$ is controlled by a feedback mechanism: if $\Delta^{2,2}$ exceeds a small fixed threshold (in either direction), some incorrect logit receives a stronger gradient, which pushes the system back toward balance. Consequently, after sufficient training: (i) $\epsilon_{\text{attn}}^{2,2} \leq C_1$; (ii) $\Delta^{2,2} \leq C_2$, for sufficiently small constants C_1, C_2 ; and since $B = \Theta(\log d)$, we still obtain $\text{Loss}_5^{2,2} \leq 1/\text{poly}(d)$.

Recursive learning for \mathcal{T}^{2^k} , $k \geq 2$. Because $\epsilon_{\text{attn}}^{2^{k-1},2}$ is already a small constant, initialization for \mathcal{T}^{2^k} satisfies $\epsilon_{\text{attn}}^{2^k,2} \leq 2\epsilon_{\text{attn}}^{2^{k-1},2}$ (still small), and $\Delta^{2^k,2} \leq \Delta^{2^{k-1},2}$. Thus the attention pattern remains close to that in $\mathcal{T}^{2^{k-1}}$, and $\mathbf{Z}^{2^k,1}$ follows a bootstrapped LEGO distribution generated by the greedy model $\hat{p}_{F^{(\tau_{k-1})}}$, which coincides with the original LEGO source. In particular, $y_1 = g_1(y_0)$ and $y_2 = g_2(y_1)$ are correct. We can therefore reuse the convergence analysis from $\mathcal{T}^{2^{k-1}}$ to show that both $\epsilon_{\text{attn}}^{2^k,2}$ and $\Delta^{2^k,2}$ decrease to a small constant, yielding stable, inductive concentration across recursive reasoning depths.

6.3 Significance of the Proof

Our proof techniques are inspired by recent advances in understanding the dynamics of feature learning in neural networks [AZL21, AZL22, WL21, WL22, HLZ⁺22, JSL22, HCL24, HWCL25], which show how gradient-based training induces useful internal patterns and representations. Building on these ideas, we analyze the dynamics of the FFN and attention layers to capture how transformers gradually acquire length-generalizable reasoning through CoT training. We conclude this section by summarizing the technical significance of our analysis for learning in the FFN and attention layers.

Learning the group actions. Most existing optimization-based analyses of CoT [HWL25, WZLZ24, KS24, HLY25] hard-code the group action, e.g., parity, as a fixed FFN or an almost-linear map. In contrast, we *train* a nonlinear FFN end-to-end to perform the group action at each step, with parity as a special $n = 2$

instance of our simply transitive framework. This formulation is strictly more general and substantially more challenging than prior setups. Our proof shows that the model learns not only the basic action for simply transitive groups but also more complex actions for richer actions that is transitive but not free. Especially, we precisely characterizing how task-relevant features emerge and spurious features are suppressed during the process. The technique we used in the proof for learning the FFN on discrete combinations of data is of independent interest beyond analyzing CoT.

Learning the attention patterns. Prior work on transformer training dynamics has established the *attention concentration* principle as a key mechanism for solving various tasks, e.g., in-context learning [HCL24], self-supervised learning [HWCL25], and graph learning [NDL24]. In those settings, only a single token needs to be retrieved, so the attention matrix is learned to be diagonal (after a suitable change of basis), and a single pattern suffices. In contrast, our task requires retrieving two different types of clauses, so the model must learn two distinct components (blocks) in the attention matrix, $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$. As we show in Section 6.2, beyond growing these blocks, we must maintain a delicate balance between them. This balance is crucial for length generalization across different group actions, introduces substantial technical challenges, and our proof provides a fine-grained control of it.

7 Experiments

In this section, we conduct synthetic experiments to verify our theoretical claim.

General Setup. We adopt an experimental setting closely aligned with our theoretical setup. Data strictly follow the LEGO distribution in the five-token-per-clause format defined in Definition 3.1 and Assumption 3.2. The simply transitive task uses the cyclic group of order 6 (denoted C_6), and the symmetry task uses the symmetry group on five elements (denoted S_5). The network is a one-layer, decoder-only transformer with two attention heads and a FFN block. Training optimizes the next-clause loss in (6a) using Adam [KB14] with a learning rate of 1e-4. We train for 300 epochs to ensure the training loss approaches zero and the model converges. For evaluation, to measure the transformer’s ability to solve CoT reasoning tasks, we eschew the teacher-forced accuracy in (5), which conditions on the ground-truth answer prefix, and instead evaluate final-answer accuracy after autoregressively generating all intermediate answer steps without teacher forcing. This better mimics CoT reasoning and is more challenging because errors can accumulate during self-rollout. For computational efficiency, we report the probability that the model predicts the value of all answer steps correctly, following [LLZM24].

Length generalization for different group actions. Figure 1a shows that when trained at the short length $L = 5$, the model exhibits strong length generalization, achieving near-perfect accuracy at much longer lengths, for example, up to $L = 160$ on the simply transitive task, which corroborates Theorem 4.1. By contrast, on the harder symmetry task the generalization is weaker, yielding only constant-factor extensions as expected.

Self-training improves reasoning length. To extend the solvable length on the harder symmetry task, we adopt a recursive self-training curriculum inspired by [LCS⁺25]. We first train at $L = 5$ with ground-truth answer supervision until convergence. At the next stage, we double the length to $L = 10$, use the $L=5$ model to greedily generate answer traces (self-labels) for the $L=10$ data, and retrain on these pseudo-labels. We repeat this doubling process for three stages ($L = 5 \rightarrow 10 \rightarrow 20 \rightarrow 40$), so that the final model is trained on self-labeled data at $L = 40$. Figure 1b reports the length-generalization curve for each stage, with the dashed line (matching the curve’s color) marking that stage’s training length. Across stages, we observe at least constant-factor generalization beyond the training length; after multiple rounds of self-training, the model achieves nearly perfect accuracy at lengths far exceeding the initial $L = 5$, ultimately matching the simply transitive task’s performance up to $L = 160$. These results validate Theorem 5.1 and demonstrate the effectiveness of recursive self-training for extending reasoning length.

Attention concentration. We visualize attention heatmaps at convergence for models trained at length $L = 5$ on both simply transitive and symmetry group tasks. Each attention matrix is averaged across the two

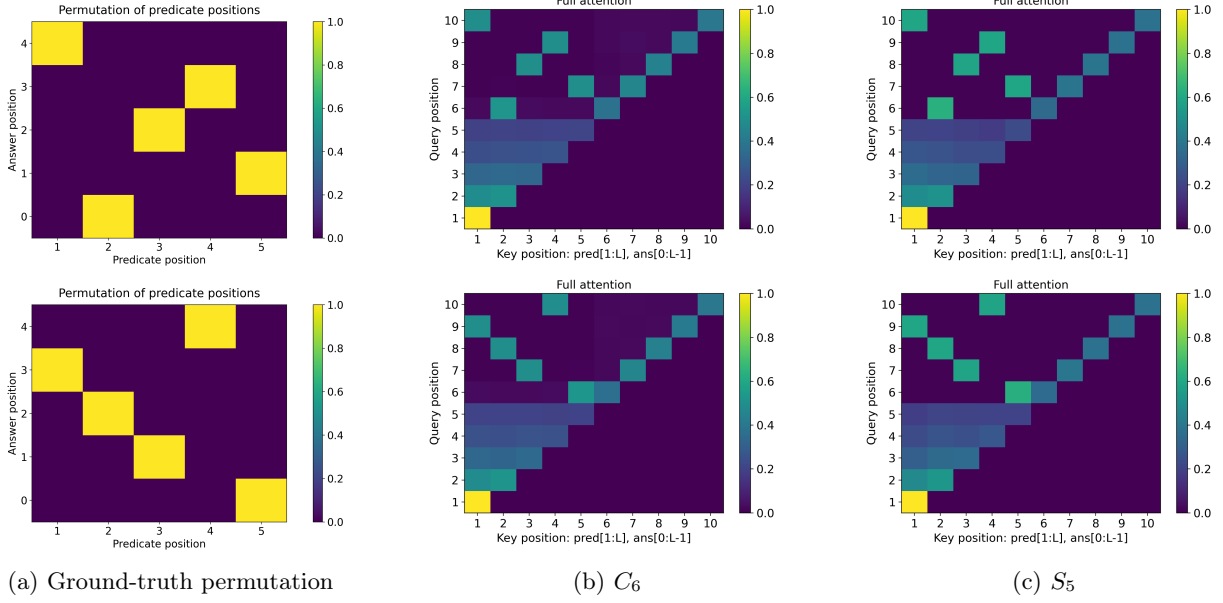


Figure 5: Attention patterns of the same trained model as Figure 2, evaluated with randomly permuted predicate-clause positions. Column (a) gives the ground-truth permutation. Column (b) and Column (c) show the attention heatmaps for the simply transitive and symmetry tasks, respectively.

heads and over 100 independently sampled LEGO sequences. Note that for a task of length L , the LEGO sequence prior to the final answer clause has length $2L$. We focus on the query positions $L+1, \dots, 2L$, which correspond to the answer clauses $Z_{\text{ans},0}, \dots, Z_{\text{ans},L-1}$ for predicting all outputs y_1, \dots, y_L .

- In Figure 2, the two diagonal structures in the upper region indicate attention concentrating on the answer clause $Z_{\text{ans},\ell}$ and on the predicate clause $Z_{\text{pred},\ell+1}$ when the query is $Z_{\text{ans},\ell}$, validating the attention concentration principle highlighted by Theorem 4.1 and Theorem 5.1.
- To ablate positional bias, we test on the trained model from Figure 2 and only change the input format by randomly permuting the locations of predicate clauses. Results in Figure 5 show that retrieval is keyed to the shared variable rather than absolute position.
- Figures 6 and 7 probe out-of-length (OOL) behavior through attention patterns at test lengths $L \in \{10, 20\}$. As the line plot indicates, the attention scores for both the target predicate and the answer clauses decrease as the task length increases from 10 to 20. This pattern matches the attention dilution predicted by our theory of length generalization. In Figure 6, for the symmetry group task, attention to the correct predicate $Z_{\text{pred},\ell+1}$ decays with position, becoming weaker at later predicates (larger ℓ , hence a larger absolute index $L+\ell$). By contrast, the cyclic group task shows no comparable decay, with attention remaining fairly uniform across lengths and positions. In Figure 7, the pattern reverses: attention over answer clauses is more stable for the symmetry group, whereas the cyclic case is more sensitive to length and location. We conjecture that length generalization in the LEGO task hinges on a robust predicate-attention pattern. Accordingly, the observed length-dependent decay of attention on predicate clauses in the symmetry task indicates a non-robust learned attention mechanism, which may underlie its poorer length generalization.

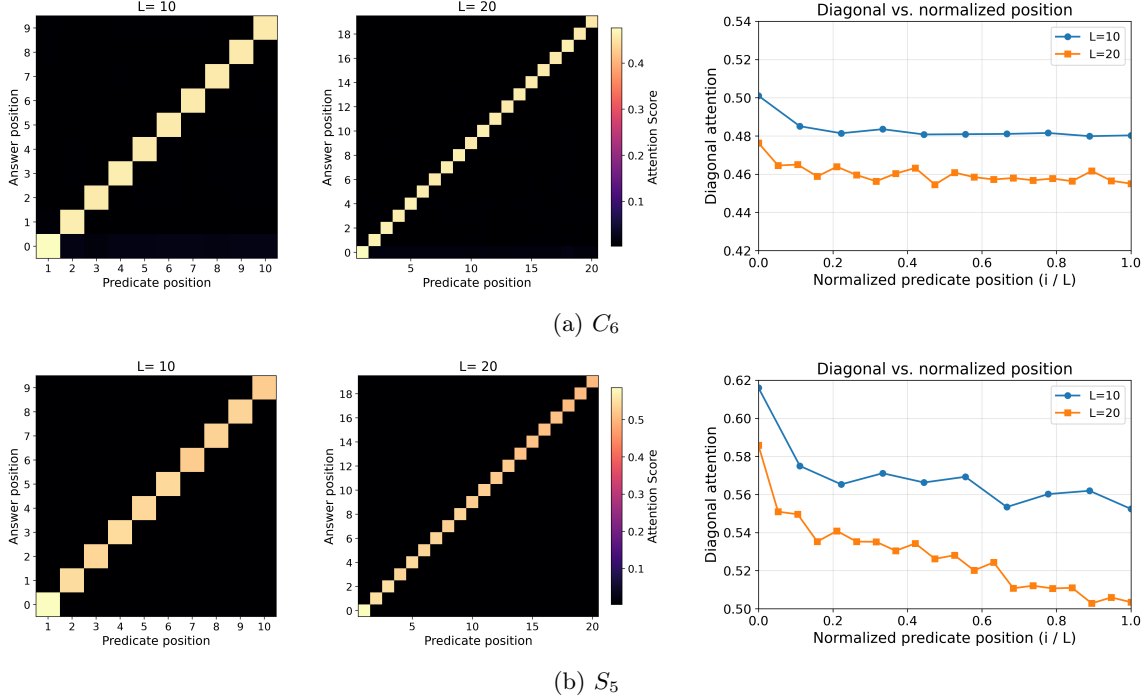


Figure 6: Predicate-clause attention for models trained with LEGO task of length $L = 5$ and tested with $L = 10, 20$. We extract the heatmap of the attention from answer queries to predicate clauses (the upper-left part of the full attention heatmap as in Figure 2), and line-plot the diagonal attention against the normalized predicate position. At longer contexts, the symmetry group task displays a visible drop in attention to the target predicate $Z_{\text{pred}, \ell+1}$ at later predicate locations (larger ℓ ; note the absolute query index $L + \ell$ also increases). By contrast, the cyclic group task maintains more consistent attention across lengths and positions.

8 Conclusions and Discussions

In this paper, we have theoretically analyzed how length-generalizable reasoning ability emerges during gradient-descent training of transformers on synthetic CoT tasks. To the best of our knowledge, our results provide the first optimization theory for learning CoT with length generalization guarantees. For tasks associated with simply transitive actions, we have proven that transformers can directly generalize from short constant-length chains to substantially longer tasks. For tasks associated with symmetry actions, we have established convergence guarantees showing that transformers can bootstrap their CoT length through recursive self-training. By addressing inherently sequential tasks beyond TC^0 , our optimization theory bridges the gap with known expressive power of transformers with CoT, while in the meantime rigorously validating the effectiveness of self-improvement training observed empirically. We wrap up this paper with a few discussions.

NoPE, recency bias, and length generalization. Empirically, standard positional embeddings often hinder length extrapolation, while NoPE has been advantageous considering its length-generalization performance [KPNR⁺23, All25]. Prior works [KS24, WZLZ24, HLY25] typically adopt fixed positional encodings, which tie the learned computation to the training horizon. Intuitively, positional embeddings inject location-specific biases that favor local neighborhoods, making longer inputs harder. In our setting, length generalization is possible because the model retrieves relevant information from long, unordered contexts by *content* (variables) rather than by position, as discussed in Section 6.2 and verified empirically in Figure 5. This provides concrete architectural guidance for practice and identifies positional embeddings as a plausible cause of observed failures to generalize to unseen lengths.

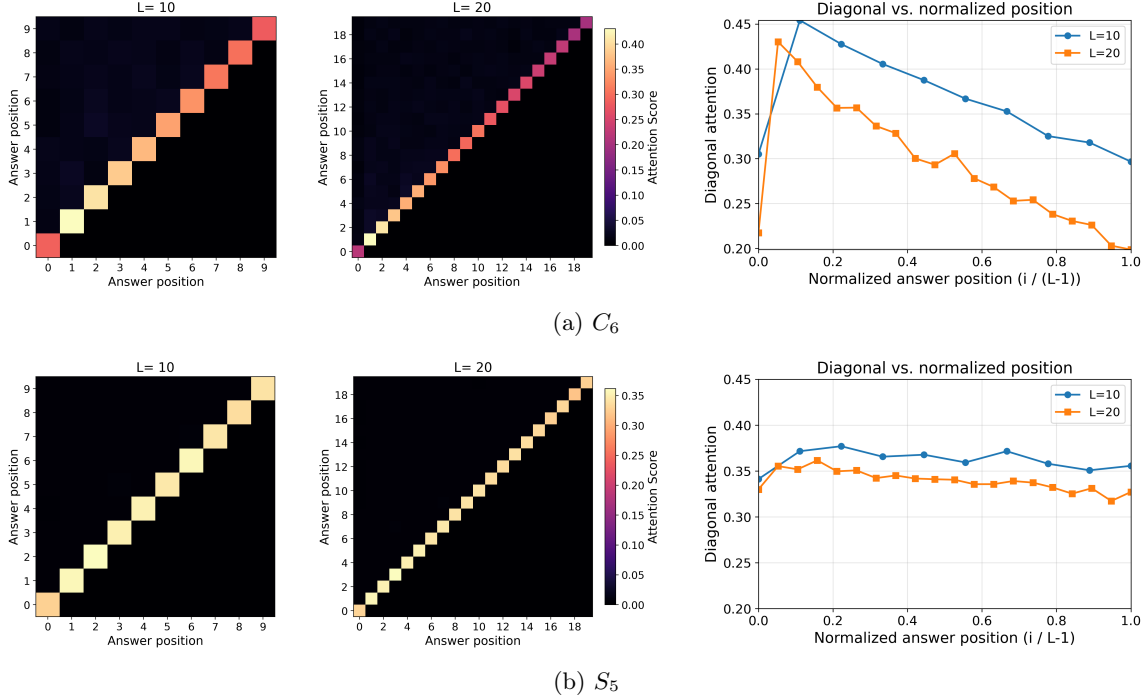


Figure 7: Answer-clause attention for models trained with LEGO task of length $L = 5$ and tested with $L = 10, 20$. We extract the heatmap of the attention from answer queries to answer clauses (the upper-right part of the full attention heatmap, as in Figure 2), and line-plot the diagonal attention against the normalized answer position. In contrast to the predicate attention in Figure 6, at longer contexts, the symmetry group task preserves a stable attention for the answer clause, while the answer attention pattern of the cyclic group task is sensitive to the lengths and positions. Nevertheless, the sharp decay of attention scores for the cyclic group task did not result in a visible performance drop for length generalization, as seen in Figure 1.

Local aggregation of tokens is beneficial. While NoPE is appealing for length generalization, it lacks position-aware horizontal mixing within a layer, which often hurts empirical performance. In practice, many strong models adopt RoPE (Rotary Positional Encoding) [SAL⁺24, PQFS23, DZZ⁺24], and recent seminal work [All25] introduces the *Cannon* layer, a short convolutional window that adds local residual links across neighboring tokens, to inject locality. Integrating Cannon layers substantially boosts the reasoning depth of NoPE transformers in synthetic experiments. Our clause-based data can be interpreted as mimicking the *output* of a Cannon layer: locality is built directly into the data structure, enabling one-layer transformers to learn induction heads, which otherwise require at least a two-layer transformer to express.

Context rot. Context rot is a phenomenon that the performance of LLMs drops as their input context grows even when the task is unchanged, which has been widely noted in practice [HTH25]. A prevailing empirical view is that longer contexts introduce many *distractors* and other irrelevant tokens, forcing the relevant signal to compete with them for attention and thereby weakening retrieval. Our attention concentration perspective characterizes this competition at a fine-grained level: we show that as irrelevant clauses accumulate, attention mass is diluted away from the correct clause, reducing performance at extended reasoning lengths. This offers a simple theoretical perspective on the origin of context rot and why long-context training hasn’t been as effective in eliminating the problem. We expect the insight from our analysis to extend to richer tasks and inspire practical mitigation strategies, e.g., *context engineering* [MYG⁺25].

Acknowledgments

The work of Z. Wen is supported in part by NSF DMS-2134080 and DMS-2134133. Y. Chen is supported in part by the Alfred P. Sloan Research Fellowship, the NSF grants IIS-2218713 and IIS-2218773, the ONR grants N00014-22-1-2354 and N00014-25-1-2344, the Wharton AI & Analytics Initiative’s AI Research Fund, and the Amazon Research Award. Z. Wen thanks Yuanzhi Li for initial contributions and helpful discussions on this project.

References

- [AB06] S. Arora and B. Barak. *Computational Complexity: A Modern Approach*. Cambridge University Press, 2006.
- [ABL⁺24] E. Abbe, S. Bengio, A. Lotfi, C. Sandon, and O. Saremi. How far can transformers reason? the globality barrier and inductive scratchpad. *Advances in Neural Information Processing Systems*, 37:27850–27895, 2024.
- [All25] Z. Allen-Zhu. Physics of Language Models: Part 4.1, Architecture Design and the Magic of Canon Layers. *SSRN Electronic Journal*, May 2025. <https://ssrn.com/abstract=5240330>.
- [AM24] K. Ahuja and A. Mansouri. On provable length and compositional generalization. *arXiv preprint arXiv:2402.04875*, 2024.
- [AWA⁺22] C. Anil, Y. Wu, A. Andreassen, A. Lewkowycz, V. Misra, V. Ramasesh, A. Slone, G. Gur-Ari, E. Dyer, and B. Neyshabur. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 35:38546–38556, 2022.
- [AZL20] Z. Allen-Zhu and Y. Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.
- [AZL21] Z. Allen-Zhu and Y. Li. Forward super-resolution: How can GANs learn hierarchical generative models for real-world distributions. *arXiv preprint arXiv:2106.02619*, 2021.
- [AZL22] Z. Allen-Zhu and Y. Li. Feature purification: How adversarial training performs robust deep learning. In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 977–988. IEEE, 2022.
- [Bar86] D. A. Barrington. Bounded-width polynomial-size branching programs recognize exactly those languages in nc. In *Proceedings of the eighteenth annual ACM symposium on Theory of computing*, pages 1–5, 1986.
- [BMR⁺20] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, and A. Askell. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- [CHX⁺25] Y. Cheng, Y. Huang, Z. Xiong, Y. Liang, and V. Y. Tan. Transformers provably learn directed acyclic graphs via kernel-guided mutual information. *arXiv preprint arXiv:2510.25542*, 2025.
- [CKB⁺21] K. Cobbe, V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, and R. Nakano. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.
- [CND⁺23] A. Chowdhery, S. Narang, J. Devlin, M. Bosma, G. Mishra, A. Roberts, P. Barham, H. W. Chung, C. Sutton, and S. Gehrmann. PaLM: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

- [CPW24] L. Chen, B. Peng, and H. Wu. Theoretical limitations of multi-layer transformer. *arXiv preprint arXiv:2412.02975*, 2024.
- [CSV84] A. K. Chandra, L. Stockmeyer, and U. Vishkin. Constant depth reducibility. *SIAM Journal on Computing*, 13(2):423–439, 1984.
- [CTJ⁺21] M. Chen, J. Tworek, H. Jun, Q. Yuan, H. P. D. O. Pinto, J. Kaplan, H. Edwards, Y. Burda, N. Joseph, and G. Brockman. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- [DA25] DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *ArXiv*, abs/2501.12948, 2025.
- [Deh11] M. Dehn. Über unendliche diskontinuierliche gruppen. *Mathematische Annalen*, 71(1):116–144, 1911.
- [Deh87] M. Dehn. On infinite discontinuous groups. In *Papers on Group Theory and Topology*, pages 133–178. Springer, 1987.
- [DLS⁺23] N. Dziri, X. Lu, M. Sclar, X. L. Li, L. Jiang, B. Y. Lin, S. Welleck, P. West, C. Bhagavatula, and R. Le Bras. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36:70293–70332, 2023.
- [DZZ⁺24] Y. Ding, L. L. Zhang, C. Zhang, Y. Xu, N. Shang, J. Xu, F. Yang, and M. Yang. Longrope: Extending llm context window beyond 2 million tokens. *arXiv preprint arXiv:2402.13753*, 2024.
- [ENO⁺21] N. Elhage, N. Nanda, C. Olsson, T. Henighan, N. Joseph, B. Mann, A. Askell, Y. Bai, A. Chen, T. Conerly, N. DasSarma, D. Drain, D. Ganguli, Z. Hatfield-Dodds, D. Hernandez, A. Jones, J. Kernion, L. Lovitt, K. Ndousse, D. Amodei, T. Brown, J. Clark, J. Kaplan, S. McCandlish, and C. Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021. <https://transformer-circuits.pub/2021/framework/index.html>.
- [Fra23] J. B. Fraleigh. *A first course in abstract algebra (8th Edition)*. Pearson India, 2023.
- [FZG⁺23] G. Feng, B. Zhang, Y. Gu, H. Ye, D. He, and L. Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:70757–70798, 2023.
- [GJB⁺25] N. Golowich, S. Jelassi, D. Brandfonbrener, S. M. Kakade, and E. Malach. The role of sparsity for length generalization in transformers. *arXiv preprint arXiv:2502.16792*, 2025.
- [GJKM25] T. Gao, J. Jin, Z. T. Ke, and G. Moryoussef. A comparison of DeepSeek and other LLMs. *arXiv preprint arXiv:2502.03688*, 2025.
- [GLL87] G. H. Godbeer, J. Lipscomb, and M. G. Luby. *On the computational complexity of finding stable state vectors in connectionist models (Hopfield nets)*. Department of Computer Science, University of Toronto, 1987.
- [GPS⁺23] C. Gulcehre, T. L. Paine, S. Srinivasan, K. Konyushkova, L. Weerts, A. Sharma, A. Siddhant, A. Ahern, M. Wang, and C. Gu. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.
- [GZA⁺23] S. Gunasekar, Y. Zhang, J. Aneja, C. C. T. Mendes, A. Del Giorno, S. Gopi, M. Javaheripi, P. Kauffmann, G. de Rosa, and O. Saarikivi. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.
- [HBF⁺24] A. Huang, A. Block, D. J. Foster, D. Rohatgi, C. Zhang, M. Simchowicz, J. T. Ash, and A. Krishnamurthy. Self-improvement in language models: The sharpening mechanism. *arXiv preprint arXiv:2412.01951*, 2024.

- [HBK⁺24] K. Hou, D. Brandfonbrener, S. Kakade, S. Jelassi, and E. Malach. Universal length generalization with turing programs. *arXiv preprint arXiv:2407.03310*, 2024.
- [HCL24] Y. Huang, Y. Cheng, and Y. Liang. In-context convergence of transformers. *International Conference on Machine Learning*, pages 19660–19722, 2024.
- [HLY25] R. Huang, Y. Liang, and J. Yang. How transformers learn regular language recognition: A theoretical study on training dynamics and implicit bias. *arXiv preprint arXiv:2505.00926*, 2025.
- [HLZ⁺22] Y. Huang, J. Lin, C. Zhou, H. Yang, and L. Huang. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International conference on machine learning*, pages 9226–9259. PMLR, 2022.
- [HSK⁺24] C.-P. Hsieh, S. Sun, S. Kriman, S. Acharya, D. Rekesh, F. Jia, Y. Zhang, and B. Ginsburg. Ruler: What’s the real context size of your long-context language models? *arXiv preprint arXiv:2404.06654*, 2024.
- [HSY22] N. Ho, L. Schmid, and S.-Y. Yun. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*, 2022.
- [HTH25] K. Hong, A. Troynikov, and J. Huber. Context rot: How increasing input tokens impacts LLM performance. Technical report, Chroma, July 2025.
- [HWCL25] Y. Huang, Z. Wen, Y. Chi, and Y. Liang. A theoretical analysis of self-supervised learning for vision transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [HWL25] J. Huang, Z. Wang, and J. D. Lee. Transformers learn to implement multi-step gradient descent with chain of thought. *arXiv preprint arXiv:2502.21212*, 2025.
- [HYB⁺24] X. Huang, A. Yang, S. Bhattamishra, Y. Sarrof, A. Krebs, H. Zhou, P. Nakkiran, and M. Hahn. A formal framework for understanding length generalization in transformers. *arXiv preprint arXiv:2410.02140*, 2024.
- [HZCY24] X. Hu, F. Zhang, S. Chen, and Z. Yang. Unveiling the statistical foundations of chain-of-thought prompting methods. *arXiv preprint arXiv:2408.14511*, 2024.
- [JB25] C. R. Jones and B. K. Bergen. Large language models pass the turing test. *arXiv preprint arXiv:2503.23674*, 2025.
- [JdDE⁺23] S. Jelassi, S. d’Ascoli, C. Domingo-Enrich, Y. Wu, Y. Li, and F. Charton. Length generalization in arithmetic transformers. *arXiv preprint arXiv:2306.15400*, 2023.
- [JSL22] S. Jelassi, M. Sander, and Y. Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.
- [JVB⁺25] N. Joshi, G. Vardi, A. Block, S. Goel, Z. Li, T. Misiakiewicz, and N. Srebro. A theory of learning with autoregressive chain of thought. *arXiv preprint arXiv:2503.07932*, 2025.
- [KB14] D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014.
- [KBA⁺24] Y. Kuratov, A. Bulatov, P. Anokhin, I. Rodkin, D. Sorokin, A. Sorokin, and M. Burtsev. Babilong: Testing the limits of llms with long context reasoning-in-a-haystack. *Advances in Neural Information Processing Systems*, 37:106519–106554, 2024.
- [KGR⁺22] T. Kojima, S. S. Gu, M. Reid, Y. Matsuo, and Y. Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

- [KPNR⁺23] A. Kazemnejad, I. Padhi, K. Natesan Ramamurthy, P. Das, and S. Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36:24892–24928, 2023.
- [KS23] N. Kim and S. Schuster. Entity tracking in language models. *arXiv preprint arXiv:2305.02363*, 2023.
- [KS24] J. Kim and T. Suzuki. Transformers provably solve parity efficiently with chain of thought. *arXiv preprint arXiv:2410.08633*, 2024.
- [KWLS25] J. Kim, D. Wu, J. Lee, and T. Suzuki. Metastable dynamics of chain-of-thought reasoning: Provable benefits of search, rl and distillation. *arXiv preprint arXiv:2502.01694*, 2025.
- [KZA⁺24] A. Kumar, V. Zhuang, R. Agarwal, Y. Su, J. D. Co-Reyes, A. Singh, K. Baumli, S. Iqbal, C. Bishop, and R. Roelofs. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024.
- [LAD⁺22] A. Lewkowycz, A. Andreassen, D. Dohan, E. Dyer, H. Michalewski, V. Ramasesh, A. Slone, C. Anil, I. Schlag, and T. Gutman-Solo. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.
- [LAG⁺22] B. Liu, J. T. Ash, S. Goel, A. Krishnamurthy, and C. Zhang. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022.
- [LCS⁺25] N. Lee, Z. Cai, A. Schwarzschild, K. Lee, and D. Papailiopoulos. Self-improving transformers overcome easy-to-hard and length generalization challenges. *arXiv preprint arXiv:2502.01612*, 2025.
- [LGA25] B. Z. Li, Z. C. Guo, and J. Andreas. (how) do language models track state? *ArXiv*, abs/2503.02854, 2025.
- [LLY⁺25] Z. Ling, K. Liu, K. Yan, Y. Yang, W. Lin, T.-H. Fan, L. Shen, Z. Du, and J. Chen. Longreason: A synthetic long-context reasoning benchmark via context expansion. *arXiv preprint arXiv:2501.15089*, 2025.
- [LLZM24] Z. Li, H. Liu, D. Zhou, and T. Ma. Chain of thought empowers transformers to solve inherently serial problems. *arXiv preprint arXiv:2402.12875*, 1, 2024.
- [LPW⁺17] Z. Lu, H. Pu, F. Wang, Z. Hu, and L. Wang. The expressive power of neural networks: A view from the width. *Advances in neural information processing systems*, 30, 2017.
- [LSG⁺23] Y. Li, K. Sreenivasan, A. Giannou, D. Papailiopoulos, and S. Oymak. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. *Advances in Neural Information Processing Systems*, 36:22021–22046, 2023.
- [Maa96] W. Maass. Lower bounds for the computational power of networks of spiking neurons. *Neural computation*, 8(1):1–40, 1996.
- [MCJ⁺24] Y. Min, Z. Chen, J. Jiang, J. Chen, J. Deng, Y. Hu, Y. Tang, J. Wang, X. Cheng, and H. Song. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*, 2024.
- [MDA⁺24] A. Marsden, E. Dogariu, N. Agarwal, X. Chen, D. Suo, and E. Hazan. Provable length generalization in sequence prediction via spectral filtering. *arXiv preprint arXiv:2411.01035*, 2024.
- [MMA⁺22] L. C. Magister, J. Mallinson, J. Adamek, E. Malmi, and A. Severyn. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*, 2022.

- [MPS24] W. Merrill, J. Petty, and A. Sabharwal. The illusion of state in state-space models. *arXiv preprint arXiv:2404.08819*, 2024.
- [MS23a] W. Merrill and A. Sabharwal. The expressive power of transformers with chain of thought. *arXiv preprint arXiv:2310.07923*, 2023.
- [MS23b] W. Merrill and A. Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023.
- [MS25a] W. Merrill and A. Sabharwal. Exact expressive power of transformers with padding. *arXiv preprint arXiv:2505.18948*, 2025.
- [MS25b] W. Merrill and A. Sabharwal. A little depth goes a long way: The expressive power of log-depth transformers. *arXiv preprint arXiv:2503.03961*, 2025.
- [MSS22] W. Merrill, A. Sabharwal, and N. A. Smith. Saturated transformers are constant-depth threshold circuits. *Transactions of the Association for Computational Linguistics*, 10:843–856, 2022.
- [MY22] A. Madaan and A. Yazdanbakhsh. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*, 2022.
- [MYG⁺25] L. Mei, J. Yao, Y. Ge, Y. Wang, B. Bi, Y. Cai, J. Liu, M. Li, Z.-Z. Li, and D. Zhang. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334*, 2025.
- [NDL24] E. Nichani, A. Damian, and J. D. Lee. How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*, 2024.
- [Ope24] OpenAI. OpenAI o1 system card. *ArXiv*, abs/2412.16720, 2024.
- [Par94] I. Parberry. *Circuit complexity and neural networks*. MIT press, 1994.
- [PLG23] B. Prystawski, M. Li, and N. Goodman. Why think step by step? reasoning emerges from the locality of experience. *Advances in Neural Information Processing Systems*, 36:70926–70947, 2023.
- [PQFS23] B. Peng, J. Quesnelle, H. Fan, and E. Shippole. Yarn: Efficient context window extension of large language models. *arXiv preprint arXiv:2309.00071*, 2023.
- [QLZ⁺24] Y. Qin, X. Li, H. Zou, Y. Liu, S. Xia, Z. Huang, Y. Ye, W. Yuan, H. Liu, and Y. Li. O1 replication journey: A strategic progress report—part 1. *arXiv preprint arXiv:2410.18982*, 2024.
- [Rin16] A. Rinaldo. 36-755: Advanced statistical theory, lecture 27. Lecture notes, December 5 2016. Scribed by Xiao Hui Tai.
- [RM21] L. Reynolds and K. McDonell. Prompt programming for large language models: Beyond the few-shot paradigm. In *Extended abstracts of the 2021 CHI conference on human factors in computing systems*, pages 1–7, 2021.
- [RS98] M. Raab and A. Steger. “balls into bins”—a simple and tight analysis. In *International Workshop on Randomization and Approximation Techniques in Computer Science*, pages 159–170. Springer, 1998.
- [RSR⁺20] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.
- [SAL⁺24] J. Su, M. Ahmed, Y. Lu, S. Pan, W. Bo, and Y. Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

- [SCRA⁺23] A. Singh, J. D. Co-Reyes, R. Agarwal, A. Anand, P. Patil, X. Garcia, P. J. Liu, J. Harrison, J. Lee, and K. Xu. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023.
- [SFH⁺24] C. Sanford, B. Fatemi, E. Hall, A. Tsitsulin, M. Kazemi, J. Halcrow, B. Perozzi, and V. Mirrokni. Understanding transformer reasoning capabilities via graph algorithms. *Advances in Neural Information Processing Systems*, 37:78320–78370, 2024.
- [SLZT25] Y. Sun, Y. Liang, Z. Zhang, and J. Teng. Theoretical modeling of llm self-improvement training dynamics through solver-verifier gap. *arXiv preprint arXiv:2507.00075*, 2025.
- [ŠO03] J. Šíma and P. Orponen. General-purpose computation with neural networks: A survey of complexity theoretic results. *Neural Computation*, 15(12):2727–2778, 2003.
- [SPHG24] M. Sabbaghi, G. Pappas, H. Hassani, and S. Goel. Explicitly encoding structural symmetry is key to length generalization in arithmetic tasks. *arXiv preprint arXiv:2406.01895*, 2024.
- [SS92] H. T. Siegelmann and E. D. Sontag. On the computational power of neural nets. In *Proceedings of the fifth annual workshop on Computational learning theory*, pages 440–449, 1992.
- [STAK92] J. S. Shawe-Taylor, M. H. Anthony, and W. Kern. Classes of feedforward neural networks and their circuit complexity. *Neural networks*, 5(6):971–977, 1992.
- [SUV18] P. Shaw, J. Uszkoreit, and A. Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.
- [SZE⁺24] Y. Song, H. Zhang, C. Eisenach, S. Kakade, D. Foster, and U. Ghai. Mind the gap: Examining the self-improvement capabilities of large language models. *arXiv preprint arXiv:2412.02674*, 2024.
- [tea25] A. A. A. team. Effective context engineering for ai agents. Technical report, Anthropic, 2025.
- [TLI⁺23] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, and F. Azhar. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- [Vol99] H. Vollmer. *Introduction to circuit complexity: a uniform approach*. Springer Science & Business Media, 1999.
- [VSP⁺17] A. Vaswani, N. M. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.
- [WL21] Z. Wen and Y. Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR, 2021.
- [WL22] Z. Wen and Y. Li. The mechanism of prediction head in non-contrastive self-supervised learning. *Advances in Neural Information Processing Systems*, 35:24794–24809, 2022.
- [WLS22] N. Wies, Y. Levine, and A. Shashua. Sub-task decomposition enables learning in sequence to sequence tasks. *arXiv preprint arXiv:2204.02892*, 2022.
- [WNB⁺25] Z. Wang, E. Nichani, A. Bietti, A. Damian, D. Hsu, J. D. Lee, and D. Wu. Learning compositional functions with transformers from easy-to-hard data. *arXiv preprint arXiv:2505.23683*, 2025.
- [WWS⁺22a] X. Wang, J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

- [WWS⁺22b] J. Wei, X. Wang, D. Schuurmans, M. Bosma, F. Xia, E. Chi, Q. V. Le, and D. Zhou. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.
- [WZLZ24] K. Wen, H. Zhang, H. Lin, and J. Zhang. From sparse dependence to sparse attention: unveiling how chain-of-thought enhances transformer sample efficiency. *arXiv preprint arXiv:2410.05459*, 2024.
- [XGZ⁺24] Y. Xie, A. Goyal, W. Zheng, M.-Y. Kan, T. P. Lillicrap, K. Kawaguchi, and M. Shieh. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*, 2024.
- [XL25] C. Xiao and B. Liu. Generalizing reasoning problems to longer lengths. In *The Thirteenth International Conference on Learning Representations*, 2025.
- [YHLC24] T. Yang, Y. Huang, Y. Liang, and Y. Chi. In-context learning with representations: Contextual generalization of trained transformers. *Advances in Neural Information Processing Systems*, 37:85867–85898, 2024.
- [YHLC25] T. Yang, Y. Huang, Y. Liang, and Y. Chi. Multi-head transformers provably learn symbolic multi-step reasoning via gradient descent. *arXiv preprint arXiv:2508.08222*, 2025.
- [YLJ⁺25] W. Yang, Z. Liu, H. Jin, Q. Yin, V. Chaudhary, and X. Han. Longer context, deeper thinking: Uncovering the role of long-context ability in reasoning. *arXiv preprint arXiv:2505.17315*, 2025.
- [YSL⁺25] Y. Yan, Y. Shen, Y. Liu, J. Jiang, M. Zhang, J. Shao, and Y. Zhuang. Infythink: Breaking the length limits of long-context reasoning in large language models. *arXiv preprint arXiv:2503.06692*, 2025.
- [ZAC⁺24] Y. Zhou, U. Alon, X. Chen, X. Wang, R. Agarwal, and D. Zhou. Transformers can achieve length generalization but not robustly. *arXiv preprint arXiv:2402.09371*, 2024.
- [ZBB⁺22] Y. Zhang, A. Backurs, S. Bubeck, R. Eldan, S. Gunasekar, and T. Wagner. Unveiling transformers with lego: a synthetic reasoning task. *arXiv preprint arXiv:2206.04301*, 2022.
- [ZBL⁺23] H. Zhou, A. Bradley, E. Littwin, N. Razin, O. Saremi, J. Susskind, S. Bengio, and P. Nakkiran. What algorithms can transformers learn? a study in length generalization. *arXiv preprint arXiv:2310.16028*, 2023.
- [ZLP⁺24] T. Zhong, Z. Liu, Y. Pan, Y. Zhang, Y. Zhou, S. Liang, Z. Wu, Y. Lyu, P. Shu, and X. Yu. Evaluation of OpenAI o1: Opportunities and challenges of AGI. *arXiv preprint arXiv:2409.18486*, 2024.
- [ZSH⁺22] D. Zhou, N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, and Q. Le. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

Appendix: Complete Proofs

A	Learning In-Context Retrieval of Variables	34
A.1	Preliminaries	34
A.2	Induction Hypothesis	34
A.3	Gradient Lemma	36
A.4	Growth of Gamma	37
A.5	Group and Value Correlations Are Not Large	38
A.6	Off-diagonal Correlations Are Small	39
A.7	Non-target Correlations Are Negligible	40
A.8	Convergence	40
B	Learning Simply Transitive Actions	41
B.1	Preliminaries	41
B.2	Phase I: Feature Emergence and Competition	46
B.3	Phase II: Feature Growth and Cancellations	49
B.4	Phase III: Convergence	55
B.5	Auxiliary Technical Tools	56
C	Learning Symmetry Group Actions	58
C.1	Preliminaries	58
C.2	Phase I: Emergence of the Feature	63
C.3	Phase II: Feature Growth and Cancellations	65
C.4	Phase III: Convergence	70
D	Learning the Attention Layer: Simply Transitive Case	71
D.1	Gradient Computations	71
D.2	Some Useful Bounds for Gradients	73
D.3	Stage 2.1: Initial Growth of Gap	81
D.4	Stage 2.2: Continual Growth of Diagonal Entries	91
D.5	Stage 2.3: Decrease of Gap and Convergence	98
D.6	Proof of Main Theorem	109
E	Learning the Attention Layer: Symmetry Case for Short-Length	110
E.1	Gradient Computations	110
E.2	Some Useful Bounds for Gradients	111
E.3	Stage 1.2.1: Initial Growth of \mathbf{Q}	116
E.4	Stage 1.2.2: Convergence with Small Wrong Attention	133
F	Recursive Learning the Attention Layer: Symmetry Case	145
F.1	Preliminaries	145
F.2	Reducing the Wrong Attention	150
F.3	Proof of Main Theorem	158

A Learning In-Context Retrieval of Variables

In this section, we focus on the learning process of \mathbf{W}_4 for the task \mathcal{T}^1 . For this task, \mathbf{W}_4 should predict the 4-th token in the answer clause $Z_{\text{ans},1}$, which is the variable x_1 . We will show that, the network learns to retrieve the target variable x_1 from the first token of the first predicate clause $Z_{\text{pred},1}$ to make an accurate prediction. Throughout the rest of the proof, we will omit the subscript for the expectation \mathbb{E} , when the context is clear.

A.1 Preliminaries

First we define some notations for the presentation of gradients.

Notations for gradient expressions For each $i \in [5], \ell \in [L], j \in [d], r \in [m]$, we denote

$$\mathcal{E}_{i,j}(\mathbf{Z}^{L,\ell-1}) \triangleq \mathbb{1}_{\mathbf{Z}_{\text{ans},\ell},i=e_j} - \text{logit}_{i,j}(F, \mathbf{Z}^{L,\ell-1}), \quad (22)$$

$$\Lambda_{i,j,r}(\mathbf{Z}^{L,\ell-1}) \triangleq \sum_{\mathbf{k} \in \mathcal{I}^{L,\ell-1}} \text{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}} \cdot \langle \mathbf{W}_{i,j,r}, \mathbf{Z}_{\mathbf{k}} \rangle + b_{i,j,r}. \quad (23)$$

where $\text{logit}_{i,j}(F, \mathbf{Z}^{L,\ell-1})$ are defined as

$$\text{logit}_{i,j}(F, \mathbf{Z}^{L,\ell-1}) := \frac{e^{F_{i,j}(\mathbf{Z}^{L,\ell-1})}}{\sum_{j' \in [d]} e^{F_{i,j'}(\mathbf{Z}^{L,\ell-1})}}.$$

Fact A.1. For any $i \in [5], j \in [d], r \in [m]$, we have the following gradient expression:

$$-\nabla_{\mathbf{W}_{i,j,r}} \text{Loss}^L = \mathbb{E} \left[\sum_{\ell=1}^L \mathcal{E}_{i,j}(\mathbf{Z}^{L,\ell-1}) \text{sReLU}'(\Lambda_{i,j,r}(\mathbf{Z}^{L,\ell-1})) \sum_{\mathbf{k} \in \mathcal{I}^{L,\ell-1}} \text{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}} \mathbf{Z}_{\mathbf{k}} \right].$$

For simplicity of notation, we will henceforth denote $\Lambda_{i,j,r}(\mathbf{Z}^{L,\ell-1})$ by $\Lambda_{i,j,r}$ and $\mathcal{E}_{i,j}(\mathbf{Z}^{L,\ell-1})$ by $\mathcal{E}_{i,j}$ when the context is clear.

Given \mathbf{Z}^L , we use $\hat{\mathcal{X}}^L$ to denote the appeared variables in the context clauses, i.e. $\hat{\mathcal{X}}^L = \{x_0, x_1, \dots, x_L\}$. We write $\hat{\mathcal{X}}^L$ as $\hat{\mathcal{X}}$ for simplicity. Throughout this section, we write $[F_i]_j$ as $F_{i,j}$ for simplicity.

A.2 Induction Hypothesis

Proof Plan. The main idea is to track the dynamics of different types of weights $\mathbf{W}_{4,j,r,p}$. Specifically, we prove that for each $j \in \tau(\mathcal{X})$, there exists certain neurons $r \in [m]$ such that the corresponding weights $\mathbf{W}_{4,j,r,1}$ grow significantly along the direction e_j , while all others remain small. Specifically, we proceed in four steps:

1. For $j \in \tau(\mathcal{X})$, define

$$\Gamma_{4,j}^{(t)} \triangleq \max_{r \in [m]} \langle \mathbf{W}_{4,j,r,1}^{(t)}, e_j \rangle + \sigma_0 \log d,$$

to track the maximal activation associated with retrieving the correct variable token.

2. **Establish rapid growth (early phase).** Let $\Lambda^- = \Theta(1/m)$, and define the hitting time

$$T_{1,j} \triangleq \min\{t > 0 : \Gamma_{4,j}^{(t)} \geq \Lambda^-\}.$$

We show that for iterations $t \leq T_1 = \Theta(d\sigma_0^{q-2}/\eta)$, the diagonal weights $\langle \mathbf{W}_{4,j,r,1}^{(t)}, e_j \rangle$ grow rapidly, causing $\Gamma_{4,j}^{(t)}$ to enter a linear growth regime. Simultaneously, the model confidently identifies the correct variable, indicated by $1 - \text{logit}_{4,\tau(x_1)}^{(t)} = 1 - o(1)$.

3. **Convergence via dominant neurons (late phase).** For each $j \in \tau(\mathcal{X})$, define active neuron sets and their total activation as:

$$\mathcal{A}_{4,j}^{(t)} \triangleq \{r \in [m] : \langle \mathbf{W}_{4,j,r,1}^{(t)}, e_j \rangle \geq \varrho \log d\}, \quad \Phi_{4,j}^{(t)} \triangleq \sum_{r \in \mathcal{A}_{4,j}^{(t)}} \langle \mathbf{W}_{4,j,r,1}^{(t)}, e_j \rangle.$$

For iterations $t > T_1$, we analyze the refined dynamics, proving that the total diagonal activation $\Phi_{4,j^*}^{(t)}$, for the weakest activated variable j^* , eventually grows to $\Theta(\log d)$, ensuring successful learning.

4. **Bounding non-target correlations.** We finally show by induction that all other correlations $\langle \mathbf{W}_{4,j,r,p}^{(t)}, e_s \rangle$, including group actions, value tokens, off-diagonal tokens, and non-target variables, remain negligible throughout the training process.

Our proof begins by positing an induction hypothesis expected to hold throughout training. We then analyze the dynamics under this hypothesis and show that it remains valid along the entire training trajectory, establishing the claim at convergence.

Induction A.1. For $t \leq T = \frac{\text{poly}d}{\eta}$, all of the following holds:

- (a). for $j \in \tau(\mathcal{X})$, $\tilde{\Omega}(\sigma_0) \leq \langle \mathbf{W}_{4,j,r,1}^{(t)}, e_j \rangle + \mu \leq \tilde{O}(1)$, where $\langle \mathbf{W}_{4,j,r,1}^{(t)}, e_j \rangle$ is non-decreasing;
- (b). for $j \in \tau(\mathcal{X})$, $g \in \mathcal{G}$

$$|\langle \mathbf{W}_{4,j,r,2}^{(t)}, e_{\tau(g)} \rangle| \leq \tilde{O}(\sigma_0) + O\left(\frac{1}{|\mathcal{G}|}\right) \max \left\{ \langle \mathbf{W}_{4,j,r,1}^{(t)}, e_j \rangle, \min_{r' \in \mathcal{A}_{4,j^*}^{(t)}} \langle \mathbf{W}_{4,j^*,r',1}^{(t)}, e_{j^*} \rangle \right\};$$

- (c). for $j \in \tau(\mathcal{X})$, $y \in \mathcal{Y}$

$$|\langle \mathbf{W}_{4,j,r,5}^{(t)}, e_{\tau(y)} \rangle| \leq \tilde{O}(\sigma_0) + O\left(\frac{1}{|\mathcal{Y}|}\right) \max \left\{ \langle \mathbf{W}_{4,j,r,1}^{(t)}, e_j \rangle, \min_{r' \in \mathcal{A}_{4,j^*}^{(t)}} \langle \mathbf{W}_{4,j^*,r',1}^{(t)}, e_{j^*} \rangle \right\};$$

- (d). else, $|\langle \mathbf{W}_{4,j,r,p}^{(t)}, e_{j'} \rangle| \leq \tilde{O}(\sigma_0)$ for any other $j, j' \in [d]$.

Claim A.1. If Induction A.1 holds at iteration t , then for an input $\mathbf{Z}^{1,0}$, we have

- 1. if $j = \tau(x_1)$,

$$\Lambda_{4,j,r}^{(t)} = \frac{1}{2} \langle \mathbf{W}_{4,j,r,2}^{(t)}, e_j \rangle + \frac{1}{2} \langle \mathbf{W}_{4,j,r,2}^{(t)}, e_{\tau(g_1)} \rangle + \frac{1}{2} \langle \mathbf{W}_{4,j,r,5}^{(t)}, e_{\tau(y_0)} \rangle + \frac{5}{2} \mu + \tilde{O}(\sigma_0);$$

- 2. else if $j \in \tau(\mathcal{X} \setminus \{x_1\})$,

$$\Lambda_{4,j,r}^{(t)} = \frac{1}{2} \langle \mathbf{W}_{4,j,r,2}^{(t)}, e_{\tau(g_1)} \rangle + \frac{1}{2} \langle \mathbf{W}_{4,j,r,5}^{(t)}, e_{\tau(y_0)} \rangle + \frac{5}{2} \mu + \tilde{O}(\sigma_0);$$

- 3. otherwise, $0 \leq \Lambda_{4,j,r}^{(t)} \leq \frac{5}{2} \mu + \tilde{O}(\sigma_0)$.

Claim A.2. If Induction A.1 holds at iteration t , then for an input $\mathbf{Z}^{1,0}$,

- 1. if $j = \tau(x_1)$, $\text{logit}_{4,j}^{(t)} = \frac{e^{O(\Phi_{4,j}^{(t)})}}{e^{O(\Phi_{4,j}^{(t)})} + d}$;
- 2. otherwise, $\text{logit}_{4,j}^{(t)} = O\left(\frac{1}{d}\right) \left(1 - \text{logit}_{\tau(x_1)}^{(t)}\right)$.

Proof. If $j = \tau(x_1)$, by Induction A.1 and Claim A.1, we have

$$\begin{aligned} 0 \leq F_{4,j}^{(t)}(\mathbf{Z}^{1,0}) &\leq \sum_{r \in [m]} [\Lambda_{4,j,r}^{(t)}]^+ \leq (\Phi_{4,j}^{(t)} + O(\frac{\max\{\Phi_{4,j}^{(t)}, \Phi_{4,j^*}^{(t)}\}}{|\mathcal{G}|})) + \tilde{O}(\sigma_0) + O(m\varrho \log d) \\ &= (\Phi_{4,j}^{(t)} + O(\frac{\Phi_{4,j}^{(t)}}{|\mathcal{G}|})) + \tilde{O}(\sigma_0) + O(\frac{1}{\text{polylog} d}). \end{aligned}$$

For $j \in \tau(\mathcal{X}) \neq \tau(x_1)$, $F_{4,j}^{(t)}(\mathbf{Z}^{1,0}) \leq \tilde{O}(\sigma_0) + O(\frac{\max\{\Phi_{4,j}^{(t)}, \Phi_{4,j^*}^{(t)}\}}{|\mathcal{G}|})$; else $F_{4,j}^{(t)}(\mathbf{Z}^{1,0}) \leq \tilde{O}(\sigma_0)$. Combining them together, we complete the proof. \square

A.3 Gradient Lemma

Starting with the gradient computation from Fact A.1:

$$-\nabla_{\mathbf{W}_{4,j,r,p}} \text{Loss}^1 = \frac{1}{2} \mathbb{E} \left[\mathcal{E}_{4,j} \mathbf{sReLU}'(\Lambda_{4,j,r}) \sum_{\mathbf{k} \in \mathcal{I}^{1,0}} \mathbf{Z}_{\mathbf{k},p} \right],$$

we first consider the gradient for $j \in \tau(\mathcal{X})$

Lemma A.1. *For $j \in \tau(\mathcal{X})$, we have*

(a) for $\mathbf{W}_{4,j,r,1}$, $s \in \tau(\mathcal{X})$

$$\begin{aligned} (1) \text{ if } s = j, \langle -\nabla_{\mathbf{W}_{4,j,r,1}}^{(t)} \text{Loss}^1, e_s \rangle &= \frac{1}{2} \mathbb{E} \left[(1 - \text{logit}_{4,j}^{(t)}) \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_1)=j} \right]; \\ (2) \text{ } s \neq j, \langle -\nabla_{\mathbf{W}_{4,j,r,1}}^{(t)} \text{Loss}^1, e_s \rangle &= \frac{1}{2} \mathbb{E} \left[-\text{logit}_{4,j}^{(t)} \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_1)=s} \right]. \end{aligned}$$

(b) for $\mathbf{W}_{4,j,r,2}$, $s = \tau(g)$ for $g \in \mathcal{G}$

$$\begin{aligned} \langle -\nabla_{\mathbf{W}_{4,j,r,2}}^{(t)} \text{Loss}^1, e_s \rangle &= \frac{1}{2} \mathbb{E} \left[(1 - \text{logit}_{4,j}^{(t)}) \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_1)=j, g_1=g} \right. \\ &\quad \left. - \text{logit}_{4,j}^{(t)} \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_1) \neq j, g_1=g} \right]. \end{aligned}$$

(c) for $\mathbf{W}_{4,j,r,p}$ with $p \in \{3, 4\}$, $s \in \tau(\mathcal{X})$

$$\begin{aligned} (1) \text{ } s = j, \langle -\nabla_{\mathbf{W}_{4,j,r,3}}^{(t)} \text{Loss}^1, e_j \rangle &= \frac{1}{2} \mathbb{E} \left[-\text{logit}_{4,j}^{(t)} \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_0)=j} \right]; \\ (2) \text{ } s \neq j \end{aligned}$$

$$\begin{aligned} \langle -\nabla_{\mathbf{W}_{4,j,r,3}}^{(t)} \text{Loss}^1, e_s \rangle &= \frac{1}{2} \mathbb{E} \left[(1 - \text{logit}_{4,j}^{(t)}) \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_0)=s, \tau(x_1)=j} \right. \\ &\quad \left. - \text{logit}_{4,j}^{(t)} \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_0)=s, j \notin \tau(\hat{X})} \right]. \end{aligned}$$

(d) for $\mathbf{W}_{4,j,r,5}$, $s = \tau(y)$ for $y \in \mathcal{Y}$

$$\begin{aligned} \langle -\nabla_{\mathbf{W}_{4,j,r,5}}^{(t)} \text{Loss}^1, e_s \rangle &= \frac{1}{2} \mathbb{E} \left[(1 - \text{logit}_{4,j}^{(t)}) \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_1)=j, y_0=y} \right. \\ &\quad \left. - \text{logit}_{4,j}^{(t)} \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_1) \neq j, y_0=y} \right]. \end{aligned}$$

Move on to $j \notin \tau(\mathcal{X})$, we can obtain

Lemma A.2. *For $j \notin \tau(\mathcal{X})$, we have*

(a) for $\mathbf{W}_{4,j,r,1}$, $s \in \tau(\mathcal{X})$,

$$\langle -\nabla_{\mathbf{W}_{4,j,r,1}^{(t)}} \text{Loss}^1, e_s \rangle = \frac{1}{2} \mathbb{E} \left[-\text{logit}_{4,j}^{(t)} \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbf{1}_{\tau(x_1)=s} \right].$$

(b) for $\mathbf{W}_{4,j,r,2}$, $s = \tau(g)$ for $g \in \mathcal{G}$,

$$\langle -\nabla_{\mathbf{W}_{4,j,r,2}^{(t)}} \text{Loss}^1, e_s \rangle = \frac{1}{2} \mathbb{E} \left[-\text{logit}_{4,j}^{(t)} \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbf{1}_{g_1=g} \right].$$

(c) for $\mathbf{W}_{4,j,r,p}$ with $p \in \{3, 4\}$, $s \in \tau(\mathcal{X})$,

$$\langle -\nabla_{\mathbf{W}_{4,j,r,p}^{(t)}} \text{Loss}^1, e_j \rangle = \frac{1}{2} \mathbb{E} \left[-\text{logit}_{4,j}^{(t)} \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbf{1}_{\tau(x_0)=s} \right].$$

(d) for $\mathbf{W}_{4,j,r,5}$, $s = \tau(y)$ for $y \in \mathcal{Y}$

$$\langle -\nabla_{\mathbf{W}_{4,j,r,5}^{(t)}} \text{Loss}^1, e_s \rangle = \frac{1}{2} \mathbb{E} \left[-\text{logit}_{4,j}^{(t)} \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbf{1}_{y_0=y} \right].$$

A.4 Growth of Gamma

Lemma A.3 (Growth). *Given $j \in \tau(\mathcal{X})$, suppose Induction A.1 holds at iteration t , when $\Phi_{4,j}^{(t)} \leq 0.01 \log d$ or $\Gamma_{4,j}^{(t)} \leq \frac{0.01 \log d}{m}$, then it satisfies*

$$\Gamma_{4,j}^{(t+1)} = \Gamma_{4,j}^{(t)} + \Theta\left(\frac{\eta}{d}\right) \text{sReLU}'(\Gamma_{4,j}^{(t)}).$$

Proof. By Lemma A.1, we have

$$\langle -\nabla_{\mathbf{W}_{4,j,r,1}^{(t)}} \text{Loss}^1, e_j \rangle = \frac{1}{2} \mathbb{E} \left[(1 - \text{logit}_{4,j}^{(t)}) \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbf{1}_{\tau(x_1)=j} \right].$$

By Claim A.2, when $\Phi_{4,j}^{(t)} \leq 0.01 \log d$, $\text{logit}_{4,j}^{(t)} = \frac{O(e^{0.01 \log d})}{O(e^{0.01 \log d}) + d} \ll 1$ when $j = \tau(x_1)$; and combining with the fact that the event $\{\tau(x_1) = j\}$ happens with probability $\frac{1}{|\mathcal{X}|}$, we complete the proof. \square

Lemma A.3, combined with the growth of the tensor power method, immediately gives the following corollary.

Lemma A.4. *Suppose Induction A.1 holds for all iterations. Define threshold $\Lambda^- = \Theta(\frac{1}{m})$. Let $T_{1,j}$ be the first iteration so that $\Gamma_{4,j}^{(t)} \geq \Lambda^-$, and $T_1 \stackrel{\text{def}}{=} \Theta(\frac{d}{\eta \sigma_0^{q-2}})$. Then we have $T_1 \geq T_{1,j}$ for every $j \in \tau(\mathcal{X})$, i.e., for $t \geq T_1$, it satisfies $\Gamma_{4,j}^{(t)} \geq \Lambda^-$.*

Lemma A.5 (Upper bound). *Suppose Induction A.1 holds for all iterations $< t$, we have $\Phi_{4,j}^{(t)} \leq \tilde{O}(1)$, for $j \in \tau(\mathcal{X})$.*

Proof. We only need to consider the time $t \geq T_1$. Notice that the gradient descent update in Lemma A.1 gives

$$\langle -\nabla_{\mathbf{W}_{4,j,r,1}^{(t)}} \text{Loss}^1, e_j \rangle = \frac{1}{2} \mathbb{E} \left[(1 - \text{logit}_{4,j}^{(t)}) \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbf{1}_{\tau(x_1)=j} \right]$$

Therefore, for sufficiently small η , we have

$$\Phi_{4,j}^{(t+1)} = \Phi_{4,j}^{(t)} + \sum_{r \in \mathcal{A}_{4,j}^{(t)}} \frac{\eta}{2} \mathbb{E} \left[(1 - \text{logit}_{4,j}^{(t)}) \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbf{1}_{\tau(x_1)=j} \right] + O(\varrho \log d) \cdot |\mathcal{A}_{4,j}^{(t+1)} \setminus \mathcal{A}_{4,j}^{(t)}|$$

$$= \Phi_{4,j}^{(t)} + \sum_{r \in \mathcal{A}_{4,j}^{(t)}} \frac{\eta}{2} \mathbb{E} \left[(1 - \mathbf{logit}_{4,j}^{(t)}) \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_1)=j} \right] + \frac{1}{\text{polylog} d}.$$

When there exists \tilde{T} , s.t., $\max_{j \in \tau(\mathcal{X})} \Phi_{4,j}^{(\tilde{T})} > \Omega(\log^{1.5} d)$, by Induction A.1 and Claim A.1, given an input sequence $\mathbf{Z}^{1,0}$ with $\tau(x_1) = \tilde{j} = \arg \max_{j \in \tau(\mathcal{X})} \Phi_{4,j}^{(\tilde{T})}$, we have

$$F_{4,j}^{(\tilde{T})}(\mathbf{Z}^{1,0}) \geq \sum_{r \in \mathcal{A}_{4,j}^{(t)}} \Lambda_{4,j,r}^{(\tilde{T})} \geq \left(1 - O\left(\frac{1}{|\mathcal{G}|}\right)\right) \Phi_{4,j}^{(\tilde{T})} - \tilde{O}(\sigma_0) > \Omega(\log^{1.5} d).$$

Following the similar analysis as Claim A.2, $F_{4,j'}^{(\tilde{T})}(\mathbf{Z}^{1,0}) \leq O\left(\frac{\Phi_{4,j'}^{(\tilde{T})}}{|\mathcal{G}|}\right)$ for other $j' \in \tau(\mathcal{X})$, and $F_{4,j'}^{(\tilde{T})}(\mathbf{Z}) \leq o(1)$ for $j' \notin \tau(\mathcal{X})$, which implies $1 - \mathbf{logit}_{4,j}^{(\tilde{T})} = e^{-\Omega(\log^{1.5} d)}$. Therefore, we derive that for $t \in [\tilde{T} + 1, \frac{\text{poly} d}{\eta}]$,

$$\Phi_{4,j}^{(t)} \leq \Phi_{4,j}^{(\tilde{T})} + \tilde{O}(\text{poly} d \cdot e^{-\Omega(\log^{1.5} d)}) + O(\rho \log d) \cdot m,$$

which implies $\Phi_{4,j}^{(t)} \leq O(\log^{1.5} d)$ since $\rho \ll \frac{1}{m \log d}$. \square

A.5 Group and Value Correlations Are Not Large

Lemma A.6. Suppose Induction A.1 holds for all iterations $< t$, then for any $j \in \tau(\mathcal{X})$ and $s = \tau(g), g \in \mathcal{G}$, we have

$$|\langle \mathbf{W}_{4,j,r,2}^{(t)}, e_{\tau(g)} \rangle| \leq \tilde{O}(\sigma_0) + O\left(\frac{1}{|\mathcal{G}|}\right) \max \left\{ \langle \mathbf{W}_{4,j,r,1}^{(t)}, e_j \rangle, \min_{r' \in \mathcal{A}_{4,j*}^{(t)}} \langle \mathbf{W}_{4,j*,r',1}^{(t)}, e_{j*} \rangle \right\}$$

Proof. By Lemma A.1, we have

$$\begin{aligned} & \langle -\nabla_{\mathbf{W}_{4,j,r,2}^{(t)}} \text{Loss}^1, e_s \rangle \\ &= \frac{1}{2} \mathbb{E} \left[(1 - \mathbf{logit}_{4,j}^{(t)}) \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_1)=j, g_1=g} - \mathbf{logit}_{4,j}^{(t)} \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_1) \neq j, g_1=g} \right]. \end{aligned}$$

Clearly, the positive gradient can be upper-bounded by $O\left(\frac{1}{|\mathcal{G}|} \langle -\nabla_{\mathbf{W}_{4,j,r,1}^{(t)}} \text{Loss}^1, e_j \rangle\right)$. Moreover, for the negative gradient, by Claim A.1, we have a naive bound

$$\mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)})|_{\tau(x_1) \neq j, g_1=g} \leq O(1) \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)})|_{\tau(x_1)=j, g_1=g}.$$

When $t \leq T_1$, by Claim A.2, we have $1 - \mathbf{logit}_{4,j}^{(t)}|_{j=\tau(x_1)} \geq \Omega(1)$ and $\mathbf{logit}_{4,j}^{(t)}|_{j \neq \tau(x_1)} \leq O(\frac{1}{d})$, which implies

$$\mathbb{E} \left[\mathbf{logit}_{4,j}^{(t)} \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_1) \neq j, g_1=g} \right] \leq O\left(\frac{1}{|\mathcal{G}|}\right) \langle -\nabla_{\mathbf{W}_{4,j,r,1}^{(t)}} \text{Loss}^1, e_j \rangle.$$

Therefore, for $t \leq T_1$, we have

$$|\langle \mathbf{W}_{4,j,r,2}^{(t)}, e_{\tau(g)} \rangle| \leq \tilde{O}(\sigma_0) + O\left(\frac{1}{|\mathcal{G}|}\right) \langle \mathbf{W}_{4,j,r,1}^{(t)}, e_j \rangle.$$

For $t \geq T_1$, notice that by Lemma A.4, $\mathcal{A}_{4,j'}^{(t)} \neq \emptyset$ for $j' \in \tau(\mathcal{X})$, thus for $r' \in \mathcal{A}_{4,j*}^{(t)}$

$$\mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)})|_{\tau(x_1) \neq j, g_1=g} \leq \mathbf{sReLU}'(\Lambda_{4,j*,r'}^{(t)})|_{\tau(x_1)=j*, g_1=g}.$$

Furthermore, $\mathbf{logit}_{4,j}^{(t)}|_{j \neq \tau(x_1)} \leq O(\frac{1}{d})(1 - \mathbf{logit}_{4,j^*}^{(t)}|_{j^* = \tau(x_1)})$, which implies

$$\mathbb{E} \left[\mathbf{logit}_{4,j}^{(t)} \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbf{1}_{\tau(x_1) \neq j, g_1 = g} \right] \leq O\left(\frac{1}{|\mathcal{G}|}\right) \langle -\nabla_{\mathbf{W}_{4,j^*,r',1}} \text{Loss}^1, e_{j^*} \rangle.$$

Due to the arbitrary of r' , we can conclude that

$$|\langle \mathbf{W}_{4,j,r,2}^{(t)}, e_{\tau(g)} \rangle| \leq \tilde{O}(\sigma_0) + O\left(\frac{1}{|\mathcal{G}|}\right) \min_{r' \in \mathcal{A}_{4,j^*}^{(t)}} \langle \mathbf{W}_{4,j^*,r',1}^{(t)}, e_{j^*} \rangle.$$

□

Lemma A.7. Suppose Induction A.1 holds for all iterations $< t$, then for any $j \in \tau(\mathcal{X})$ and $s = \tau(y), y \in \mathcal{Y}$, we have

$$|\langle \mathbf{W}_{4,j,r,5}^{(t)}, e_{\tau(y)} \rangle| \leq \tilde{O}(\sigma_0) + O\left(\frac{1}{|\mathcal{Y}|}\right) \max \left\{ \langle \mathbf{W}_{4,j,r,1}^{(t)}, e_j \rangle, \min_{r' \in \mathcal{A}_{4,j^*}^{(t)}} \langle \mathbf{W}_{4,j^*,r',1}^{(t)}, e_{j^*} \rangle \right\}.$$

Proof. The proof is similar to Lemma A.6, and we omit the details here. □

A.6 Off-diagonal Correlations Are Small

Lemma A.8 (off-diagonal bound). Given $j \in \tau(\mathcal{X})$, suppose Induction A.1 holds at all iterations $< t$, for $s \in \tau(\mathcal{X}) \neq j$

$$|\langle \mathbf{W}_{4,j,r,1}^{(t)}, e_s \rangle| \leq \tilde{O}(\sigma_0).$$

Proof. By Lemma A.1

$$\langle -\nabla_{\mathbf{W}_{4,j,r,1}^{(t)}} \text{Loss}^1, e_s \rangle = \frac{1}{2} \mathbb{E} \left[-\mathbf{logit}_{4,j}^{(t)} \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbf{1}_{\tau(x_1) = s} \right].$$

Notice that by Claim A.1,

$$\mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)})|_{\tau(x_1)=s} \leq O(1) \mathbf{sReLU}'(\Lambda_{4,s,r}^{(t)})|_{\tau(x_1)=s},$$

combined with Claim A.2, $\mathbf{logit}_{4,j} \leq O(\frac{1}{d})(1 - \mathbf{logit}_{4,s}^{(t)})$ when $s = \tau(x_1)$, thus

$$\begin{aligned} \mathbb{E} \left[\mathbf{logit}_{4,j}^{(t)} \mathbf{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbf{1}_{\tau(x_1)=s} \right] &\leq \mathbb{E} \left[O\left(\frac{1}{d}\right)(1 - \mathbf{logit}_{4,s}^{(t)}) \mathbf{sReLU}'(\Lambda_{4,s,r}^{(t)}) \mathbf{1}_{\tau(x_1)=s} \right] \\ &\leq O\left(\frac{1}{d}\right) \langle -\nabla_{\mathbf{W}_{4,s,r,1}} \text{Loss}, e_s \rangle. \end{aligned}$$

From Induction A.1, we have

$$|\langle \mathbf{W}_{4,j,r,1}^{(t)}, e_s \rangle| \leq O\left(\frac{1}{d}\right) |\langle \mathbf{W}_{4,s,r,1}^{(t)}, e_s \rangle| + \tilde{O}(\sigma_0) \leq \tilde{O}\left(\frac{1}{d}\right) + \tilde{O}(\sigma_0) = \tilde{O}(\sigma_0).$$

□

Lemma A.9. Given $j \in \tau(\mathcal{X})$, suppose Induction A.1 holds at all iterations $< t$, we have

$$|\langle \mathbf{W}_{4,j,r,p}^{(t)} \text{Loss}^1, e_s \rangle| \leq \tilde{O}(\sigma_0), \quad \text{for } p \in \{3, 4\} \text{ and all } s \in \tau(\mathcal{X})$$

Proof. When $s = j$, we have

$$\begin{aligned}\langle -\nabla_{\mathbf{W}_{4,j,r,p}} \text{Loss}^1, e_j \rangle &= \frac{1}{2} \mathbb{E} \left[-\text{logit}_{4,j}^{(t)} \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbf{1}_{\tau(x_0)=j} \right] \\ &= \frac{1}{2} \mathbb{E} \left[-\text{logit}_{4,j}^{(t)} \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \sum_{s \neq j} \mathbf{1}_{\tau(x_0)=j, \tau(x_1)=s} \right].\end{aligned}$$

Therefore, we can bound the above gradient in the similar way as the off-diagonal case, and obtain

$$|\langle \mathbf{W}_{4,j,r,p}^{(t)}, e_j \rangle| \leq O\left(\frac{1}{d}\right) \max_{s \in \tau(\mathcal{X})} |\langle \mathbf{W}_{4,s,r,1}^{(t)}, e_s \rangle| + \tilde{O}(\sigma_0) \leq \tilde{O}(\sigma_0).$$

When $s \neq j$,

$$\begin{aligned}\langle -\nabla_{\mathbf{W}_{4,j,r,p}} \text{Loss}^1, e_s \rangle &= \frac{1}{2} \mathbb{E} \left[(1 - \text{logit}_{4,j}^{(t)}) \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbf{1}_{\tau(x_0)=s, \tau(x_1)=j} - \text{logit}_{4,j}^{(t)} \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbf{1}_{\tau(x_0)=s, j \notin \tau(\hat{\mathcal{X}})} \right].\end{aligned}$$

Noticing that $\{\tau(x_0) = s, \tau(x_1) = j\}$ happens with probability $\frac{1}{|\mathcal{X}|(|\mathcal{X}|-1)}$, thus the positive gradient can be upper bounded by $O(\frac{1}{d}) \cdot |\langle -\nabla_{\mathbf{W}_{4,j,r,1}} \text{Loss}^1, e_j \rangle|$. Furthermore, the negative part can be upper bounded in the similar way as previous off-diagonal negative gradient. Putting it together, we complete the proof. \square

A.7 Non-target Correlations Are Negligible

Lemma A.10. Suppose Induction A.1 holds at all iterations $< t$, for $j' \notin \tau(\mathcal{X})$, for $p \in [5]$ and $s \in [d]$

$$|\langle \mathbf{W}_{4,j',r,p}^{(t)}, e_s \rangle| \leq \tilde{O}(\sigma_0).$$

Proof. By Lemma A.2, $\langle \mathbf{W}_{4,j',r,p}^{(t)}, e_s \rangle$ for $p \in \{1, 3, 4\}$ and $s \in \tau(\mathcal{X})$ can be bounded in the similar way previous off-diagonal negative gradient.

We can observe that for $j' \notin \tau(\mathcal{X})$, all the non-zero gradient on the different directions are negative gradient, which implies $\langle \mathbf{W}_{4,j',r,p}^{(t)}, e_s \rangle \leq \langle \mathbf{W}_{4,j',r,p}^{(0)}, e_s \rangle = \tilde{O}(\sigma_0)$. Moreover, $\Lambda_{4,j',r}^{(t)} \leq \tilde{O}(\sigma_0)$ is also non-increasing.

For $s = \tau(g)$, $g \in \mathcal{G}$, whenever $\langle \mathbf{W}_{4,j',r,2}^{(t)}, e_s \rangle$ reaches -3μ , we have $\Lambda_{4,j',r}^{(t)}|_{g_1=g} \leq -3\mu + \frac{5}{2}\mu + \tilde{O}(\sigma_0) \leq 0$, and thus $\langle -\nabla_{\mathbf{W}_{4,j',r,2}^{(t)}} \text{Loss}^1, e_s \rangle = \frac{1}{2} \mathbb{E} \left[-\text{logit}_{4,j}^{(t)} \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbf{1}_{g_1=g} \right] = 0$, which implies $\langle \mathbf{W}_{4,j',r,2}^{(t)}, e_s \rangle \geq -3\mu$. Hence, $|\langle \mathbf{W}_{4,j',r,2}^{(t)}, e_s \rangle| \leq \tilde{O}(\sigma_0)$. Following the similar argument, we can prove the result for $\langle \mathbf{W}_{4,j',r,5}^{(t)}, e_s \rangle$ for $s \in \tau(\mathcal{Y})$. \square

A.8 Convergence

Lemma A.11. For $|\mathcal{G}| \geq |\mathcal{Y}| \geq \Omega(\frac{\log \log d}{\log \log \log d})$, $\text{poly} \log d \geq m \geq |\mathcal{Y}|$, $\varrho \ll \frac{1}{m \log d}$ and sufficiently small $\eta \leq \frac{1}{\text{poly} d}$, Induction A.1 holds for all iterations $t \leq T = \frac{\text{poly} d}{\eta}$.

Proof. Putting the results in Lemmas A.5 to A.8 and A.10, we can directly establish the results in Induction A.1. \square

Lemma A.12 (Convergence). For sufficiently large $T_1 \leq t = \frac{\text{poly} d}{\eta}$, we have

- (a) Objective convergence: $\text{Loss}^1 \leq \frac{1}{\text{poly} d}$;
- (b) Successful learning of diagonal feature: $\Phi_{4,j}^{(t)} \geq \Omega(\log d)$ for any $j \in \tau(\mathcal{X})$.

Proof. Assuming for some sufficiently large constant $n > 0$, $\mathbb{E}[(1 - \mathbf{logit}_{4,j^*}^{(t)} \mid \tau(x_1) = j^*)] \geq \Omega(\frac{1}{d^n})$ for $t \in (T_1, T_1 + \frac{d^{n+1} \log^2 d}{\eta}]$ then by Lemma A.1, we have

$$\Gamma_{4,j^{(*)}}^{(T_1 + \frac{d^2 \log^2 d}{\eta})} \geq \Omega\left(\frac{\eta}{d^{n+1}}\right) \cdot \frac{d^{n+1} \log^2 d}{\eta} + \Gamma_{4,j^{(*)}}^{(t)} \geq \Omega(\log^2 d),$$

which contradicts with $\Gamma_{4,j^{(*)}}^{(t)} \leq \Phi_{4,j^{(*)}}^{(t)} \leq O(\log^{1.5} d) = \tilde{O}(1)$ in the polynomial time. This implies after sufficiently large iteration t , we must have $\mathbb{E}[(1 - \mathbf{logit}_{4,j}^{(t)} \mid \tau(x_1) = j)] \leq O(\frac{1}{d^n})$ for $j \in \tau(\mathcal{X})$. Hence

$$\begin{aligned} \text{Loss}^1 &= \mathbb{E}[-\log \mathbf{logit}_{4,\tau(x_1)}^{(t)}] = \sum_{j \in \tau(\mathcal{X})} \mathbb{E}[-\log \mathbf{logit}_{4,j}^{(t)} \mathbb{1}_{\tau(x_1)=j}] \\ &\leq \sum_{j \in \tau(\mathcal{X})} \mathbb{E}[O(1)(1 - \mathbf{logit}_{4,j}^{(t)}) \mathbb{1}_{\tau(x_1)=j}] \quad (\mathbf{logit}_{4,j}^{(t)} \text{ is very close to } 1) \\ &\leq O\left(\frac{1}{\text{poly}d}\right). \end{aligned}$$

By Claim A.2, at the time of convergence, we must have $\Phi_{4,j}^{(t)} \geq \Omega(\log d)$. \square

B Learning Simply Transitive Actions

B.1 Preliminaries

Definition B.1 (Combinations ϕ, Φ, Φ^\dagger). Let \mathcal{G}, \mathcal{Y} be defined as in Definition 2.2. For any pair $(g, y) \in \mathcal{G} \times \mathcal{Y}$, we call $\phi = (g, y)$ a **combination**. We write $\phi_1 = g$ and $\phi_2 = y$ to denote the corresponding components. We further define the set of **correct combinations** Φ^* by

$$\Phi_j^* = \{\phi = (g, y) \mid j = \tau(\alpha(g, y))\}, \quad \forall j \in \tau(\mathcal{Y}); \quad \Phi = \bigcup_{j \in \tau(\mathcal{Y})} \Phi_j^* \quad (24)$$

The combinations $\phi \in \Phi_j^*$ are the ones that can be composed to get $\tilde{y} = \tau^{-1}(j) \in \mathcal{Y}$. We also define the set of **incorrect combinations** Φ_j^\dagger for each $j \in \tau(\mathcal{Y})$ to be

$$\Phi_j^\dagger = \{\phi \in \Phi \setminus \Phi_j^*\} \quad (25)$$

For simply transitive action $\alpha : \mathcal{G} \times \mathcal{Y} \rightarrow \mathcal{Y}$, any combination $\phi \in \Phi_j^\dagger$ is a *incorrect* pair of g and y . That is, it holds that any $y' \neq y \in \mathcal{Y}$ and $g' \neq g \in \mathcal{G}$ satisfy $\alpha(g, y') = \alpha(y, g') \neq \tau^{-1}(j)$.

For each combination, we define the following neural-features:

Definition B.2 (combination features $\psi, \Psi, \Psi^*, \Psi^\dagger$). Given a combination $\phi = (g, y) \in \Phi$, token index $j \in \tau(\mathcal{Y})$ and neuron index $r \in [m]$, we define

$$V_{j,r}(g) = \langle \mathbf{W}_{5,j,r,2}, e_g \rangle, \quad V_{j,r}(y) = \langle \mathbf{W}_{5,j,r,5}, e_y \rangle, \quad V_{j,r}(\phi) := \frac{1}{2}(V_{j,r}(g) + V_{j,r}(y))$$

and we call $V_{j,r}(\phi)$ the *features for combination* ϕ in neuron $(j, r) \in \tau(\mathcal{Y}) \times [m]$. We write $\psi = (j, r, \phi)$ and $V_\psi \equiv V_{j,r}(\phi)$ to make the notation concise. Similar to above, we further define

$$\Psi := \{\psi = (j, r, \phi) \times \tau(\mathcal{Y}) \times [m] \times \Phi\}$$

and Ψ^* that contains the desirable features for each class $j \in \tau(\mathcal{Y})$:

$$\Psi_{j,r}^* = \{\psi = (j, r, \phi) \mid \phi \in \Phi_j^*\}, \quad \Psi^* = \bigcup_{(j,r) \in \tau(\mathcal{Y}) \times [m]} \Psi_{j,r}^* \quad (26)$$

and Ψ^\dagger that contains the incorrect feature combinations:

$$\Psi_{j,r}^\dagger = \{\psi = (j, r, \phi) \mid \phi \in \Phi \setminus \Phi_j^*\}, \quad \Psi^\dagger = \bigcup_{(j,r) \in \tau(\mathcal{Y}) \times [m]} \widehat{\Psi}_{j,r} \quad (27)$$

Events of combination appearance. Let us first define some useful notations. For a combination $\phi = (g, y)$, we write \mathcal{H}_ϕ to denote the event when ϕ appears in the sequence \mathbf{Z}^1 :

$$\mathcal{H}_\phi := \{(g_1, y_0) = \phi\}, \quad \mathcal{H}_{\phi_1} = \mathcal{H}_g = \{g_1 = g\}, \quad \mathcal{H}_{\phi_2} = \mathcal{H}_y = \{y_0 = y\}.$$

Further, we write

$$\mathcal{H}_{\phi,1}^\dagger = \{g_1 \neq g, y_0 = y\}, \quad \mathcal{H}_{\phi,2}^\dagger = \{g_1 = g, y_0 \neq y\}, \quad \mathcal{H}_\phi^\dagger = \mathcal{H}_{\phi,1}^\dagger \cup \mathcal{H}_{\phi,2}^\dagger$$

to denote the event where ϕ did not appear but its group element or value is the same.

Finally we define a notion called *learning curriculum*, that sits at the center of our proof.

Definition B.3 (Learning curriculum). We define an order on the set Σ , defined by the following process: Let $\Sigma_0 = \Psi$. At each $i \in [n_y^2]$, we choose

$$\psi_i = \arg \max_{\psi \in \Sigma_i} V_\psi^{(0)}$$

Let's write $\psi_i = (j, r, \phi)$ where $\phi = (g, y)$, then we define Σ_{i+1} by excluding the following features:

1. Exclude the confusing combinations in neuron (j, r) ;

$$\Sigma_{\psi_i}^{\dagger,1} \equiv \Sigma_i^{\dagger,1} = \{\psi' = (j, r, \phi') \in \Sigma_i, \phi' = (g', y') \mid (g' = g) \text{ XOR } (y' = y)\}$$

2. Exclude the unselected combinations in neuron (j, r) ;

$$\Sigma_{\psi_i}^{\dagger,2} \equiv \Sigma_i^{\dagger,2} = \{\psi' = (j, r, \phi') \in \Sigma_i, \phi' = (g', y') \mid g' \neq g, y' \neq y\}$$

3. Exclude the feature indices of the same combination in other neurons $(j, r'), r' \neq r$;

$$\Sigma_{\psi_i}^{\dagger,3} \equiv \Sigma_i^{\dagger,3} = \{\psi' = (j, r', \phi) \in \Sigma_i, \forall r' \neq r \in [m]\}.$$

Which returns $\Sigma_{i+1} = \Sigma_i \setminus \{\psi = (j, r', \phi') \in \Sigma_i \mid \text{either } r' = r \text{ or } \phi' = \phi\}$. Iterate over the whole set Ψ , we will arrive at $\Sigma_{n_y^2} = \emptyset$. This gives rise to an ordered sequence Σ^* and a set Σ^\dagger :

$$\Sigma^* := (\psi_1, \psi_2, \dots, \psi_{n_y^2}); \quad \Sigma^\dagger := \bigcup_{\psi \in \Sigma^*} \Sigma_\psi^\dagger, \quad \Sigma_\psi^\dagger := \Sigma_\psi^{\dagger,1} \cup \Sigma_\psi^{\dagger,2} \cup \Sigma_\psi^{\dagger,3}$$

By our construction $V_{\psi_i}^{(0)} \geq V_{\psi_{i+1}}^{(0)}$. We write $\psi \prec_\Sigma \psi'$ to denote that ψ is ahead of ψ' in Σ^* .

Intuitively, Σ^* encodes the order at which the features V_ψ grow in magnitude and Σ^\dagger contains all the unlearned features during the process.

Throughout the analysis of the FFN layer, we further make the following assumptions.

Assumption B.1. For $(j, r, p) \in [d] \times [m] \times [5]$, we fix $\mathbf{W}_{5,j,r,p}^{(t)} \equiv \mathbf{W}_{5,j,r,p}^{(0)}$ for $p \in \{1, 3, 4\}$ at initialization for simplicity of proof.

Assumption B.2. We use a modified smoothed ReLU as our activation function. This technical assumption is to avoid many pathologies appearing in the learning dynamics.

$$\mathbf{sReLU}(x) := \begin{cases} \varpi B & x \leq -B \\ -\varpi x & x \in (-B, -\varpi] \\ \frac{1}{2}x^2 & x \in (-\varpi, 0]; \\ x^q / (\varrho^{q-1}q) & x \in (0, \varrho]; \\ x - \varrho \left(1 - \frac{1}{q}\right) & x \in (\varrho, B] \\ B - \varrho \left(1 - \frac{1}{q}\right) & x > B \end{cases}$$

where $q = O(1)$ is a large even integer and $\varrho = \Theta(1/\text{polylog}(d))$, $\varpi \in (d^2\mu^{q-1}, \frac{\lambda}{d^{q/3}})$, $B = C \log d$, $C = \Theta(1) \in (5, \frac{q-1}{3})$ and $\lambda = \frac{d-1}{d-1+e^B}$.

B.1.1 Theorem Statement

We will try to prove the following theorem:

Theorem B.1 (Learning Simply Transitive Actions). *Suppose $F^{(t)}$ is returned by Algorithm 1 at $t = T_1$, and let $\delta_1 = d^{c_1}\mu$, $\delta_2 = \varpi^{\frac{1}{q-1}}$, it holds that with probability $\geq 1 - o(1)$:*

- A. For all $\psi \in \Sigma^*$, we have $V_\psi^{(T_1)} = B \pm O(\delta_1)$;
- B. For any $\psi \in \Sigma^{\dagger,1}$, $|V_\psi^{(T_1)}| \leq \tilde{O}(\delta_2)$;
- C. For any $\psi \in \Sigma^{\dagger,2}$, $V_\psi^{(T_1)} \leq -B + O(\delta_1)$;
- D. For any $\psi \in \Sigma^{\dagger,3}$, $V_\psi^{(T_1)} \leq O(\delta_1)$;
- E. For $\psi = (j, r, (g, y)) \in \Sigma^*$, $|V_{j,r}^{(t)}(g) - V_{j,r}^{(t)}(y)| \leq \tilde{O}(\delta_2)$.

In fact, we can decompose the statement of theorem B.1 into the following statement for every feature $\psi \in \Sigma^*$.

Definition B.4 (Feature Shape). Let $\delta = (\delta_1, \delta_2)$ be a tuple of error parameters. Let $\psi \in \Sigma^*$ be a feature in the learning curriculum. We say the feature ψ reached *feature shape* $\mathcal{F}_\psi(\delta)$ with error δ if:

1. $\mathcal{F}_{\psi,1}(\delta_1)$: $V_\psi^{(t)} \geq B - O(\delta_1)$;
2. $\mathcal{F}_{\psi,2}(\delta_2)$: For any $\psi' \in \Sigma_\psi^{\dagger,1}$, it holds that $|V_{\psi'}^{(t)}| \leq \tilde{O}(\delta_2)$;
3. $\mathcal{F}_{\psi,3}(\delta_1)$: For any $\psi' \in \Sigma_\psi^{\dagger,2}$, it holds that $V_{\psi'}^{(t)} \leq -B \pm O(\delta_1)$;
4. $\mathcal{F}_{\psi,4}(\delta_1)$: For any feature $\psi' \in \Sigma_\psi^{\dagger,3}$, it holds that $V_{\psi'}^{(t)} \leq O(\delta_1)$.
5. $\mathcal{F}_{\psi,5}(\delta_2)$: Writing $\psi = (j, r, (g, y))$, then it holds that $|V_{j,r}^{(t)}(g) - V_{j,r}^{(t)}(y)| \leq \tilde{O}(\delta_2)$;

Obviously if we proved the condition $\mathcal{F}_\psi(\delta)$ of all $\psi \in \Sigma^*$ are reached for $\delta = (d^{c_1}\mu, \varpi^{\frac{1}{q-1}})$ at T_1 , then theorem B.1 is proven. We do this by following a induction process sequentially for each $\psi \in \Sigma^*$, and provide a guarantee that all features in Σ^* reach the claimed convergence condition together at $t = T_1$.

B.1.2 Induction Hypotheses and Phase Decomposition

To prove theorem B.1, or equivalently, to prove the convergence condition in Definition B.4 is reached for every feature $\psi \in \Sigma^*$, we shall characterize the dynamics of each feature and prove that they eventually arrive at the desired shape.

Phase Decomposition. Let us define the following timestamps for different phases of learning each $\psi = (j, r, \phi) \in \Psi \cup \widehat{\Psi}$. Let $c_1 = \frac{1}{1000}$ be a small constant:

- (a) Phase I: $t \in [0, T_\psi^1]$, where $T_\psi^1 := \min\{t \geq 0 \mid V_\psi^{(t)} \geq d^{c_1} \mu\}$;
- (b) Phase II.1: $t \in (T_\psi^1, T_\psi^{2,1}]$, where $T_\psi^{2,1} := \min\{t \geq 0 \mid V_\psi^{(t)} \geq \frac{1}{2} \log d\}$;
- (c) Phase II.2: $t \in (T_\psi^{2,1}, T_\psi^{2,2}]$, where $T_\psi^{2,2} := \min\{t \geq 0 \mid \mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mathbf{1}_{\mathcal{H}_\phi}] \geq 1 - \frac{1}{\sqrt{d}}\}$;
- (d) Phase II.3: $t \in (T_\psi^{2,2}, T_\psi^2]$, where $T_\psi^2 = \max\{T_\psi^{2,3}, T_\psi^{2,2} + \Omega(\frac{1}{\eta(d^{c_1} \mu)^{q-2}})\}$ and $T_\psi^{2,3}$ is defined as:

$$T_\psi^{2,3} := \min \left\{ t \geq T_\psi^{2,2} \mid \mathcal{F}_\psi(\delta_1, \delta_2) \text{ holds, where } \delta_1 = d^{c_1} \sigma_0, \delta_2 = \left(\frac{n_y d \varpi}{\lambda} \right)^{\frac{1}{q-1}} \right\}$$

- (e) Phase III.1: $t \in (T_\psi^2, T_{1,1}]$, where $T_{1,1} = T_{\psi_{n_y^2}}^2$ and $\psi_{n_y^2}$ is the last feature in Σ^* . This is the convergence phase where the rest of the features are learned and the feature converge to a perfect shape.
- (f) Phase III.2: $t \in (T_{1,1}, T_1]$, the end phase where all features are in perfect shape and have stablized.

Now we state the following induction hypotheses which naturally results in theorem B.1.

Induction B.1 (Simply Transitive Actions, All Phases). *For all $t \leq T_1$, the following holds:*

- (a) $V_\psi^{(t)} + b_{i,j,r} \geq \Omega(\mu)$.
- (b) $T_{\psi'}^2 < T_\psi^1$ if $\psi' \prec_\Sigma \psi$ for $\psi, \psi' \in \Sigma^*$, i.e., the intervals $\{[T_\psi^1, T_\psi^2]\}_{\psi \in \Sigma^*}$ are non-overlapping;

In order to prove Induction B.1 and theorem B.1, we introduce the following feature based induction hypothesis to prove that \mathcal{F}_ψ defined in Definition B.4 holds for some error parameters at $t = T_\psi^2$.

Induction B.2 (Induction for Individual Feature). *Let $\psi = (j, r, \phi) \in \Sigma^*$, $\phi = (g, y)$, at $t \leq T_1$, the following holds:*

- (a) At $t \leq T_\psi^1$, any $g \neq g', y' \neq y$ has $V_{j,r}(g'), V_{j,r}(y') \leq \widetilde{O}(\mu)$.
- (b) During $t \leq T_\psi^1$, there are at most $\widetilde{O}(\sqrt{d}/\eta)$ iterations where $\mathbf{logit}^{(t)} \geq \frac{1}{\sqrt{d}}$ conditioned on \mathcal{H}_ϕ .
- (c) Let $\psi' \prec \psi \in \Sigma^*$, the feature shape $\mathcal{F}_{\psi'}(d^{c_1} \sigma_0, (\frac{n_y d \varpi}{\lambda})^{\frac{1}{q-1}})$ holds throughout $t \in [T_\psi^1, T_1]$;
- (d) Let $\psi' \succ \psi \in \Sigma^*$, then $V_\psi^{(t)} \leq \widetilde{O}(\sigma_0)$ throughout $t \in [0, T_\psi^2]$.

B.1.3 Technical Calculations

Let us recall some facts about gradient computation for \mathbf{W} .

Fact B.1 (gradient computation). The gradient with respect to $\mathbf{W}_{i,j,r,p}$ for $i = 5, j \in [d], r \in [m], p \in [5]$ and $v \in \mathcal{V}$ is

$$\langle -\nabla_{\mathbf{W}_{5,j,r,p}} \text{Loss}, e_v \rangle = \frac{1}{2} \mathbb{E} \left[\mathcal{E}_{5,j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \sum_{\mathbf{k} \in \mathcal{I}^{1,0}} \langle \mathbf{Z}_{\mathbf{k},p}, e_v \rangle \right] \quad (28)$$

where $\mathcal{E}_{5,j} = \mathbf{1}_{v=Z_{\text{ans},1,p}} - \mathbf{logit}_{5,j}$ is the loss derivatives. We compute more precise expressions when j, p and v varies:

- (a) when $p = 2$, for $g \in \mathcal{G}$, let $\phi = (g, y) \in \Phi_j^*$ be the combination with $\phi_1 = g$, we have for all $j \in \tau(\mathcal{Y}), r \in [m]$,

$$\begin{aligned} \langle -\nabla_{\mathbf{w}_{5,j,r,2}} \text{Loss}, e_g \rangle &= \frac{1}{2} \mathbb{E} \left[(1 - \text{logit}_{5,j}) \text{sReLU}'(\Lambda_{5,j,r}) \mathbf{1}_{\mathcal{H}_\phi} \right] \\ &\quad - \frac{1}{2} \mathbb{E} \left[\text{logit}_{5,j} \text{sReLU}'(\Lambda_{5,j,r}) \mathbf{1}_{\mathcal{H}_{\phi,2}^\dagger} \right] \end{aligned} \quad (29)$$

- (b) when $p = 5$, for $y \in \mathcal{Y}$, let $\phi = (g, y) \in \Phi_j^*$ be the combination with $\phi_2 = y$, we have for all $j \in \tau(\mathcal{Y}), r \in [m]$,

$$\begin{aligned} \langle -\nabla_{\mathbf{w}_{5,j,r,5}} \text{Loss}, e_y \rangle &= \frac{1}{2} \mathbb{E} \left[(1 - \text{logit}_{5,j}) \text{sReLU}'(\Lambda_{5,j,r}) \mathbf{1}_{\mathcal{H}_\phi} \right] \\ &\quad - \frac{1}{2} \mathbb{E} \left[\text{logit}_{5,j} \text{sReLU}'(\Lambda_{5,j,r}) \mathbf{1}_{\mathcal{H}_{\phi,1}^\dagger} \right] \end{aligned} \quad (30)$$

- (c) For any combination of p, v not included above, we have due to our assumptions of the distribution and update rule:

$$\langle -\nabla_{\mathbf{w}_{5,j,r,p}} \text{Loss}, e_v \rangle \equiv 0$$

Definition B.5 (Γ -notations). To simplify gradient expression, we define the following notation: let $\psi = (j, r, \phi) \in \Psi$, we write

$$\begin{aligned} \Gamma_{j,r,\phi}^{+, (t)} &= \Gamma_{\psi}^{+, (t)} := \mathbb{E} \left[(1 - \text{logit}_{5,j}) \text{sReLU}'(\Lambda_{5,j,r}) \mathbf{1}_{\mathcal{H}_\phi} \right] \\ \Gamma_{j,r,\phi}^{-, (t)} &= \Gamma_{\psi}^{-, (t)} := \mathbb{E} \left[\text{logit}_{5,j} \text{sReLU}'(\Lambda_{5,j,r}) \mathbf{1}_{\mathcal{H}_\phi^\dagger} \right] \end{aligned} \quad (31)$$

we can further define Γ for each $g \in \mathcal{G}$ and $y \in \mathcal{Y}$: Let $v \in \mathcal{G} \cup \mathcal{Y}$, we define

$$\begin{aligned} \Gamma_{j,r,v}^{+, (t)} &= \Gamma_{\psi,v}^{+, (t)} := \mathbb{E} \left[(1 - \text{logit}_{5,j}) \text{sReLU}'(\Lambda_{5,j,r}) \mathbf{1}_{\mathcal{H}_v} \right] \\ \Gamma_{j,r,v}^{-, (t)} &= \Gamma_{\psi,v}^{-, (t)} := \mathbb{E} \left[\text{logit}_{5,j} \text{sReLU}'(\Lambda_{5,j,r}) \mathbf{1}_{\mathcal{H}_v^\dagger} \right] \end{aligned} \quad (32)$$

This allows us to define the following gradient condition:

Definition B.6 (gradient condition). At $t \leq T_1$, letting $\psi = (j, r, \phi) \in \Psi$, we write $\mathcal{C}_\psi(\delta)$ for $\delta > 0$ to denote that the gradient update for V_ψ is smaller than δ , formally:

$$\begin{aligned} \mathcal{C}_{j,r,v}(\delta) \text{ holds} &\implies |\Gamma_{j,r,v}^{+, (t)} - \Gamma_{j,r,v}^{-, (t)}| \leq \delta \\ \mathcal{C}_{j,r,v}^+(\delta) \text{ holds} &\implies |\Gamma_{j,r,v}^{+, (t)}| \leq \delta \\ \mathcal{C}_{j,r,v}^-(\delta) \text{ holds} &\implies |\Gamma_{j,r,v}^{-, (t)}| \leq \delta \end{aligned}$$

We further write

$$\begin{aligned} \mathcal{C}_\psi(\delta) \text{ holds} &\implies \mathcal{C}_{j,r,g}(\delta) \wedge \mathcal{C}_{j,r,y}(\delta) \text{ holds} \\ \mathcal{C}_\psi^+(\delta) \text{ holds} &\implies \mathcal{C}_{j,r,g}^+(\delta) \wedge \mathcal{C}_{j,r,y}^+(\delta) \text{ holds} \\ \mathcal{C}_\psi^-(\delta) \text{ holds} &\implies \mathcal{C}_{j,r,g}^-(\delta) \wedge \mathcal{C}_{j,r,y}^-(\delta) \text{ holds} \end{aligned}$$

Now we proceed to characterize the lower bound of logits

Fact B.2 (logit lower bound). We list some basic facts about the architecture Definition 3.5 here.

1. For any $Z^1 \in \text{supp}(\mathcal{D}^1)$, we have

$$\lambda := \min_{i,j,F,Z} (1 - \text{logit}_{i,j}(F, Z)) = \frac{d-1}{d-1+e^B}$$

2. Suppose for some $i \in [5]$, $F_{i,j} = o(1)$ for all $j \notin \tau(\mathcal{Y})$, then

$$\min_{i,j,F,Z} \text{logit}_{i,j}(F, Z) = \frac{1}{(1+o(1))d + n_y e^B} = O\left(\frac{\lambda}{n_y d}\right)$$

B.2 Phase I: Feature Emergence and Competition

We shall prove the following properties at initialization.

Fact B.3 (initialization range). At $t = 0$, the following holds with probability $\geq 1 - o(1)$:

- (a) $|\langle \mathbf{W}_{5,j,r,p}^{(0)}, e_v \rangle| \leq O(\sigma_0 \sqrt{\log d})$ For all j, r, p and $v \in \mathcal{V}$;
- (b) $V_\psi^{(0)} \geq V_{\psi'}^{(0)} + \gamma$, where $\gamma = \Omega(\frac{\sigma_0}{n_y^4 m^2 \log d})$, for any pair $\psi \prec \psi' \in \Sigma^*$ in Def. B.3;

Proof. Fact B.3a can easily verified from our initialization $\mathbf{W}_{i,j,r,p}^{(0)} \sim \mathcal{N}(0, \sigma_0 I_d)$. For Fact B.3b, we give a straightforward proof that every pair of $V_\psi^{(0)}, \psi \in \Sigma^*$ has a gap of $\frac{\sigma_0}{n_y^4 m^2 \log d}$. First note that $V_\psi^{(0)}$ of different $\psi \in \Sigma^*$ are independent and identically distributed on the randomness of $\mathbf{W}^{(0)}$, due to the orthogonality of embeddings $e_v, v \in \mathcal{V}$. Then, by the basic property of a Gaussian variable (notice that $V_\psi^{(0)} - V_{\psi'}^{(0)}$ is also Gaussian with variance $2\sigma_0$) (all though different pairs could be dependent), we have with probability $1 - \frac{1}{n_y^4 m^2 \log d}$ that their gap is at least $\gtrsim \frac{\sigma_0}{n_y^4 m^2 \log d}$ for each pair. Then by a union bound over $O(m^2 n_y^4)$ -many all possible pairs we can conclude the proof. \square

We give another characterization of the initial activations.

Fact B.4 (activation magnitude). At $t = 0$, with high probability it holds that

$$\Lambda_{5,j,r}^{(0)} = (1 + O(\frac{1}{\sqrt{\log d}}))\mu = \Theta(\sigma \log d) \ll \varrho$$

and thus $\text{sReLU}(\Lambda_{5,j,r}) = \Theta(\frac{1}{q}\mu^q)$, $\text{sReLU}'(\Lambda_{5,j,r}) = \Theta(\mu^{q-1})$ and $F_{5,j} = \Theta(\frac{m}{q}\mu^q)$.

Proof. Combining Definitions 3.4 and 3.5, Assumption 3.3, and Fact B.3 gives the fact. \square

We establish some properties in Phase I.

Lemma B.1 (Key properties in Phase I). *Let $\psi = (j, r, (g, y)) \in \Sigma^*$, and assume Inductions B.1 and B.2 holds at $t \leq T_\psi^1$, then*

- (a) $|V_{j,r}^{(t)}(v)| \leq O(\mu/\sqrt{\log d})$ if $v \neq g, v \neq y$;
- (b) $\Lambda_{5,j,r}^{(t)}(\mathbf{Z}) \geq 0$ for $\mathbf{Z} \in \mathcal{H}_\phi$;
- (c) $\text{logit}_{5,j} = O(1/d)$ whenever \mathcal{H}_ϕ^\dagger happens.

Proof. lemma B.1a is from both the initialization of weights Fact B.3 and Induction B.2a. lemma B.1b is from lemma B.1a and Induction B.1a. lemma B.1c is due to both Induction B.1b and Induction B.1b. \square

B.2.1 Competition between feature combinations.

In this phase we compare the growth of different features using a proxy.

Lemma B.2 (Approximating gradient with proxy). *Assume Induction B.1. For a fixed $\psi = (j, r, \phi) \in \Sigma^*$, define the auxiliary sequence*

$$\tilde{V}_\psi^{(0)} := V_\psi^{(0)}, \quad \tilde{V}_\psi^{(t+1)} := \tilde{V}_\psi^{(t)} + \eta \mathbb{E}[\mathbf{sReLU}'(\tilde{\Lambda}_{5,j,r}^{(t)}) \mathbf{1}_{\mathcal{H}_\phi}],$$

where in $\tilde{\Lambda}_{5,j,r}^{(t)}$ the weight V_ψ is replaced by \tilde{V}_ψ (all other weights are unchanged). Then for any feature $\psi \succ \psi_i$ or $\psi \in \Sigma_i \setminus \{\psi_i\}$ and all $t \leq T_\psi^1$, $|V_\psi^{(t)} - \tilde{V}_\psi^{(t)}| \leq O(\tilde{V}_\psi^{(t)}/d^{1/2-3c_1})$.

Proof. We need to prove the result by induction. Firstly we notice that the newly defined proxy sequence $\tilde{V}_\psi^{(t)}$ is monotonically increasing. Moreover, as $\tilde{V}_\psi^{(t)}$ is initialized the same as $V_\psi^{(t)}$, we can use lemmas B.11 and B.12 to argue its growth by taking

$$x_t = \tilde{V}_\psi^{(t)} + b_{i,j,r}, \quad \xi = \sum_{p=3,4} \langle \mathbf{W}_{5,j,r,p}, e_{x_0} \rangle + \langle \mathbf{W}_{5,j,r,1}, e_{x_1} \rangle, \quad x_{t+1} \geq x_t + \eta \frac{1}{n_y^2} \mathbb{E}[(x_t + \mu)^{q-1}]$$

And lemmas B.12 and B.13 guarantee $\tilde{V}_\psi^{(t)} \geq d^{c_1} \mu$ after $\tilde{O}(\frac{1}{\eta \mu^{q-2}})$ steps. Now assume the bound of difference between $V_\psi^{(t)}$ and $\tilde{V}_\psi^{(t)}$ holds at a $t \leq T_\psi^1$. Since $t \leq T_\psi^1$, we have $V_\psi^{(t)} \leq d^{c_1} \sigma_0$ by definition of T_ψ^1 . By Fact B.1, the exact update of V_ψ is

$$V_\psi^{(t+1)} = V_\psi^{(t)} + \eta \left(\mathbb{E}[(1 - \mathbf{logit}_{5,j}^{(t)}) \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbf{1}_{\mathcal{H}_\phi}] - \frac{1}{2} \mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbf{1}_{\mathcal{H}_\phi^\dagger}] \right).$$

Compare the above with the definition of $\tilde{V}_\psi^{(t)}$, we first notice that $\tilde{V}_\psi^{(t)}$ is dominating $V_\psi^{(t)}$ for all iterations $t \leq T_\psi^1$. Then, we can bound the difference between the two sequences by

$$\begin{aligned} & |\tilde{V}_\psi^{(t+1)} - V_\psi^{(t+1)}| \\ & \leq |\tilde{V}_\psi^{(t)} - V_\psi^{(t)}| + \eta \left| \mathbb{E}[\mathbf{sReLU}'(\tilde{\Lambda}_{5,j,r}^{(t)}) - \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbf{1}_{\mathcal{H}_\phi}] \right| \\ & \quad + \eta \left| \mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbf{1}_{\mathcal{H}_\phi}] \right| + \frac{\eta}{2} \left| \mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbf{1}_{\mathcal{H}_\phi^\dagger}] \right| \end{aligned}$$

We bound the three error terms for $t \leq T_\psi^1$:

- A. Activation perturbation: by the smoothness of \mathbf{sReLU}' on $[0, \varrho]$ and that both pre-activations are in this range during Phase I, we have

$$\left| \mathbb{E}[(\mathbf{sReLU}'(\tilde{\Lambda}_{5,j,r}^{(t)}) - \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})) \mathbf{1}_{\mathcal{H}_\phi}] \right| \leq |\tilde{V}_\psi^{(t)} - V_\psi^{(t)}| \mathbb{E}[(\tilde{\Lambda}_{5,j,r}^{(t)})^{q-2} \mathbf{1}_{\mathcal{H}_\phi}]$$

where we have used the fact that $\tilde{\Lambda}_{5,j,r} \geq \Lambda_{5,j,r}$ because of the dominance of \tilde{V}_ψ .

- B. Outside a set of iterations $\mathcal{A}_\phi = \{t \leq T_\psi^1 \mid \mathbf{logit}_{5,j}^{(t)} \geq \frac{1}{\sqrt{d}} \text{ conditioned on } \mathcal{H}_\phi\}$ whose cardinality is no more than $\tilde{O}(d^{\frac{1}{2}+c_1}/\eta)$ by Induction B.2a, we have $\mathbf{logit}_{5,j}^{(t)} \leq d^{-1/2}$. Hence for $t \notin \mathcal{A}_\phi$

$$\left| \mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbf{1}_{\mathcal{H}_\phi}] \right| \leq O(d^{-1/2}) \mathbb{E}[\mathbf{sReLU}'(\tilde{\Lambda}_{5,j,r}^{(t)}) \mathbf{1}_{\mathcal{H}_\phi}]$$

Now we can sum over the iterations to get

$$\left(\sum_{s \leq t, s \in \mathcal{A}_\phi} + \sum_{s \leq t, s \notin \mathcal{A}_\phi} \right) \eta \left| \mathbb{E}[\mathbf{logit}_{5,j}^{(s)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbf{1}_{\mathcal{H}_\phi}] \right|$$

$$\begin{aligned}
&\leq O\left(\frac{d^{\frac{1}{2}+c_1}}{\eta}\right) \cdot \eta(d^{c_1}\mu)^{q-1} + O(d^{-1/2})(\tilde{V}_\psi^{(t)} - \tilde{V}_\psi^{(0)}) \\
&\leq O(\mu/d^{1/2-c_1})
\end{aligned}$$

C. Confusing events: before $V_\psi^{(t)} \geq \frac{1}{2} \log d$, we have $\mathbf{logit}_{5,j} = O(1/d)$ in Phase I by lemma B.1. So we get

$$\sum_{s \leq t} \eta \mathbb{E}[\mathbf{logit}_{5,j}^{(t)} | \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) | \mathbb{1}_{\mathcal{H}_\phi^\dagger}] \leq O\left(\frac{1}{\eta \mu^{q-2}}\right) \eta \left(\frac{(d^{c_1}\mu)^{q-1}}{d}\right) \ll O(\mu/\sqrt{d})$$

So we have that the accumulated errors from the last two terms are smaller than $O(\mu/d^{1/2-c_1})$. This allow us to bound the difference by using the more naive approach: we can assume the difference started with $O(\mu/d^{1/2-c_1})$ and now we only need to bound the following sequence for $t \leq T_\psi^1$

$$\delta_{t+1} = \delta_t + O(\eta \delta_t) \mathbb{E} \left[\frac{1}{\tilde{\Lambda}_{5,j,r}^{(t)}} \mathbf{sReLU}'(\tilde{\Lambda}_{5,j,r}^{(t)}) \mathbb{1}_{\mathcal{H}_\phi} \right], \quad \delta_0 \leq O(\mu/d^{1/2-c_1})$$

Now let $\delta' > \eta$ be some small parameter, we can do some slicing of time t_0, t_1, \dots, t_n where $t_i = \min\{t \geq 0 \mid \tilde{V}_\psi^{(t)} \geq (1 + \delta')^i \mu\}$ and $(1 + \delta')^n \mu \leq \tilde{V}_\psi^{(t_{n+1})} \leq (1 + \delta')^{n+1} \mu$. This produce the following bound:

$$\begin{aligned}
&\log \delta_{t+1} - \log \delta_0 \\
&= \sum_{1 \leq i \leq n} \sum_{s \in [t_i, t_{i+1}-1]} \log \left(1 + O(\eta) \mathbb{E} \left[\frac{1}{\tilde{\Lambda}_{5,j,r}^{(s)}} \mathbf{sReLU}'(\tilde{\Lambda}_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{H}_\phi} \right] \right) \\
&\leq \sum_{1 \leq i \leq n} \sum_{s \in [t_i, t_{i+1}-1]} O(\eta) \mathbb{E} \left[\frac{1}{\tilde{\Lambda}_{5,j,r}^{(s)}} \mathbf{sReLU}'(\tilde{\Lambda}_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{H}_\phi} \right] \quad (\log(1+x) \leq x \text{ when } x > 0) \\
&\leq \sum_{1 \leq i \leq n} O\left(\frac{1}{\min_{s \in [t_i, t_{i+1}-1], \mathbf{Z} \in \mathcal{H}_\phi} \tilde{\Lambda}_{5,j,r}^{(s)}(\mathbf{Z})}\right) \sum_{s \in [t_i, t_{i+1}-1]} \eta \left| \mathbb{E}[\mathbf{sReLU}'(\tilde{\Lambda}_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{H}_\phi}] \right| \\
&\leq \sum_{1 \leq i \leq n} O\left(\frac{1}{(1 + \delta')^i \mu}\right) (\tilde{V}_\psi^{(t_{i+1})} - \tilde{V}_\psi^{(t_i)}) \\
&\leq \frac{\delta' \log(\tilde{V}_\psi^{(t)} / \mu)}{\log(1 + \delta')} \\
&\leq 2 \log(\tilde{V}_\psi^{(t)} / \mu) \quad (\log(1 + \delta')^{\frac{1}{\delta'}} < 2 \text{ if } \delta' \text{ is small enough})
\end{aligned}$$

So when $\tilde{V}_\psi^{(t)} \in [d^{c_1}\mu, 2d^{c_1}\mu]$, it holds that $\delta_t \lesssim d^{2c_1} \delta_0 \lesssim \mu/d^{1/2-3c_1}$. This proves the claim. \square

Lemma B.3 (competition). *Assuming Induction B.1, and let $\psi \in \Sigma^*$, we shall have for any feature $\psi' \succ \psi \in \Sigma^*$ or $\psi' \in \Sigma_\psi \setminus \{\psi\}$ that $V_{\psi'}^{(T_\psi^1)} \leq \tilde{O}(\sigma_0)$.*

Proof. We can approximate the sequence V_{ψ_i} using $\tilde{V}_{\psi_i}^{(t)}$ and straightforwardly use lemma B.12 to compare ψ_i with any $\psi' \succ \psi_i \in \Sigma^*$. In fact, let's compared the time they reach $V_\psi^{(t)} \geq d^{c_1}\sigma_0$. By lemma B.2, we can just compare $\tilde{V}_\psi^{(t)}$ before $t = T_\psi^1$, now since by Fact B.3 we know $\tilde{V}_\psi^{(0)} \geq \tilde{V}_{\psi'}^{(0)} + \gamma$, we can use lemma B.12 to get that the first time when $\tilde{V}_{\psi_i}^{(t)} \geq d^{c_1}\sigma_0$, we have $\tilde{V}_{\psi'}^{(t)} = \tilde{O}(\sigma_0)$. This combined with lemma B.2 concludes the result. Moreover, we can also obtain bounds on all $\psi \in \Psi$ by discussing different ψ :

- For all $\phi' \in \Phi^*$ such that there exists $\psi' = (j', r', \phi') \prec \psi_i$, ϕ' is learned. That means any ψ contains ϕ' has gradient too small because of B.2c and therefore $\mathbb{E}[\mathcal{E}_{5,j} \mathbb{1}_{\mathcal{H}_{\phi'}}] \leq \tilde{O}(\lambda |\Gamma_\psi^{(t)}|)$ and therefore wouldn't be learned.

- All the rest of $\phi' \in \Phi^*$ has their corresponding $\psi' \succ \psi \in \Sigma^*$ and therefore lost in the competition and has small growth.
- All other $\psi \in \Psi \setminus \cup_{i' \leq i} \Sigma_{i'}$ lost the competition as well for the same reason.

This concludes the proof. \square

Proof of Inductions B.1 and B.2 in Phase I. Let $\psi_i \in \Sigma^*$. Assume Induction B.1 holds for some $t \leq T_{\psi_{i-1}}^3$, then we shall prove that it continues to hold for $t \in [T_{\psi_{i-1}}^3, T_{\psi_i}^1]$.

- Induction B.1a: This is simple as all V_ψ for $\psi \succ \psi_i$ has updates approximated by $\tilde{V}_\psi^{(t)}$, which is monotonically increasing, and has minimal update by lemma B.2.
- Induction B.1b: We proved in lemma B.3 that whenever $\psi \prec \psi' \in \Sigma^*$, $T_\psi^1 \leq T_{\psi'}^1$. This does not violate Induction B.1b so the non-overlapping is preserved.
- Induction B.2a: This is proven in lemma B.3.
- Induction B.2b-c: They will be proven when we analyze the end phase of feature learning. They are the corollaries of the end phase guarantee of the previous features.
- Induction B.2c:
- Induction B.1e: We did not violate Induction B.1e either.
- Induction B.2a-d will be proved in Phase III where we argue the end phase of each feature learned.

\square

B.3 Phase II: Feature Growth and Cancellations

Below we present the induction hypothesis for phase II.

Induction B.3 (Phase II). *For all $\psi = (j, r, \phi) \in \Sigma^*$ and $t \in [T_\psi^1, T_\psi^2]$, the following holds*

- (a) *For $t \in [T_\psi^1, T_\psi^{2,1}]$, $\mathcal{E}_{5,j}^{(t)} \geq 1 - \frac{1}{\sqrt{d}}$ conditioned on \mathcal{H}_ϕ ;*
- (b) *For $t \leq T_\psi^{2,1}$, for all $j' \neq j$, we have $\mathbf{logit}_{5,j'}^{(t)} \leq O(\sqrt{d}) \cdot \mathbf{logit}_{5,j'}^{(T_\psi^1)}$.*
- (c) *For any $\phi' = (g', y') \neq \phi$, $V_{j,r}^{(t)}(g'), V_{j,r}^{(t)}(y') \in (-B - O(1), \tilde{O}(\mu))$;*
- (d) *For $t \leq T_\psi^{2,2}$, the number of iterations where $V_\psi^{(t+1)} - V_\psi^{(t)} < 0$ is bounded by $\tilde{O}(1/d^{c_1}\eta)$.*

B.3.1 Technical Lemmas

Lemma B.4 (gradient estimation). *Assuming Inductions B.1 and B.3, let $\psi = (j, r, \phi) \in \Sigma^*$, $\phi = (g, y)$ and $t \in [T_\psi^{2,1}, T_\psi^2]$, then*

- (a) *At any $t \geq T_\psi^{2,1}$, if $V_\psi^{(t)} \in (\frac{1}{3} \log d, \frac{2}{3} \log d)$, then $\Gamma_{j,r,\phi}^{(t)} \geq \Omega(1)$.*
- (b) *Any $g' \neq g \in \mathcal{G}, y' \neq y \in \mathcal{Y}$ has the following gradient approximation*

$$\begin{aligned}\Gamma_{j,r,g'}^{(t)} &= -\mathbb{E}\left[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbf{1}_{\mathcal{H}_{(g',y)}}\right] \pm O(\varpi/n_y) \\ \Gamma_{j,r,y'}^{(t)} &= -\mathbb{E}\left[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbf{1}_{\mathcal{H}_{(g,y')}}\right] \pm O(\varpi/n_y)\end{aligned}$$

(c) let $\phi' = (g', y') \neq \phi \in \Phi_j^*$ if $V_{j,r}(g, y) \geq B - o(1)$ and $V_{j,r}(g', y') < -B - 2\mu$, we have

$$\Gamma_{j,r,g'}^{(t)} = \sum_{\tilde{y} \in \mathcal{Y}, V_{j,r}(g', \tilde{y}) \in (-B, -\varpi - \mu)} \Omega\left(\frac{\lambda \varpi}{dn_y^3}\right) - \mathbb{E}\left[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbb{1}_{\mathcal{H}_{(g', y)}}$$

and similarly for $\Gamma_{j,r,y'}^{(t)}$.

Proof. Firstly (a) is simple to prove, because when $V_{\psi}^{(t)} \in (\frac{1}{3} \log d, \frac{2}{3} \log d)$, we have $\mathbf{logit}_{5,j} \mathbb{1}_{\mathcal{H}_{\phi}^*} \leq O(\frac{1}{d^{c_1}})$, which is from the calculation $F_{5,j} \leq \frac{2}{3} \log d + O(1)$ guaranteed by Inductions B.2 and B.3. For (b), we can estimate the gradient update of $V_{j,r}^{(t)}(g')$ where $g' \neq g$. In fact

$$\begin{aligned} -\nabla_{V_{j,r}(g')} \mathbf{Loss}^{(t)} &= \Gamma_{j,r,g'}^{(t)} = \Gamma_{j,r,g'}^{+, (t)} - \Gamma_{j,r,g'}^{-, (t)} \\ &= \Gamma_{j,r,g'}^{+, (t)} - \sum_{y' \in \mathcal{Y}} \mathbb{E}\left[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbb{1}_{\mathcal{H}_{(g', y')}}\right] \\ &= -\mathbb{E}\left[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbb{1}_{\mathcal{H}_{(g', y)}}\right] \pm \max\{\tilde{O}(\mu^{q-1}), \varpi\} \end{aligned}$$

where the last one is because for $y' \neq y$, the activation gradient $\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})$ is either positive and bounded by $\tilde{O}(\mu^{q-1})$ or negative and bounded by ϖ while $\mathbf{logit}_{5,j} \mathbb{1}_{\mathcal{H}_{\phi'}} \leq 1$. Similarly we can also get the same for $V_{j,r}(y')$ where $y' \neq y$:

$$-\nabla_{V_{j,r}(y')} \mathbf{Loss}^{(t)} = \Gamma_{j,r,y'}^{(t)} = -\mathbb{E}\left[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbb{1}_{\mathcal{H}_{(g', y)}}\right] \pm O(\varpi/n_y^2)$$

To prove (c), notice that when $V_{j,r}(g', y') < -B - 2\mu$, the part of $\Gamma_{j,r,g'}^{+, (t)} = 0$ due to $\Lambda_{5,j,r}$ exceeding the boundary $-B$. We can group different $\tilde{y} \neq y \in \mathcal{Y}$ by their feature magnitude and obtain the lower bound. The factor of $\frac{\lambda}{dn_y}$ is due to the logit lower bound from Fact B.2. \square

Lemma B.5 (Feature Magnitude). *Assume Inductions B.1 and B.3 holds, then for any $\psi \in \Sigma^*$,*

(a) at $t = T_{\psi}^{2,1}$, we have $V_{j,r}^{(t)}(g) = \frac{1}{4} \log d \pm o(1)$, $V_{j,r}^{(t)}(y) = \frac{1}{4} \log d \pm o(1)$;

(b) when $\mathcal{C}_{\psi}^+(\delta)$ hold for some $\delta \leq O(\lambda/d^{c_1})$ at some $t \geq T_{\psi}^{2,1}$, we have $V_{\psi}^{(t)} \geq B - O(d^{c_1} \mu)$, that is, $\mathcal{F}_{\psi,1}(d^{c_1} \mu)$ holds.

Proof. Let's proceed the proof one by one.

- Proof of (a) : This can be computed by comparing the gradient of $V_{j,r}^{(t)}(g)$ which is $\Gamma_{j,r,g}^{(t)} = \Gamma_{j,r,g}^{+, (t)} - \Gamma_{j,r,g}^{-, (t)}$ and the gradient of $V_{j,r}^{(t)}(y)$ which is $\Gamma_{j,r,y}^{(t)} = \Gamma_{j,r,y}^{+, (t)} - \Gamma_{j,r,y}^{-, (t)}$. Now since

$$\Gamma_{j,r,g}^{+, (t)} = \Gamma_{j,r,y}^{+, (t)} \geq (1 - O(\frac{1}{\sqrt{d}})) \mathbf{sReLU}'(V_{j,r}^{(t)}(g) + V_{j,r}^{(t)}(y) + O(\mu))$$

and that

$$\Gamma_{j,r,g}^{-, (t)} = O(\frac{1}{\sqrt{d}} \mathbf{sReLU}'(V_{j,r}^{(t)}(g) + O(\mu))), \quad \Gamma_{j,r,y}^{-, (t)} = O(\frac{1}{\sqrt{d}} \mathbf{sReLU}'(V_{j,r}^{(t)}(y) + O(\mu)))$$

one could see that $\sum_{t \leq T_{\psi}^{2,1}} |\Gamma_{j,r,g}^{(t)} - \Gamma_{j,r,y}^{(t)}| \leq \tilde{O}(\frac{1}{\sqrt{d}})$ and thus (a) holds.

- Proof of (b): from lemma B.4a we know that when $\mathcal{C}_{\psi}^+(\delta)$ holds for some small $\delta = O(\frac{\lambda}{d^{c_1}})$ it is not because $V_{\psi}^{(t)}$ dropped below $\frac{1}{2} \log d$, thus when Inductions B.1 and B.3 holds, it can only be because

$V_\psi^{(t)} \geq B - O(d^{c_1}\mu)$, otherwise

$$\Gamma_\psi^{+, (t)} = \mathbb{E}[(1 - \mathbf{logit}_{5,j}^{(t)}) \mathbb{1}_{\mathcal{H}_\phi}] \geq \lambda \Pr(\mathcal{H}_\phi) \gg \frac{\lambda}{d^{c_1}} \quad (\text{if } V_\psi^{(t)} \leq B - O(d^{c_1}\mu))$$

which concludes the proof. \square

Lemma B.6 (Gradient Stationarity). *Let $\psi \in \Sigma^*$, assume Inductions B.1 and B.2 holds at $t \in [T_\psi^1, T_\psi^2]$, and choose $\delta \gg n_y \varpi$, then we have the following guarantees:*

- (a) Define $\mathcal{B}_\delta^+ = \{t \in [T_\psi^{2,1}, T_\psi^2] \mid \mathcal{C}_\psi^+(\delta) \text{ doesn't hold}\}$, then $|\mathcal{B}_\delta^+| \leq O(\frac{n_y \log^2 d}{\eta \delta})$;
- (b) Define $\mathcal{B}_\delta^- = \{t \in [T_\psi^{2,1}, T_\psi^2] \mid \mathcal{C}_\psi^-(\delta) \text{ doesn't hold}\}$, then $|\mathcal{B}_\delta^-| \leq O(\frac{n_y \log^3 d}{\eta \delta})$;
- (c) Finally, at $t = T_\psi^2$, the feature shape $\mathcal{F}_\psi(\delta_1, \delta_2)$ holds with $\delta_1 = d^{c_1}\mu$ and $\delta_2 = (\frac{n_y^3 d \varpi}{\lambda})^{\frac{1}{q-1}}$.

Proof. Let us start with (a). First by writing $\psi = (j, r, \phi)$, we define the following quantity

$$\Upsilon^{(t)} := V_{j,r}^{(t)}(\phi) - \sum_{\phi' \in \Phi_j^* \setminus \{\phi\}} V_{j,r}^{(t)}(\phi') \quad (33)$$

which sits at the central of our proof of counter-argument. Now we have for any $T \geq T_\psi^{2,1}$

$$\begin{aligned} \Upsilon^{(t)} &= V_{j,r}^{(T)}(\phi) - \sum_{\phi' \in \Phi_j^* \setminus \{\phi\}} V_{j,r}^{(T)}(\phi') \\ &= V_{j,r}^{(T_\psi^{2,1})}(\phi) - \sum_{\phi' \in \Phi_j^* \setminus \{\phi\}} V_{j,r}^{(T_\psi^{2,1})}(\phi') + \sum_{s \in [T_\psi^{2,1}, T]} \eta \Gamma_{j,r,\phi}^{(t)} - \sum_{\phi' \in \Phi_j^* \setminus \{\phi\}} \sum_{s \in [T_\psi^{2,1}, t]} \eta \Gamma_{j,r,\phi'}^{(t)} \end{aligned}$$

We can further rewrite the Γ s on the RHS to

$$\begin{aligned} &\sum_{t \in [T_\psi^{2,1}, T]} \eta \left(\Gamma_{j,r,\phi}^{(t)} - \sum_{\phi' \in \Phi_j^* \setminus \{\phi\}} \eta \Gamma_{j,r,\phi'}^{(t)} \right) \\ &= \sum_{t \in [T_\psi^{2,1}, T]} \eta \left(\Gamma_{j,r,\phi}^{+, (t)} - \Gamma_{j,r,\phi}^{-, (t)} - \sum_{\phi' \neq \phi \in \Phi_j^*} \Gamma_{j,r,\phi'}^{+, (t)} + \sum_{\phi' \neq \phi \in \Phi_j^*} \Gamma_{j,r,\phi'}^{-, (t)} \right) \end{aligned} \quad (34)$$

Now we can start to estimate the RHS. Define $P_1 = \{\phi' \neq \phi \in \Phi^* \mid \phi' \text{ learned}\}$ ⁷ and $P_2 = \{\phi' \neq \phi \in \Phi^* \mid \phi' \text{ not learned}\}$. The sum $\sum_{\phi' \neq \phi \in \Phi_j^*} \Gamma_{j,r,\phi'}^{+, (t)}$ can be decomposed and bounded by

$$- O(\eta n_y \varpi) \leq \sum_{\phi' \in P_1} \eta \Gamma_{j,r,\phi'}^{+, (t)} + \sum_{\phi' \in P_2} \eta \Gamma_{j,r,\phi'}^{+, (t)} \leq \tilde{O}(\eta \lambda \mu^{q-2})$$

and then by the same methods of bounding the gradients, we have that

$$\begin{aligned} &\sum_{\phi' \neq \phi \in \Phi_j^*} \Gamma_{j,r,\phi'}^{-, (t)} \\ &= \sum_{\phi' = (g', y') \in \Phi_j^* \setminus \{\phi\}} \mathbb{E} \left[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \left(\sum_{g'' \neq g' \in \mathcal{G}} \mathbb{1}_{\mathcal{H}_{(g'', y')}} + \sum_{y'' \neq y' \in \mathcal{Y}} \mathbb{1}_{\mathcal{H}_{(g', y'')}} \right) \right] \end{aligned}$$

⁷ By ϕ' is learned we mean there is a feature $\psi' \prec \psi \in \Sigma^*$ such that $\psi' = (j, r, \phi')$.

$$\begin{aligned}
&= \sum_{g' \neq g} \mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)} \mathbb{1}_{\mathcal{H}_{(g',y)}})] + \sum_{y' \neq y} \mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)} \mathbb{1}_{\mathcal{H}_{(g,y')}})] \pm O(\varpi) \\
&= \Gamma_{j,r,\phi}^{-(t)} \pm O(\varpi)
\end{aligned}$$

where ① is because for any incorrect feature combination $\phi' = (g', y') \notin \Phi_j^*$ and does not share a component with ϕ , the activation of $\Lambda_{5,j,r}$ conditioned on $\mathcal{H}_{\phi'}$ is within $[-B, \tilde{O}(\mu^{q-1})]$ and thus the negative gradient is upper bounded by ϖ/n_y^2 and lower bounded by $-\tilde{O}(\mu^{q-1})$. Inserting this back to Eq. (34), we shall have

$$\sum_{t \in [T_{\psi}^{2,1}, T)} \eta \left(\Gamma_{j,r,\phi}^{(t)} - \sum_{\phi' \in \Phi_j^* \setminus \{\phi\}} \eta \Gamma_{j,r,\phi'}^{(t)} \right) = \sum_{t \in [T_{\psi}^{2,1}, T)} \eta \Gamma_{j,r,\phi}^{+, (t)} \pm O((T - T_{\psi}^{2,1}) \eta \varpi)$$

Now we have arrived at the desired estimate of the update, finally, we have the following counter-argument: if $\mathcal{C}_{\psi}^+(\delta)$ does not hold for $\Omega(\frac{n_y \log^2 d}{\eta \delta})$ iterations, letting $T - 1$ be the last iteration which $\mathcal{C}_{\psi}^+(\delta)$ does not hold, we have

$$\begin{aligned}
\Upsilon^{(T)} &\geq \sum_{t \in [T_{\psi}^{2,1}, T)} \eta \Gamma_{j,r,\phi}^{+, (t)} \\
&\geq \sum_{t \in [T_{\psi}^{2,1}, T] \setminus B_{\delta}^+} \eta \delta + \Omega((T - T_{\psi}^{2,1}) \eta \varpi) \quad (\text{For } t \in B_{\delta}^+, \Gamma_{j,r,\phi}^{+, (t)} \geq 0) \\
&\geq \sum_{t \in [T_{\psi}^{2,1}, T] \setminus B_{\delta}^+} \eta \delta - O(\frac{n_y \log^2 d}{\eta \delta} \eta \varpi) \\
&\geq \Omega(\frac{n_y \log^2 d}{\eta \delta}) \eta \delta \quad (\text{by choosing } \delta \gg \varpi) \\
&\geq \Omega(n_y \log^2 d)
\end{aligned}$$

which is impossible because the gradient would have vanished when $\Upsilon^{(T)} \geq n_y \log^{1.5} d$ as that would mean $\exists \phi \in \Phi_j^*$ such that $|V_{j,r}^{(t)}(\phi)| \geq \log^2 d$ and it is forbidden in our setting, which proves the claim via contradiction.

Proof of lemma B.6b. Actually, by applying (a) combined with lemma B.5b, we know that for some $T \geq T_{\psi}^{2,1} + \tilde{\Omega}(\frac{1}{\eta \lambda})$ it holds $V_{\psi}^{(T)} \geq B - O(d^{c_1} \mu)$. From this iteration forward we shall prove that $V_{\psi}^{(t)}$ stays above $B - O(d^{c_1} \mu)$ and can help. In fact, at $t = T$, $V_{j,r}^{(t)}(V) \in (-O(B), \tilde{O}(\mu))$ for $v \in \mathcal{G} \cup \mathcal{Y} \setminus \{g, y\}$ by Induction B.3. Say $v = g' \neq g$, then due to lemma B.4, its gradient is approximated by

$$-\mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)} \mathbb{1}_{\mathcal{H}_{(g',y)}})] \pm O(\varpi/n_y) \quad (35)$$

Let $\delta \geq \varpi$, we will find an iteration $t \leq T_{\psi}^{2,1} + O(\frac{n_y \log^3 d}{\eta \delta})$ that satisfy $\mathcal{C}_{\psi}^-(\delta)$. By applying (a), we first get that $\mathcal{C}_{\psi}^+(\delta)$ holds for no more than $O(\frac{n_y \log^2 d}{\eta \delta})$ iters after $T_{\psi}^{2,1}$ and before T_{ψ}^2 .

Let's assume this is the case, that $\mathcal{C}_{\psi}^-(\delta)$ doesn't hold for more than $\Omega(\frac{n_y \log^3 d}{\eta \delta})$ iterations. In this case, it must be that the gradient $\Gamma_{j,r,\phi}^{-(t)} < -\delta$ is negative, since when the gradient is positive it has an upper bound $O(\varpi/n_y) = o(\delta)$ and therefore cannot violate the condition $\mathcal{C}_{\psi}^+(\delta)$. Therefore the updates $\sum_{s \leq t} \Gamma_{j,r,\phi}^{-(s)}$ must accumulate to lower than

$$V_{j,r}^{(t)}(\phi) \leq \Upsilon^{(t)} - \Omega(\frac{n_y \log^3 d}{\eta \delta}) \times \eta \delta = -\Omega(n_y \log^3 d)$$

which is impossible in our setting. This proved the result for any $\delta \gg n_y \varpi$.

Proof of lemma B.6c. Let $\delta = n_y^2 \varpi$, which is covered by lemma B.6a-b, we can find an iteration where

$|\Gamma_{j,r,\phi'}^{(t)}| \leq \delta$ $\phi' \in \Phi_j^* \setminus \{\phi\}$. This implies we have found a t such that

$$|\mathbb{E}[\text{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}) \mathbf{1}_{\mathcal{H}_{\phi'}}]| \leq n_y^2 \varpi \implies |\mathbf{sReLU}'(V_{j,r}^{(t)}(\phi') + O(\mu))| \leq O(\frac{n_y^3 d \varpi}{\lambda})$$

(We lower bounded the logit by Fact B.2)

Thus we arrive at $|V_{j,r}(\phi')| \leq O((\frac{n_y^3 d \varpi}{\lambda})^{\frac{1}{q-1}})$. Remember in the proof of b, $V_{\psi}^{(t)} = \frac{1}{2}(V_{j,r}^{(t)}(g) + V_{j,r}^{(t)}(y)) \geq B - O(d^{c_1} \mu)$ at $t \geq T + \tilde{\Omega}(\frac{d^{c_1}}{\eta \lambda})$. So we have for any $\phi' = (g', y') \in \Phi_j^* \setminus \{\phi\}$,

$$\frac{1}{2}(V_{j,r}^{(t)}(g') + V_{j,r}^{(t)}(y')) \leq -\frac{1}{2}(V_{j,r}^{(t)}(g) + V_{j,r}^{(t)}(y)) + O(\frac{1}{d^{1/3q}}) \leq -B + O((\frac{n_y^3 d \varpi}{\lambda})^{\frac{1}{q-1}})$$

Now we have verified that given $\delta = n_y^2 \varpi$, there is a iteration $T_{\psi}^{2,3}$ which gives the near-perfect feature shape $\mathcal{F}_{\psi}(\delta_1, \delta_2)$ with $\delta_1 = d^{c_1} \mu$, $\delta_2 = (\frac{n_y^3 d \varpi}{\lambda})^{\frac{1}{q-1}}$. We shall prove that this feature shape will hold until $t = T_2$.

For $t \geq T_{\psi}^{2,3}$, we verify $\mathcal{F}_{\psi}(\delta_1, \delta_2)$ in Definition B.4 one by one:

1. $\mathcal{F}_{\psi,1}(d^{c_1} \mu)$: By lemma B.6a we shall find a step after $t \geq T_{\psi}^{2,1} + O(\frac{d^{c_1}}{\eta \lambda})$ at which $\mathcal{F}_{\psi,1}(d^{c_1} \mu)$ is satisfied. In fact, the majority of the steps after $T_{\psi}^{2,1} + O(\frac{d^{c_1}}{\eta \lambda})$ will satisfy this. We shall prove that combined with the above feature shape guarantee, $\mathcal{F}_{\psi,1}(d^{c_1} \mu)$ will hold true until $t = T_1$.
2. $\mathcal{F}_{\psi,2}((\frac{n_y^3 d \varpi}{\lambda})^{\frac{1}{q-1}})$: From the above argument that applies lemma B.6b, we shall find a time step where $\mathcal{F}_{\psi,1}(d^{c_1} \mu)$ and $\mathcal{F}_{\psi,2}(d^{-1/3q})$ hold. Now we discuss the possible values of the gradient $\Gamma_{\psi'}^{-(t)}$, $\psi' \in \Sigma_{\psi}^{\dagger,1}$. In fact, let $\phi' = (g', y') \in \Phi_j^* \setminus \{\phi\}$, we have

- Regime A: If $V_{j,r}(\phi') < -B - 2\mu$, then there is no negative gradient for $V_{j,r}(g')$ and $V_{j,r}(y')$ unless $V_{j,r}(g', y)$ or $V_{j,r}(g, y')$ is positive. In fact, only one of the feature can be positive to have negative gradient, and in that case the gradient of $\Gamma_{j,r,v}^{(t)}$, $v \in \{g', y'\}$ is pointing to the convergence direction $|V_{j,r}(g', y)| \rightarrow 0$.
- Regime B: If $V_{j,r}(\phi') \geq -B - 2\mu$, then there is a possibility of negative gradient for $V_{j,r}(g')$ and $V_{j,r}(y')$ even when $V_{j,r}(g', y)$ or $V_{j,r}(g, y')$ is negative, In this case the feature $V_{j,r}(g', y)$ or $V_{j,r}(g, y')$ might not go the desired direction to converge to zero, but given a $\delta > n_y^2 \varpi$, $\Gamma_{j,r,(g',y)}^{-(t)}$ or $\Gamma_{j,r,(g,y')}^{-(t)}$ cannot stay above δ , otherwise the term will dominate and revert the gradient direction.

On the other hand, when one side has $V_{j,r}(g, y') < -\Omega((\frac{n_y^3 d \varpi}{\lambda})^{\frac{1}{q-1}})$, we have

$$V_{j,r}(\phi') \leq -V_{j,r}(\phi) - \Omega((\frac{n_y^3 d \varpi}{\lambda})^{\frac{1}{q-1}}) \leq -B - 2\mu$$

which is back in regime A and therefore will converge back.

Induction over the above two regimes for all $t \in [T_{\psi}^{2,3}, T_{\psi}^2]$ showed that the feature will indeed satisfy $\mathcal{F}_{\psi,2}(\delta)$ for all $\delta > n_y^2 \varpi$ at t .

3. $\mathcal{F}_{\psi,3}((\frac{n_y^3 d \varpi}{\lambda})^{\frac{1}{q-1}})$: We need to use both $\mathcal{F}_{\psi,1}$ and $\mathcal{F}_{\psi,2}$ to get the desired result. In fact, we have for any $\psi' = (j, r, (g', y')) \in \Sigma^{\dagger,2}$,

$$V_{\psi'}^{(t)} = V_{j,r}^{(t)}(g') + V_{j,r}^{(t)}(y') + V_{j,r}^{(t)}(g) + V_{j,r}^{(t)}(y) - V_{j,r}^{(t)}(\phi) = -B + O((\frac{n_y^3 d \varpi}{\lambda})^{\frac{1}{q-1}})$$

because $V_{j,r}^{(t)}(g) + V_{j,r}^{(t)}(y')$ and $V_{j,r}^{(t)}(g') + V_{j,r}^{(t)}(y)$ are bounded due to $\mathcal{F}_{\psi,2}$.

4. $\mathcal{F}_{\psi,4}(d^{c_1} \mu)$: The number of iterations $[T_{\psi}^1, T_{\psi}^2]$ is no more than $\tilde{O}(\frac{1}{\eta(d^{c_1} \mu)^{q-2}})$, so by using lemma B.13 along with the induction from the result at T_{ψ}^1 shall conclude the proof.

5. $\mathcal{F}_{\psi,5}((\frac{n_y^3 d\varpi}{\lambda})^{\frac{1}{q-1}})$: By using lemma B.7 and $\mathcal{F}_{\psi,2}$ after $t \geq T_{\psi}^{2,3}$, we have the result.

This proved (c). \square

Lemma B.7 (Symmetry of Features). *Let $\psi = (j, r, (g, y)) \in \Sigma^*$ and $t \geq T_{\psi}^{2,3}$, we have that if $|V_{\psi'}| \leq \delta$ for all $\psi' \in \Sigma^{\dagger,1}$, then*

$$|V_{j,r}^{(t)}(g) - V_{j,r}^{(t)}(y)| \leq O(\delta)$$

Proof. For a feature $\psi = (j, r, \phi) \in \Sigma^*$, $\phi = (g, y)$, their update can be represented as:

$$\begin{aligned} V_{j,r}^{(t)}(g) &= U_{g,y} - \sum_{y' \neq y} R_{g,y'} + V_{j,r}^{(0)}(g) \\ V_{j,r}^{(t)}(y) &= U_{g,y} - \sum_{g' \neq g} R_{g',y} + V_{j,r}^{(0)}(y) \end{aligned}$$

where $U_{g,y} = \sum_{s \leq t} \eta \Gamma_{j,r,\phi}^{+, (s)}$ and $R_{g,y'}$ is

$$R_{g,y'} = \sum_{s \leq t} \eta \mathbb{E}[\text{logit}_{5,j}^{(s)} \text{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbf{1}_{\mathcal{H}_{g',y}}]$$

Therefore:

$$\begin{aligned} \sum_g V_{j,r}^{(t)}(g) &= \sum_{\phi \in \Phi_j^*} U_{\phi} - \sum_g \sum_{y' \neq y} R_{g,y'} = \sum_{\phi \in \Phi_j^*} U_{\phi} - \sum_{\phi^{\dagger} \in \Phi_j^{\dagger}} R_{\phi^{\dagger}} + V_{j,r}^{(0)}(g) \\ \sum_y V_{j,r}^{(t)}(y) &= \sum_{\phi \in \Phi_j^*} U_{\phi} - \sum_y \sum_{g' \neq g} R_{g',y} = \sum_{\phi \in \Phi_j^*} U_{\phi} - \sum_{\phi^{\dagger} \in \Phi_j^{\dagger}} R_{\phi^{\dagger}} + V_{j,r}^{(0)}(y) \end{aligned}$$

Thus

$$\sum_{g'} V_{j,r}^{(t)}(g') = \sum_{y'} V_{j,r}^{(t)}(y') \pm O(\mu)$$

Moreover, we have assumed

$$|V_{j,r}^{(t)}(g) + V_{j,r}^{(t)}(y')| = O(\delta), \forall y' \neq y, \quad |V_{j,r}^{(t)}(y) + V_{j,r}^{(t)}(g')| = O(\delta), \forall g' \neq g$$

Thus

$$\sum_{g'} V_{j,r}^{(t)}(g') = V_{j,r}^{(t)}(g) - (n_y - 1)V_{j,r}^{(t)}(y) \pm O(n_y \delta) = V_{j,r}^{(t)}(y) - (n_y - 1)V_{j,r}^{(t)}(g) \pm O(n_y \delta)$$

Therefore

$$V_{j,r}^{(t)}(g) = V_{j,r}^{(t)}(y) + O(\delta) + O(\mu)$$

inserting $\delta = O((\frac{n_y^2 d\varpi}{\lambda})^{\frac{1}{q-1}})$ which is from lemma B.6 proves the claim. \square

Lemma B.8 (Arrival time estimates). *Assuming Inductions B.1 and B.3 for $\psi \in \Sigma^*$, we have*

$$(a) \quad T_{\psi}^{2,1} - T_{\psi}^1 \leq \tilde{O}(1/\eta(d^{c_1} \sigma_0)^{q-2});$$

$$(b) \quad T_{\psi}^{2,2} - T_{\psi}^{2,1} \leq O(\frac{d^{\frac{1}{2}+c_1}}{\eta});$$

$$(c) \quad T_{\psi}^{2,3} - T_{\psi}^{2,2} \leq \tilde{O}(\frac{1}{\eta \varpi});$$

$$(d) \ T_\psi^2 - T_\psi^1 \leq \tilde{O}\left(\frac{1}{\eta(d^{2c_1}\mu)^{q-2}}\right)$$

Proof. • Proof of (a): Once $V_\psi^{(t)} \geq d^{c_1}\sigma_0$, we can apply lemma B.13 to compute the number of iteration needed to reach $V_\psi^{(t)} \geq \frac{1}{2} \log d$. In fact, by Induction B.3a-b we know that

$$V_\psi^{(t+1)} \geq V_\psi^{(t)} + \eta(1 - O(\frac{1}{\sqrt{d}}))\text{sReLU}'(V_\psi^{(t)} + O(\mu)) \geq V_\psi^{(t)} + \Omega(\eta(V_\psi^{(t)})^{q-1})$$

Since $V_\psi^{(T_\psi^1)} \geq d^{c_1}\sigma_0$, we have by lemma B.13 that $T_\psi^{2,1} - T_\psi^1 \leq O(\frac{1}{\eta(d^{c_1}\sigma_0)^{q-2}})$.

- Proof of (b): Simply apply lemma B.6 with $\delta = \frac{1}{d^{\frac{1}{2}+c_1}}$;
- Proof of (c): Again by applying lemma B.6 with $\delta = n_y^2\varpi$;
- Proof of (d): Follow the definition of T_ψ^2 .

□

Proof of Induction B.3. The proof of time steps are all in lemma B.8 and lemma B.6. We do not repeat here. □

B.4 Phase III: Convergence

We present the induction hypothesis in this phase.

Induction B.4 (Phase III, final). *Let $\psi = (j, r, \phi) \in \Sigma^*$, for $t \in [T_\psi^2, T_1]$, the following holds:*

- (a) *At $t \in [T_\psi^2, T_{1,1}]$, we have $\mathcal{F}_\psi(\delta_1, \delta_2)$ holds with $\delta_1 = d^{c_1}\mu$, $\delta_2 = (\frac{n_y^2 d \varpi}{\lambda})^{\frac{1}{q-1}}$;*
- (b) *At $t \in [T_\psi^3, T_1]$, we have $\mathcal{F}_\psi(\delta_1, \delta_2)$ holds with $\delta_1 = d^{c_1}\mu$, $\delta_2 = (n_y^2 \varpi)^{\frac{1}{q-1}}$.*

We need a lemma to describe the logit distribution when all features are learned.

Lemma B.9 (Logit shape at convergence). *Assuming Inductions B.1 and B.4. Let $\psi \in \Sigma^*$ be the last of Σ^* , then at $t = T_\psi^2$, the followings hold:*

1. *For $\phi \in \Phi_j^*$, $\text{logit}_{5,j}^{(t)} \geq 1 - O(\lambda)$ conditioned on \mathcal{H}_ϕ .*
2. *For $\phi \notin \Phi_j^*$, $\text{logit}_{5,j}^{(t)} \leq O(\lambda/d)$ conditioned on \mathcal{H}_ϕ .*

Proof. Since ψ is the last feature in the learning curriculum Σ^* , all feature at $t = T_\psi^2$ has satisfied $\mathcal{F}_\psi(\delta_1, \delta_2)$ for $\delta_1 = d^{c_1}\mu$, $\delta_2 = (\frac{n_y^2 d \varpi}{\lambda})^{\frac{1}{q-1}}$. Thus we can bound the logit by

$$\text{logit}_{5,j}^{(t)} = \frac{e^{F_{5,j}^{(t)}(\mathbf{Z})}}{e^{F_{5,j}^{(t)}(\mathbf{Z})} + d} = \frac{e^{B-o(1)}}{e^{B-o(1)} + d} = 1 - O(\lambda) \quad (\text{conditioned on } \mathbf{Z} \in \mathcal{H}_\phi, \phi \in \Phi_j^*)$$

and

$$\text{logit}_{5,j}^{(t)} = \frac{e^{F_{5,j}^{(t)}(\mathbf{Z})}}{e^{F_{5,j}^{(t)}(\mathbf{Z})} + d} = \frac{e^{o(1)}}{e^{B-o(1)} + d - 1 + e^{o(1)}} = O(\frac{\lambda}{d}) \quad (\text{conditioned on } \mathbf{Z} \in \mathcal{H}_\phi, \phi \in \Phi_j^*)$$

which is the desired result. □

Lemma B.10 (Gradient Bounds, Phase III). *Assuming Inductions B.1 and B.4 holds, and let $\psi = (j, r, \phi) \in \Sigma^*$, then at $t \in [T_{1,1}, T_1]$, the followings hold*

(a) $\Gamma_{\psi}^{+, (t)} \geq \Omega(\lambda)$ when $V_{\psi}^{(t)} \leq B - d^{c_1} \mu$;

(b) $|\Gamma_{j,r,v}^{-, (t)}| \leq O(\frac{\lambda \varpi}{dn_y})$ for any $v \in \mathcal{G} \cup \mathcal{Y}$;

(c) For $\phi' \in \Phi_j^* \setminus \{\phi\}$, we have

$$\Gamma_{j,r,\phi'}^{+, (t)} = \begin{cases} -\Theta(\lambda \varpi / n_y^2), & \text{if } V_{j,r,\phi'}^{(t)} \in [-B + 2\mu, -\varpi] \\ \in [-\Theta(\lambda \varpi / n_y^2), -\Theta(\lambda \varpi / dn_y^2)] \cup \{0\}, & \text{if } V_{j,r,\phi'}^{(t)} \in [-B - 2\mu, -B + 2\mu] \end{cases}$$

Proof. lemma B.10a is simple when lemma B.9a holds. lemma B.10b is basically redoing the calculations in lemma B.4. lemma B.10c is because when $V_{j,r,\phi'}^{(t)} \geq -B + 2\mu$, we have

$$\begin{aligned} \mathbb{E}[\mathcal{E}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\mathcal{H}'_{\phi}}] &= \mathbb{E}[(1 - \mathbf{logit}_{5,j}^{(t)}) \mathbf{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\mathcal{H}'_{\phi}}] \\ &\geq \lambda \mathbf{sReLU}'(V_{j,r,\phi'}^{(t)} + \mu \pm o(\mu)) \Pr(\mathcal{H}_{\phi'}) \\ &\geq -\Theta(\lambda \varpi / n_y^2) \end{aligned}$$

And the rest is simply the property of \mathbf{sReLU}' at the negative boundary when some activations of x_0, x_1 crosses the boundary. \square

B.4.1 Proof of Induction in Phase III and theorem B.1

Proof of Inductions B.1, B.2 and B.4. We prove the induction by reusing the proof of lemma B.6.

- Proof of Induction B.4a: The same as in the proof of lemma B.6c, with longer time. We do not use any property of T_{ψ}^3 so no additional handling is needed.
- Proof of Induction B.4b: When all feature is learned, that is, after $t = T_{\psi}^2$ for the last feature $\psi \in \Sigma^*$, we have perfect feature shape and therefore have much better logit shape as described by lemma B.9. We can reuse the proof of lemma B.6a-b to get the following result: for any $\delta = \omega(\frac{\lambda \varpi}{dn_y})$, we can guarantee that $\mathcal{C}_{\psi}(\delta)$ doesn't hold for more than $\tilde{O}(\frac{1}{\eta \delta})$ iterations. And beyond that, we can just reuse the argument in the proof of lemma B.6c to obtain the desired result.
- Proof of Induction B.1a-b: They are correct in Phase III because there is no violation.
- Proof of Induction B.2a-b: They are proven in previous phases.
- Proof of Induction B.2c: For predecessor feature $\psi' \prec \psi \in \Sigma^*$, we proved in this phase that they maintain feature shape until $T_{\psi'}^3$, and improved their feature at $T_{\psi'}^3$.
- Proof of Induction B.2d: This is proven in previous phases.

\square

Proof of theorem B.1. Since all feature in Σ^* have shape $\mathcal{F}_{\psi}(\delta_1, \delta_2)$ for $\delta_1 = d^{c_1} \mu, \delta_2 = \varpi^{\frac{1}{q-1}}$ at the end of Induction B.4, we have proven theorem B.1. \square

B.5 Auxiliary Technical Tools

First we need a Bernstein inequality for U-statistics

Lemma B.11 (concentration inequality for pseudo-U-statistics). *Let x_1, \dots, x_n be different symbols, and let $m \ll n$ be such that $n \equiv 0 \pmod{m}$. Suppose for some function h with $|h| \leq M$ the random variables $h(x_{i_1}, x_{i_2}, \dots, x_{i_m})$ and $h(x_{i'_1}, x_{i'_2}, \dots, x_{i'_m})$ are independent and identically distributed as long as*

$\{x_{i_1}, x_{i_2}, \dots, x_{i_m}\} \cap \{x_{i'_1}, x_{i'_2}, \dots, x_{i'_m}\} = \emptyset$, then the pseudo- U -statistic

$$U_{m,n} = \frac{1}{\binom{n}{m}} \sum_{0 \leq i_1 < i_2 < \dots < i_m \leq n} h(x_{i_1}, x_{i_2}, \dots, x_{i_m})$$

satisfies $\Pr(|U_{n,m} - \mathbb{E}[U_{n,m}]| \geq t) \leq e^{-\frac{nt^2}{mM^2}}$

Proof. The proof is the same as in [Rin16]. \square

We present two lemmas related to the tensor power method.

Lemma B.12 (TPM in expectation). *Let $p > 3$ be an integer, and $\mu, \sigma > 0$ satisfying $\mu \gg \sigma\sqrt{\log d}$. Now suppose $\xi \sim \mathcal{N}(0, \sigma)$ is a Gaussian variable and there are two sequences x_t, y_t defined by the following update rules:*

- $x_{t+1} \geq x_t + \eta C_t \mathbb{E}_\xi[(\xi + x_t)^p]$ for some $C_t = \Theta(1)$;
- $y_{t+1} \leq y_t + \eta C_t \mathbb{E}_\xi[(\xi + y_t)^p]$ for some $S = \Theta(1)$.

Then if $x_0, y_0 \geq \mu, x_0 - y_0 \geq \varepsilon = \Omega(\frac{1}{\text{polylog}(d)})$, for every $t \geq 0$ such that $x_t \in [d^{0.01}x_0, O(1)]$, we shall have $y_t \leq O(\text{polylog}(d))$.

Proof. Let $\delta > 0$ be such that $\delta \gg \frac{\mu}{\sigma}$ and $\delta = o(1)$. Let $T_i, i \geq 0$ be defined as the first time when $x_t \geq (1 + \delta)^i x_0$, and $b := \min\{i | (1 + \delta)^i x_0 \geq A\}$. Now we can compute a growth lower bound for $\varepsilon_t := x_t - y_t$ at each step:

$$\begin{aligned} \varepsilon_{T_{i+1}} &\geq \varepsilon_{T_i} + \sum_{t \in [T_i, T_{i+1})} \eta C_t \mathbb{E}_\xi[(\xi + x_t)^p - (\xi + y_t)^p] \\ &\geq \varepsilon_{T_i} + \varepsilon_{T_i} \sum_{t \in [T_i, T_{i+1})} \eta C_t \mathbb{E}_\xi[(\xi + x_t)^{p-1}] \\ &\geq \varepsilon_{T_i} + \varepsilon_{T_i} \sum_{t \in [T_i, T_{i+1})} \eta C_t \mathbb{E}_\xi[(\xi + x_t)^p] \cdot \frac{1}{(1 + \delta)x_{T_{i+1}}} \\ &\geq \varepsilon_{T_i} + \varepsilon_{T_i} \delta (1 + \delta)^i x_0 \frac{1}{(1 + \delta)^{i+2} x_0} \\ &\geq (1 + \frac{\delta}{(1 + \delta)^2}) \varepsilon_{T_i} \end{aligned}$$

Therefore let i' be such that $(1 + \delta)^{i'} \geq \frac{\mu}{\varepsilon}$, we have that

$$\varepsilon_{T_{i'}} \geq (1 + \frac{\delta}{(1 + \delta)^2})^{i'} \varepsilon_0 \geq (1 - o(1))\mu \gg \sigma\sqrt{\log d} \quad (36)$$

At $t = T_{i'}$ we have $y_{T_{i'}} \leq x_{T_{i'}} \leq (1 + o(1))\frac{\mu}{\varepsilon} x_0$. After $t = T_{i'}$, we can construct a surrogate sequence $\tilde{y}_t \geq y_t$ for $t \geq T_{i'}$ by initializing $\tilde{y}_{T_{i'}} = y_{T_{i'}} + \frac{\mu}{2}$, and updates by

$$\tilde{y}_{t+1} = \tilde{y}_t + \eta C_t \tilde{y}_t^p$$

This update guarantees that $\tilde{y}_t \geq y_t$ because $\tilde{y}_t^p \geq \mathbb{E}_\xi[(\xi + y_t)^p]$ at every step. Now by applying corollary B.1 to the sequence of x_t, \tilde{y}_t using the fact that $x_{T_{i'}} \geq \tilde{y}_{T_{i'}} + \frac{\mu}{2} = \tilde{y}_{T_{i'}}(1 + \frac{1}{\text{polylog}(d)})$, we can get the desired result for \tilde{y}_t and thus y_t at $t = T_b$, i.e, when $x_t \geq A$. \square

Lemma B.13 (TPM, adapted from [AZL20]). *Consider an increasing sequence $x_t \geq 0$ defined by $x_{t+1} = x_t + \eta C_t x_t^{q-1}$ for some integer $q \geq 3$ and $C_t = \Theta(1) > 0$, then we have for every $A > x_0$, for every $\delta > 0$,*

and every $\eta \in (0, 1)$:

$$\begin{aligned} \sum_{t \geq 0, x_t \leq A} \eta C_t &\geq \left(\frac{\delta(1+\delta)^{-1}}{(1+\delta)^{q-2} - 1} \left(1 - \left(\frac{(1+\delta)x_0}{A} \right)^{q-2} \right) - \frac{O(\eta A^{q-1}) \log(A/x_0)}{x_0 \log(1+\delta)} \right) \cdot \frac{1}{x_0^{q-2}} \\ \sum_{t \geq 0, x_t \leq A} \eta C_t &\leq \left(\frac{(1+\delta)^{q-2}}{q-2} + \frac{O(\eta A^{q-1}) \log(A/x_0)}{x_0 \log(1+\delta)} \right) \cdot \frac{1}{x_0^{q-2}} \end{aligned}$$

This lemma has a corollary:

Corollary B.1 (TPM, from [AZL20]). *Let $q \geq 3$ be a constant and $x_0, y_0 = o(1)$ and $A = O(1)$. Let $\{x_t, y_t\}_{t \geq 0}$ be two positive sequences updated as*

- $x_{t+1} = x_t + \eta C_t x_t^{q-1}$ for some $C_t = \Theta(1)$;
- $y_{t+1} = y_t + \eta S C_t y_t^{q-1}$ for some $S = \Theta(1)$.

Suppose $x_0 \geq y_0 S^{\frac{1}{q-2}} (1 + \frac{1}{\text{polylog}(d)})$, letting T_x be the first iteration s.t., $x_t \geq A$, then $y_{T_x} \leq \tilde{O}(y_0)$.

C Learning Symmetry Group Actions

Roadmap. We begin by introducing notation and defining fibers, followed by a description of the learning curriculum and the key appearance events. We then break the training process into clearly defined phases and prove the main results for each phase: In Phase I, features emerge and compete; In Phase II, features grow and mutually cancel, with precise tracking of when different outcomes occur; In Phase III, features converge to a desired shape.

Let us restate the action structure of \mathcal{G} and \mathcal{Y} here.

Assumption C.1 (Assumption 5.1, restated). Let $\mathbf{LEGO}(\mathcal{X}, \mathcal{G}, \mathcal{Y})$ be the LEGO language, where $|\mathcal{Y}| = n_Y$ ⁸ and \mathcal{G} is a group acting on \mathcal{Y} . We assume the (left) group action $\alpha : \mathcal{G} \times \mathcal{Y} \rightarrow \mathcal{Y}$ is transitive but not free. In particular, we assume $n_Y = |\mathcal{Y}| \in [\omega(1), O(\log d)]$ and the group \mathcal{G} satisfy $n_G = |\mathcal{G}| = \Theta(\text{polylog}(d)) > \frac{1}{\epsilon}$ for all $y \in \mathcal{Y}$.

Remark C.1. *When the action $\alpha : \mathcal{G} \times \mathcal{Y} \rightarrow \mathcal{Y}$ is transitive, there is only one orbit $\mathcal{Y} = \{g \cdot y \mid g \in \mathcal{G}\} \simeq \mathcal{G}/G_y$ for all $y \in \mathcal{Y}$. By orbit-stabilizer theorem, the stabilizer G_y at any point y has the same cardinality.*

Remark C.2. *The canonical action of a symmetry group S_n on \mathbb{Z}_n satisfy Assumption C.1 with the choice $n_y = \Theta(\frac{\log \log d}{\log \log \log d})$ that keeps $n_y! = O(\text{polylog}(d))$.*

C.1 Preliminaries

We work with fiber-indexed features. We first define fibers and super-combinations φ , then we define neuron features by its index $\psi = (j, r, \varphi)$, and finally set the curriculum and appearance events in this notation.

Definition C.1 (Fiber of values). Assuming \mathcal{G} follows Assumption C.1. For each $j \in \tau(\mathcal{Y})$ and $y \in \mathcal{Y}$, define the **fiber** of j at y as:

$$\text{Fiber}_{j,y} := \{g \in \mathcal{G} \mid \tau(g \cdot y) = j\}$$

which collects all group elements that send y to $y' = \tau^{-1}(j)$.

Remark C.3. *We denote $\text{Fiber}_{j,y}$ to be the set that transport y to $y' = \tau^{-1}(j)$ because it is the pre-image of the predictor map $\phi_y(\cdot) : g \mapsto \tau(g \cdot y)$ over $y \in \mathcal{Y}$. Algebraically the fibers are the left-cosets $\{gG_y\}_{g \in \mathcal{G}}$, where $G_y := \{g \in \mathcal{G} \mid g \cdot y = y\}$ is the stablizer subgroup. Moreover, we have the following fact from the orbit-stablizer theorem.*

⁸ We use n_y or n_Y interchangeably in the proof below.

Fact C.1 (cardinality of the fibers). For every $j \in \tau(\mathcal{Y})$ and $y \in \mathcal{Y}$, $|\text{Fiber}_{j,y}| = |\mathcal{G}|/|\mathcal{Y}| = n_G/n_Y$ from the orbit-stablizer theorem.

Definition C.2 (Super-combinations). Let $(j, y) \in \tau(\mathcal{Y}) \times \mathcal{Y}$, we define the **super-combinations**, which is a set of combinations:

$$\varphi_{j,y} := \text{Fiber}_{j,y} \times \{y\} \subseteq \mathcal{G} \times \mathcal{Y}, \quad \varphi_{j,y}^1 = \text{Fiber}_{j,y}, \quad \varphi_{j,y}^2 = \{y\}$$

Let $\Phi := \cup_{j \in \tau(\mathcal{Y}), y \in \mathcal{Y}} \varphi_{j,y}$ be the full set. For a fixed class $j \in \tau(\mathcal{Y})$, we denote the **correct** set to be $\Phi_j^\star := \cup_{y \in \mathcal{Y}} \varphi_{j,y}$, and the **incorrect** set $\Phi_j^\dagger := \cup_{j' \neq j, y \in \mathcal{Y}} \varphi_{j',y}$. A base pair $\phi = (g, y)$ belongs to $\varphi_{j,y}$ if and only if $g \in \text{Fiber}_{j,y}$. To simplify the proof, we also define for each $\varphi_{j,y} \subset \Phi$, the confounding combinations

$$\Phi_\varphi^\dagger = \bigcup_{j' \neq j \text{ XOR } y' \neq y} \varphi_{j',y'}$$

That is, the set of combinations that intersect with exactly one component with $\varphi_{j,y}$.

Definition C.3 (Neuron feature indices). Let us index neuron features by $\psi = (j, r, \varphi)$ with $\varphi \in \Phi$ and $(j, r) \in \tau(\mathcal{Y}) \times [m]$, and define

$$\Psi_{j,r}^\star := \{(j, r, \varphi) : \varphi \in \Phi_j^\star\}, \quad \Psi_{j,r}^\dagger := \{(j, r, \varphi) : \varphi \in \Phi_j^\dagger\}.$$

And we further define

$$\Psi := \tau(\mathcal{Y}) \times [m] \times \Phi, \quad \Psi^\star = \bigcup_{(j,r) \in \tau(\mathcal{Y}) \times [m]} \Psi_{j,r}^\star, \quad \Psi^\dagger = \bigcup_{(j,r) \in \tau(\mathcal{Y}) \times [m]} \Psi_{j,r}^\dagger$$

The set Ψ contains all the neuron indices we care about, and Ψ^\star and Ψ^\dagger contain the correct and incorrect indices for class j .

V-Notations We call a pair $\phi = (g, y)$ with $g \in \mathcal{G}$ and $y \in \mathcal{Y}$ a *base pair*, or following the naming of appendix B, a *base combination*. For neuron (j, r) , we define

$$V_{j,r}(g) := \langle \mathbf{W}_{5,j,r,2}, e_g \rangle, \quad V_{j,r}(y) := \langle \mathbf{W}_{5,j,r,5}, e_y \rangle, \quad V_{j,r}(\phi) := \frac{1}{2}(V_{j,r}(g) + V_{j,r}(y)).$$

For a fixed value y and class j , the fiber $\text{Fiber}_{j,y}$ collects all permutations sending y to $\tau^{-1}(j)$. For $\psi = (j, r, \varphi_{j,y})$, we define:

$$V_\psi := \frac{1}{n_G/n_Y} \sum_{g \in \text{Fiber}_{j,y}} V_{j,r}(g, y), \quad \bar{V}_\psi := \max_{g \in \text{Fiber}_{j,y}} V_{j,r}(g, y), \quad \underline{V}_\psi := \min_{g \in \text{Fiber}_{j,y}} V_{j,r}(g, y).$$

For compactness of notation, let us also write $V_{j,r}(x) = \sum_{p=3,4} \langle \mathbf{W}_{5,j,r,1}, e_{x_0} \rangle + \langle \mathbf{W}_{5,j,r,1}, e_{x_1} \rangle$ which is a random variable depending on the randomness of x_0, x_1 .

The definition of learning curriculum for the symmetry group actions is essentially the same with the simply transitive case, with a slight difference of using the fiber-combination φ instead of the base combination ϕ . We repeat here for reference.

Definition C.4 (Learning curriculum, symmetry group actions). We generate an feature index set Σ^\star over fiber-indexed features. Letting $\Sigma_0 = \Psi$, at each $i \in [n_y^2]$, we choose

$$\psi_i = \arg \max_{\psi=(j,r,\varphi) \in \Sigma_i, \varphi \in \Phi_j^\star} V_\psi^{(0)}.$$

Write $\psi_i = (j, r, \varphi)$, we define the next iteration set Σ_{i+1} by excluding:

(1) The confusing combinations for φ in the same neuron:

$$\Sigma_{\psi_i}^{\dagger,1} \equiv \Sigma_i^{\dagger,1} = \{(j, r, \varphi') \in \Sigma_i, \varphi' \in \Phi_\varphi^\dagger\};$$

(2) All other combinations in the same neuron (j, r) :

$$\Sigma_{\psi_i}^{\dagger,2} \equiv \Sigma_i^{\dagger,2} = \{(j, r, \varphi') \mid \varphi' \in \Phi \setminus (\Phi_\varphi^\dagger \cup \{\varphi\})\};$$

(3) The same combination $\varphi_{j,y}$ in other neurons of the same class $j, r', r' \neq r$:

$$\Sigma_{\psi_i}^{\dagger,3} \equiv \Sigma_i^{\dagger,3} = \{(j, r', \varphi_{j,y}) \in \Sigma_i : r' \neq r\}.$$

The next iteration is given by

$$\Sigma_{i+1} = \Sigma_i \setminus (\Sigma_{\psi_i}^{\dagger,1} \cup \Sigma_{\psi_i}^{\dagger,2} \cup \Sigma_{\psi_i}^{\dagger,3} \cup \{\psi_i\})$$

Eventually we obtain $\Sigma^* = (\psi_1, \dots, \psi_{n_y^2})$ and $\Sigma^\dagger := \bigcup_{\psi \in \Sigma^*} \Sigma_\psi^\dagger$. We write $\psi \prec_\Sigma \psi'$ (or simply $\psi \prec \psi'$) if ψ precedes ψ' in Σ^* . More generally, we also write $\psi \prec \psi'$ for $\psi, \psi' \in \Psi$ if $\psi \in \Sigma^*$ and $\psi' \in \bigcup_{\psi'' \succ \psi} \Sigma_{\psi''}^\dagger$ is from the .

Events of combination appearance. These events specify whether a specific base combination $\phi = (g, y)$ appears, or only one of its component matches.

Definition C.5 (Events of appearance). For $\phi = (g, y) \in \varphi_{j,y}$ for some $j \in \tau(\mathcal{Y})$ and $y \in \mathcal{Y}$, we define

$$\mathcal{H}_\phi := \{g_1 = g, y_0 = y\}, \quad \mathcal{H}_g := \{g_1 = g\}, \quad \mathcal{H}_y := \{y_0 = y\},$$

For the mismatched events \mathcal{H}^\dagger , the definition is a little different from the simply transitive actions: since $\varphi_{j,y} = \text{Fiber}_{j,y} \times \{y\}$ and $g \in \text{Fiber}_{j,y}$, we write

$$\mathcal{H}_\varphi^\dagger(g) := \{g_1 = g, y_0 \neq y\}, \quad \mathcal{H}_\varphi^\dagger(y) := \{g_1 \notin \text{Fiber}_{j,y}, y_0 = y\},$$

When considering the event of all the base pair $\phi \in \varphi_{j,y}$, we also define the union events

$$\mathcal{H}_\varphi := \bigcup_{\phi \in \varphi} \mathcal{H}_\phi, \quad \mathcal{H}_\varphi^\dagger := \mathcal{H}_\varphi^\dagger(y) \cup \left(\bigcup_{\phi=(g,y) \in \varphi} \mathcal{H}_\varphi^\dagger(g) \right).$$

C.1.1 Theorem Statement

Now we present the theorem for learning symmetric group actions.

Theorem C.1 (Learning Symmetric Group Actions). *Assuming Assumption 5.1, and let $F^{(t)}$ be obtained by Algorithm 2 at $t = T_1$, then the loss for the 5th-token is minimized, i.e., $\text{Loss}_5^{(T_1)} \leq \frac{1}{\text{poly}(d)}$, and for each $\psi = (j, r, \varphi) \in \Sigma^*$, the followings hold:*

A. For any $\phi \in \varphi$, we have $V_{j,r}^{(T_1)}(\phi) \in [B - O(d^{c_1}\mu), B + O(\mu)]$;

B. For any $\psi' = (j, r, \varphi') \in \Sigma^{\dagger,1}$ and $\phi' \in \varphi'$, we have $|V_{j,r}^{(T_1)}(\phi')| \leq \tilde{O}(\varpi^{\frac{1}{q-1}})$. Or more concretely:

$$\frac{1}{2} \left| V_{j,r}^{(T_1)}(g') + V_{j,r}^{(T_1)}(y') \right| \leq \tilde{O}(\varpi^{\frac{1}{q-1}}), \quad \forall (g', y') \in \varphi'$$

C. For any $\phi = (g, y) \in \varphi$ the following relation holds:

$$\left| V_{j,r}^{(T_1)}(g) - C_\alpha V_{j,r}^{(T_1)}(y) \right| \leq \tilde{O}(\varpi^{\frac{1}{q-1}})$$

where $C_\alpha = \frac{1+n_G(n_Y-1)/n_Y}{(n_Y-1)+n_G/n_Y} = \Theta(n_Y)$ due to our choice of n_G, n_Y in Assumption C.1.

D. For any $\psi' \in \Sigma^{\dagger,3}$, it holds that $\bar{V}_{\psi'}^{(t)} = \tilde{O}(\mu)$.

The assertions of Theorem C.1 described the *feature shapes* of the neurons, which we formally defined in the proof of learning simply transitive actions (Definition B.1). The main difference between the results here and Theorem B.1 is that Theorem C.1C described a different symmetry of features without the factor $\Theta(n_Y)$.

Let us follow the proof in Appendix B and the feature shape for symmetry group actions.

Definition C.6 (Feature shape). Let $\psi = (j, r, \varphi) \in \Psi$, and let $\delta = (\delta_1, \delta_2)$ be the error parameters. We say the feature ψ has shape $\mathcal{F}_\psi(\delta)$ if the followings are satisfied:

1. $\mathcal{F}_{\psi,1}(\delta_1)$: $\underline{V}_\psi^{(t)} \geq B - O(\delta_1)$ and $\bar{V}_\psi^{(t)} \leq B + O(\delta_1)$;
2. $\mathcal{F}_{\psi,2}(\delta_2)$: For any $\psi' = (j, r, \varphi') \in \Sigma_{\psi}^{\dagger,1}$, it holds that $|V_{j,r}^{(t)}(\phi')| \leq O(\delta_2)$ for all $\phi' \in \varphi'$;
3. $\mathcal{F}_{\psi,3}(\delta_2)$: $|V_{j,r}^{(t)}(g) - C_Y V_{j,r}^{(t)}(y)| \leq O(\delta_2)$ for any $\phi = (g, y) \in \varphi$;
4. $\mathcal{F}_{\psi,4}$: For any feature $\psi' \in \Sigma_{\psi}^{\dagger,3}$, it holds that $\bar{V}_{\psi'}^{(t)} \leq \tilde{O}(\mu)$.

Indeed, we aim to prove that $\mathcal{F}_\psi(\delta)$ holds for suitable parameter δ at the end of training, for every $\psi \in \Sigma^*$. And this is indeed what the theorem states.

Remark C.4. A major difference between Definition C.6 and Definition B.4 is that now the conditions involve not just one combination of weights, but all combinations $\phi \in \varphi$ given a combination $\varphi \in \Phi$. This creates some significant difficulties for the proof.

C.1.2 Facts and Basic Calculations

We now define the gradient notation and conditions.

Definition C.7 (gradient notation). Let $\psi = (j, r, \varphi)$, we write Γ^+ and Γ^- to denote the different terms in the gradient computation. For any base combination $\phi = (g, y) \in \varphi_{j,y}$ where $\varphi_{j,y} \in \Phi$, we write

$$\Gamma_{j,r,\phi}^{+, (t)} := \mathbb{E}[(1 - \text{logit}_{5,j}) \text{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\mathcal{H}_\phi}], \quad \Gamma_{j,r,\phi}^{-, (t)} := \mathbb{E}[\text{logit}_{5,j} \text{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\mathcal{H}_\phi}].$$

When we consider the gradient of $V_{j,r}(g)$ or $V_{j,r}(y)$ for individual tokens $g \in \mathcal{G}$ and $y \in \mathcal{Y}$, we use the following notations: let $\phi = (g, y) \in \varphi_{j,y}$, then we define

$$\begin{aligned} \Gamma_{j,r,g}^{+, (t)} &:= \mathbb{E}[\text{logit}_{5,j} \text{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\mathcal{H}_\phi}] \\ \Gamma_{j,r,g}^{-, (t)} &:= \mathbb{E}[\text{logit}_{5,j} \text{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\mathcal{H}_\phi^\dagger(g)}] \\ \Gamma_{j,r,y}^{-, (t)} &:= \mathbb{E}[\text{logit}_{5,j} \text{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\mathcal{H}_\phi^\dagger(y)}]. \end{aligned}$$

Finally, we sum up over $\phi \in \varphi$ for the gradient of combination $\psi = (j, r, \varphi)$:

$$\begin{aligned} \Gamma_\psi^{+, (t)} &:= \mathbb{E}[(1 - \text{logit}_{5,j}) \text{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\mathcal{H}_\varphi}], \\ \Gamma_\psi^{-, (t)} &:= \mathbb{E}[\text{logit}_{5,j} \text{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\mathcal{H}_\varphi^\dagger}]. \end{aligned}$$

Let $\psi = (j, r, \varphi)$ where $\varphi_1 = \text{Fiber}_{j,y}$. One can check that the defined $\Gamma_\psi^{+, (t)}$ is connected to the base pair Γ by the following relation:

$$\Gamma_\psi^{+, (t)} = \sum_{g \in \text{Fiber}_{j,y}} \Gamma_{j,r,g}^{+, (t)}, \quad \Gamma_\psi^{-, (t)} = \Gamma_{j,r,y}^{-, (t)} + \sum_{g \in \text{Fiber}_{j,y}} \Gamma_{j,r,g}^{-, (t)} \quad (37)$$

Definition C.8 (gradient condition). At $t \leq T_1$ and $\delta > 0$, write for $\psi = (j, r, \varphi)$

$$\mathcal{C}_\psi(\delta) : |\Gamma_\psi^{+, (t)} - \Gamma_\psi^{-, (t)}| \leq \delta, \quad \mathcal{C}_\psi^+(\delta) : |\Gamma_\psi^{+, (t)}| \leq \delta, \quad \mathcal{C}_\psi^-(\delta) : |\Gamma_\psi^{-, (t)}| \leq \delta.$$

These conditions control the magnitude of the gradient of the feature at iteration t . Next we compute the expressions of the gradient using

Fact C.2 (gradient expressions). Equipped with Definition C.8, the gradient with respect to $\mathbf{W}_{i,j,r,p}$ for $i = 5$, $j \in [d]$, $r \in [m]$, $p \in [5]$ and $v \in \mathcal{G} \cup \mathcal{Y}$ can be computed by

(a) when $j \in \tau(\mathcal{Y})$, $r \in [m]$, $p = 5$, for $g \in \mathcal{G}$, let $\phi = (g, y) \in \varphi_{j,y}$ be the combination with $\phi_1 = g$, we have

$$\langle -\nabla_{\mathbf{W}_{5,j,r,2}} \text{Loss}, e_g \rangle = \frac{1}{2} (\Gamma_{j,r,\phi}^{+, (t)} - \Gamma_{j,r,g}^{-, (t)}) \quad (38)$$

(b) when $j \in \tau(\mathcal{Y})$, $r \in [m]$, $p = 5$, for $y \in \mathcal{Y}$, let $\varphi_{j,y} \in \Phi_j^*$ be the combination, we have

$$\langle -\nabla_{\mathbf{W}_{5,j,r,5}} \text{Loss}, e_y \rangle = \frac{1}{2} (\Gamma_\psi^{+, (t)} - \Gamma_{j,r,y}^{-, (t)}) \quad (39)$$

(c) For any combination of p , v not included above, we have due to our assumptions of the distribution and update rule:

$$\langle -\nabla_{\mathbf{W}_{5,j,r,p}} \text{Loss}, e_v \rangle \equiv 0$$

C.1.3 Induction Hypothesis and Training Phases

We define the following intermediate time-steps:

Definition C.9 (Phase decomposition). Define timestamps for ψ using the feature:

- (a) Phase I: $t \in [0, T_\psi^1]$, where $T_\psi^1 := \min\{t \geq 0 : \underline{V}_\psi^{(t)} \geq d^{c_1} \mu\}$.
- (b) Phase II.1: $t \in (T_\psi^1, T_\psi^{2,1}]$, where $T_\psi^{2,1} := \min\{t \geq 0 : V_\psi^{(t)} \geq \frac{1}{2} \log d\}$.
- (c) Phase II.2: $t \in (T_\psi^{2,1}, T_\psi^{2,2}]$, where $T_\psi^{2,2} := \min\{t \geq 0 : \mathbb{E}[\mathcal{E}_{5,j}^{(t)} \mathbb{1}_{\mathcal{H}_\varphi}] \leq d^{-1/2}\}$.
- (d) Phase II.3: $t \in (T_\psi^{2,2}, T_\psi^2]$, where $T_\psi^2 = \min\{T_\psi^{2,3}, T_{2,1} + O(\frac{1}{\eta(d^{c_1} \mu)^{q-2}})\}$, and $T_\psi^{2,3}$ is defined by

$$T_\psi^{2,3} := \min\{t \geq 0 : \mathcal{F}_\psi(\delta_1, \delta_2) \text{ holds, where } \delta_1 = d^{c_1} \mu, \delta_2 = \tilde{O}((d\varpi/\lambda)^{\frac{1}{q-1}})\}$$

- (e) Phase III.1: $t \in (T_\psi^2, T_{1,1}]$, where $T_{1,1} = T_{\psi_{n_y}^2}$ and ψ_{n_y} is the last feature in Σ^* .
- (f) Phase III.2: $t \in (T_{1,1}, T_1]$, the end of the training where the feature shape has stabilized

We give several induction hypotheses, each characterizing different aspects of the process.

Induction C.1 (learning symmetric group actions). *Let $t \leq T_1$. The following holds:*

- (a) $\underline{V}_\psi^{(t)} + b_{i,j,r} \geq \Omega(\mu)$ for all $\psi \in \Sigma^*$.
- (b) $T_\psi^2 < T_{\psi'}^1$, if $\psi \prec \psi' \in \Sigma^*$. Thus the intervals $\{[T_\psi^1, T_\psi^2]\}_{\psi \in \Sigma^*}$ are non-overlapping.
- (c) Let $\psi' \prec \psi \in \Sigma^*$, then the feature shape $\mathcal{F}_{\psi'}(d^{c_1} \sigma_0, \varpi^{\frac{1}{q-1}})$ holds throughout $t \in [T_\psi^1, T_1]$;
- (d) For any $\psi = (j, r, \varphi) \in \Sigma^*$ There are at most $\tilde{O}(\sqrt{d}/\eta)$ iterations where $\mathbf{logit}^{(t)} \geq \frac{1}{\sqrt{d}}$ conditioned on \mathcal{H}_φ in $t \in [0, T_\psi^1]$.

Interpretations of Induction C.1. We explain what the inductions hypothesis mean here:

- (a) This property simply asserts that the *good* features never shrink below a certain level which is crucial for the optimization to work.
- (b) This property describes the *curriculum* nature of Σ^* : each feature V_ψ grows in different time intervals within $0 \leq t \leq T_1$.
- (c) This property asserts that the feature ψ' previous to ψ in the curriculum Σ^* has stable and good feature shapes, which helps with learning the current feature.
- (d) This property means there are very few iteration where the gradient of $V_{j,r}(\phi)$, $\phi \in \varphi$ is disturbed before $T_{j,r,\varphi}^1$ is reached.

C.2 Phase I: Emergence of the Feature

We first prove some properties at the beginning of phase I.

Lemma C.1 (Initialization range for features). *Under random initialization and Assumption C.1, with probability at least $1 - o(1)$, the following holds*

- (a) For any $v \in \mathcal{V}$, we have $|\langle \mathbf{W}_{5,j,r,p}^{(0)}, e_x \rangle| \leq O(\mu/\sqrt{\log d})$ for any j, r, p .
- (b) Let $\psi \in \Sigma^*$, $\psi' \in \Psi$ and $\psi \prec \psi'$, then $V_\psi^{(0)} \geq V_{\psi'}^{(0)} + \Delta_0$ for a gap $\Delta_0 = \Omega(\sigma_0/\text{polylog} d)$.
- (c) We have $\Lambda_{5,j,r}^{(0)}(\mathbf{Z}) = \Omega(\mu)$ for all $j \in [d]$, $r \in [m]$.

Proof. The proof is basically the same as that of Fact B.3. The first one uses basic concentration properties of Gaussian distribution and the second uses a union bound with the anti-concentration property of the Gaussian distribution. (c) is based on the fact that

$$\Lambda_{5,j,r}^{(0)}(\mathbf{Z}) = b_{5,j,r} + \frac{1}{2} \sum_{\mathbf{k} \in \mathcal{I}^{1,0}} \sum_{p \in [5]} \langle \mathbf{W}_{5,j,r,p}^{(0)}, \mathbf{Z}_{\mathbf{k},p} \rangle = \mu \pm O(\mu/\sqrt{\log d})$$

which concludes the proof. \square

Corollary C.1. *For any $\psi = (j, r, \varphi) \in \Sigma^*$ and $y \in \varphi^2$, we have with probability $\geq 1 - o(1)$:*

$$\left| \frac{1}{|\varphi|} \sum_{g \in \varphi^1} \mathbb{E}_x[\text{sReLU}'(\Lambda_{5,j,r}^{(0)}) \mid \mathcal{H}_{g,y}] - \mathbb{E}_x[\text{sReLU}'(V_{j,r}^{(0)}(y) + V_{j,r}(x) + b_{5,j,r})] \right| \\ \leq O\left(\frac{1}{\sqrt{n_G/n_Y}}\right) \mathbb{E}_x[\text{sReLU}'(V_{j,r}^{(0)}(y) + V_{j,r}(x) + b_{5,j,r})]$$

Proof. This is true because of lemma C.1a and Hoeffding's inequality. \square

Now we present an induction for phase I.

Induction C.2. *Let $\psi = (j, r, \varphi) \in \Sigma^*$, for $t \leq T_\psi^1$, the gradients satisfy $\Gamma_{j,r,y}^{+, (t)} \geq \Omega(\frac{n_G}{n_Y}) \Gamma_{i,j,g}^{+, (t)}$ for any $(g, y) \in \varphi$.*

Using Fact B.1 with a representative $\phi = (g, y) \in \varphi_{j,y}$ and the definition $V_{j,r}(\phi) = \frac{1}{2}(V_{j,r}(g) + V_{j,r}(y))$, a single gradient step with step size η yields

$$V_{j,r}^{(t+1)}(\phi) - V_{j,r}^{(t)}(\phi) = \frac{\eta}{2} \left(\langle -\nabla \mathbf{w}_{5,j,r,5} \text{Loss}, e_y \rangle + \langle -\nabla \mathbf{w}_{5,j,r,2} \text{Loss}, e_g \rangle \right)$$

$$= \frac{\eta}{2} \left(\Gamma_{\psi}^{+, (t)} + \Gamma_{j,r,\phi}^{+, (t)} - \Gamma_{j,r,g}^{-, (t)} - \Gamma_{j,r,y}^{-, (t)} \right) \quad (40)$$

Similar to the proof of the simply transitive actions, we define a proxy term and argue that the trajectory of V_{ψ} is similar to the proxy.

Lemma C.2 (Approximating the gradient with proxy). *Assuming Inductions C.1 and C.2 holds at t . Let $\psi = (j, r, \varphi) \in \Psi$, for every $\phi = (g, y) \in \varphi$, we define two proxy sequences $\tilde{V}_{j,r}^{(t)}(v), v \in \{g, y\}$ by the update rule:*

$$\tilde{V}_{j,r}^{(t+1)}(y) = \tilde{V}_{j,r}^{(t)}(y) + \frac{\eta}{2} \tilde{\Gamma}_{\psi}^{+, (t)}, \quad \tilde{V}_{j,r}^{(t+1)}(g) = \tilde{V}_{j,r}^{(t)}(g) + \frac{\eta}{2} \tilde{\Gamma}_{j,r,g}^{+, (t)};$$

where both $\tilde{V}_{j,r}(g)$ and $\tilde{V}_{j,r}(y)$ start from $\tilde{V}_{j,r}^{(0)}(y) := V_{j,r}^{(0)}(y)$, $\tilde{V}_{j,r}^{(0)}(g) := V_{j,r}^{(0)}(g)$. And the $\tilde{\Gamma}$ are defined by replacing the features of $V_{j,r}(v), v \in \varphi^1 \cup \{y\}$ with $\tilde{V}_{j,r}(v)$. Now for all $t \leq T_{\psi}^1$, we have

$$|\tilde{V}_{j,r}^{(t)}(\phi) - V_{j,r}^{(t)}(\phi)| \leq \tilde{O}(\mu/d^{\frac{1}{2}-3c_1}), \quad \text{where } \tilde{V}_{j,r}^{(t)}(\phi) = \tilde{V}_{j,r}(g) + \tilde{V}_{j,r}(y).$$

Proof. Fix $\psi = (j, r, \varphi_{j,y})$. (40) gives

$$|\tilde{V}_{j,r}^{(t)}(y) - V_{j,r}^{(t)}(y)| = \frac{\eta}{2} \Gamma_{j,r,y}^{-, (t)}, \quad |\tilde{V}_{j,r}^{(t)}(g) - V_{j,r}^{(t)}(g)| = \frac{\eta}{2} \Gamma_{j,r,g}^{-, (t)}$$

Since Induction C.2 holds at t , we have that $\tilde{V}_{j,r}^{(t+1)}(y) - \tilde{V}_{j,r}^{(t)}(y) \geq \Omega(\frac{n_G}{n_Y})(\tilde{V}_{j,r}^{(t+1)}(g) - \tilde{V}_{j,r}^{(t)}(g))$. Moreover, since $\sum_{g \in \varphi^1} \Gamma_{j,r,g}^{+, (t)} = \Gamma_{j,r,y}^{+, (t)}$ by Definition C.7. This, combined with Induction C.1, implies that

$$\tilde{V}_{j,r}^{(t+1)}(y) - \tilde{V}_{j,r}^{(t)}(y) \geq \sum_{g \in \varphi^1} \Gamma_{j,r,g}^{+, (t)} \geq \Omega(\frac{n_G}{n_Y})(\tilde{V}_{j,r}^{(t+1)}(g) - \tilde{V}_{j,r}^{(t)}(g)), \quad \forall g \in \varphi^1$$

The above update bound imply, for all $t \leq T_{\psi}^1$, the following property hold: $\tilde{V}_{j,r}^{(t)}(g) - \tilde{V}_{j,r}^{(0)}(g) \leq O(\frac{1}{n_G/n_Y})(\tilde{V}_{j,r}^{(t)}(y) - \tilde{V}_{j,r}^{(0)}(y))$ for all $g \in \varphi^1$. Now, by the same induction procedure in the proof of Lemma B.2, we know that the sum

$$\left| V_{j,r}^{(t)}(y) + \sum_{g \in \varphi^1} V_{j,r}^{(t)}(g) - \tilde{V}_{j,r}^{(t)}(y) + \sum_{g \in \varphi^1} \tilde{V}_{j,r}^{(t)}(g) \right| \leq O(\mu/d^{\frac{1}{2}-3c_1})$$

And then by the above relations we know

$$|\tilde{V}_{j,r}^{(t)}(y) - V_{j,r}^{(t)}(y)| \leq O(\mu/d^{\frac{1}{2}-3c_1})$$

Thus the conclusion holds. \square

Lemma C.3 (competition). *Let $\psi \in \Sigma^*$ Assuming Inductions C.1 and C.2 holds at $t \leq T_{\psi}^1$. Then at $t = T_{\psi}^1$, we have for any feature $\psi' \in \Psi$, $\psi' \succ \psi$, $\bar{V}_{\psi'}^{(t)} \leq \tilde{O}(\mu)$.*

Proof. From lemma C.2, we know that

$$\tilde{V}_{j,r}^{(t+1)}(y) = \tilde{V}_{j,r}^{(t)}(y) \geq \frac{\eta}{2} \sum_{g \in \varphi^1} \mathbb{E}_x[(\tilde{V}_{j,r}^{(t)}(y) + \tilde{V}_{j,r}^{(t)}(g) + b_{5,j,r} + V_{j,r}(x))^{q-1} \mathbb{1}_{\mathcal{H}_{g,y}}]$$

Suppose there is another $\psi' = (\varphi', y') \in \Sigma^*$ such that $\psi' \succ \psi$, then we can also write a comparison based on lemma C.1:

$$\tilde{V}_{j,r}^{(0)}(y) \geq \tilde{V}_{j,r}^{(0)}(y') + \Delta_0,$$

And therefore one can compute by absolutely bounding the average over $g \in \varphi$ using corollary C.1:

$$\begin{aligned}\Gamma_{\psi}^{+, (0)} &= \frac{1}{n_Y |\varphi^1|} \sum_{g \in \varphi^1} \mathbb{E}_x[(\tilde{V}_{j,r}^{(0)}(y) + \tilde{V}_{j,r}^{(0)}(g) + b_{5,j,r} + V_{j,r}(x))^{q-1}] \\ &\geq \frac{1}{n_Y} (1 - \frac{1}{\sqrt{n_G/n_Y}}) \mathbb{E}_x[(\tilde{V}_{j,r}^{(0)}(y) + b_{5,j,r} + V_{j,r}(x))^{q-1}]\end{aligned}$$

Similarly, for $\psi' = (j', r', \varphi')$ and $\phi' \in \varphi'$, we have

$$\begin{aligned}\Gamma_{\psi'}^{+, (0)} &= \frac{1}{|\varphi', 1|} \sum_{g \in \varphi', 1} \mathbb{E}_x[(\tilde{V}_{j',r'}^{(0)}(y') + \tilde{V}_{j',r'}^{(0)}(g) + b_{5,j',r'} + V_{j',r'}(x))^{q-1} \mid \mathcal{H}_{g,y'}] \\ &\leq (1 + \frac{1}{\sqrt{n_G/n_Y}}) \mathbb{E}_x[(\tilde{V}_{j',r'}^{(0)}(y') + b_{5,j',r'} + V_{j',r'}(x))^{q-1}]\end{aligned}$$

Now we can compare the proxies $\tilde{V}_{j,r}(y) + b_{5,j,r}$ with $\tilde{V}_{j,r}(y')$ using lemma B.12 to get that when $V_{j,r}(y) \geq d^{c_1} \mu$ we have $V_{j',r'}(y') \leq \tilde{O}(\mu)$. Now due to lemma C.2 and Induction C.2, we have the desired result. \square

Proof of Inductions C.1 and C.2 in Phase I. Fix $\psi = (j, r, \varphi) \in \Sigma^*$. We check items (a)-(d) in Induction C.1:

- Induction C.1a: This is simple by combining lemmas C.1 and C.2, where each $V_{j,r}(\phi)$ is well approximated by a monotonically growing proxy $\tilde{V}_{j,r}(\phi)$.
- Induction C.1b: This is proved later but here lemma C.3 showed that for any pair $\psi \prec \psi' \in \Sigma^*$, that $T_{\psi}^1 < T_{\psi'}^1$, which is a precondition for Induction C.1b.
- Induction C.1c-d: These two are proved later in future phases and are not violated in phase I.
- Induction C.2: We shall show that this is the case. Consider any feature $g \in \varphi^1$, its gradient is clearly bounded by $O(n_Y/n_G) \Gamma_{\psi}^{+, (t)}$ initially before the feature $V_{j,r}(g) \geq \Omega(\mu)$. In this range the difference of its growth with the update of $V_{j,r}(y)$ averaged over all $g \in \varphi^1$ is bounded by $O(n_Y/n_G) (\tilde{V}_{j,r}^{(t)}(y) + b_{5,j,r})^{q-1}$ because it grows slower than $O(n_Y/n_G) \tilde{V}_{j,r}(y)$ and thus Hoeffding's inequality with variance parameter μ would work to bound the difference between the averaged gradient and the gradient with $V_{j,r}(g)$ in expectation, and thus we obtain a bound $\Gamma_{\psi}^{+, (t)} \geq \Omega(n_G/n_Y) \Gamma_{j,r,g}^{+, (t)}$. When $V_{j,r}(g) \geq \Omega(\mu)$, the feature $V_{j,r}(y)$ has grown more than $\Omega(n_G/n_Y) \mu$ and the growth ratio still holds. This proved the induction for $t \leq T_{\psi}^1$.

\square

C.3 Phase II: Feature Growth and Cancellations

Below we present an additional induction hypothesis for phase II.

Induction C.3 (Phase II). *For all features $\psi = (j, r, \varphi) \in \Sigma^*$ and $t \in [T_{\psi}^1, T_{\psi}^2]$, then for any $g' \notin \varphi^1, y' \notin \varphi^2$, we have $V_{j,r}^{(t)}(g'), V_{j,r}^{(t)}(y') \in (-B - \tilde{O}(\mu), \tilde{O}(\mu))$.*

C.3.1 Technical Lemmas

Lemma C.4 (gradient estimation). *Assume Inductions C.1 and C.3. Let $\psi = (j, r, \varphi_{j,y}) \in \Sigma^*$, let $\phi = (g, y) \in \varphi_{j,y}$, then at $t \in [T_{\psi}^1, T_{\psi}^2]$, we have*

- (a) *For $t \in [T_{\psi}^1, T_{\psi}^{2,1}]$, we have $\Gamma_{\psi}^{(t)} \geq \Omega(\Pr(\mathcal{H}_{\varphi})) \mathbf{sReLU}'(V_{\psi}^{(t)})$;*
- (b) *If $V_{j,r}^{(t)}(\phi) \in (\frac{1}{3} \log d, \frac{2}{3} \log d)$, then $\Gamma_{j,r,\phi}^{(t)} = \Gamma_{j,r,\phi}^{+, (t)} - \Gamma_{j,r,\phi}^{-, (t)} \geq \Omega(\Pr(\mathcal{H}_{\phi}))$;*

(c) For any $g' \notin \varphi^1$ and any $y' \neq y$,

$$\begin{aligned}\Gamma_{j,r,g'}^{(t)} &= -\mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbf{1}_{\mathcal{H}_{(g',y)}}] \pm O(\varpi n_Y / n_G), \\ \Gamma_{j,r,y'}^{(t)} &= -\sum_{g \in \varphi^1} \mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbf{1}_{\mathcal{H}_{(g,y')}}] \pm O(\varpi).\end{aligned}$$

(d) Any individual component $V_{j,r}^{(t)}(v), v \in \varphi^1 \cup \varphi^2$ cannot fall below $O(\mu)$, and $\Gamma_{j,r,v}^{+, (t)} \geq 0$.

Proof. Write $\psi = (j, r, \varphi_{j,y})$.

(a) By the definition of $T_\psi^{2,1}$, we know there exist $\phi \in \varphi$ such that $V_{j,r}(\phi) \leq \frac{1}{2} \log d$. By the same arguments in lemma C.2, we can get that all $\phi \in \varphi$ satisfy $\frac{1}{2} \log d + o(1)$. Given that Induction C.1 holds at t , conditioned on \mathcal{H}_ϕ , all other neuron $(j', r') \neq (j, r)$ has activation smaller than ϱ . Therefore $F_{5,j'} \leq o(1)$, and $\mathbf{logit}_{5,j'} = o(1)$ for any $j' \neq j$ conditioned on \mathcal{H}_ϕ . So conditioned on $\mathbf{Z} \in \mathcal{H}_\phi$, when $V_{j,r}^{(t)}(\phi) \leq \frac{2}{3} \log d$, we have

$$\mathbf{logit}_{5,j} = \frac{e^{F_{5,j}(\mathbf{Z})}}{e^{F_{5,j}(\mathbf{Z})} + (1 + o(1))(d-1)} \leq O(d^{-\Omega(1)})$$

Now we can simply bound all $\Gamma_{j,r,v}^{-, (t)}$ for $v \in \varphi^1 \cup \varphi^2$ to be smaller than $\tilde{O}(\frac{1}{\sqrt{d}}) \Gamma_{j,r,v}^{+, (t)}$. Thus the proof is obtained by computing $\Gamma_{j,r,v}^{+, (t)}$ and sum to get $\Gamma_\psi^{+, (t)}$.

(b) Similar to (a), here we also notice that $\mathbf{sReLU}'(\Lambda_{5,j,r}(\mathbf{Z})) = 1$ because $\Lambda_{5,j,r}(\mathbf{Z}) \in [\varrho, B]$ is in the linear regime. Now we can see that the terms $\Gamma_{j,r,\phi}^{+, (t)} \geq \Omega(\Pr(\mathcal{H}_\phi))$ and $\Gamma_{j,r,\phi}^{-, (t)} \leq O(d^{-\Omega(1)})$ and thus we have the result.

(c) For $g' \notin \text{Fiber}_{j,y}$, let $y' \in \mathcal{Y}$ such that $g' \in \text{Fiber}_{j,y'}$, then we have

$$\begin{aligned}\Gamma_{j,r,g'}^{(t)} &= \mathbb{E}[(1 - \mathbf{logit}_{5,j}) \mathbf{sReLU}'(\Lambda_{5,j}) \mathbf{1}_{\mathcal{H}_{(g',y')}}] - \sum_{y' \neq y, y} \mathbb{E}[\mathbf{logit}_{5,j} \mathbf{sReLU}'(\Lambda_{5,j}) \mathbf{1}_{\mathcal{H}_{(g',y')}}] \\ &\quad - \mathbb{E}[\mathbf{logit}_{5,j} \mathbf{sReLU}'(\Lambda_{5,j}) \mathbf{1}_{\mathcal{H}_{(g',y)}}]\end{aligned}$$

Since $\mathcal{H}_{(g',y)}$ is confusing, the negative term dominates; other appearances contribute at most $O(\varpi n_Y / n_G)$ or $O(\mu^{q-1})$. Similar arguments hold for y' .

(d) This one is proven similarly to (a) but can adapt to any level of $V_{j,r}(\phi)$ for the corresponding $\phi \in \varphi$ for $v \in \varphi^1 \cup \varphi^2$. From Induction C.3c we know that all other feature $V_{j,r}(v') \leq \tilde{O}(\mu)$ for $v' \notin \varphi^1 \cup \varphi^2$. So for any $v \in \varphi^1 \cup \varphi^2$ such that $V_{j,r}(v) \leq \tilde{O}(\mu)$, we have that the combination $\phi = (v, v') \in \Phi \setminus \varphi$ has $V_{j,r}(\phi) < O(\mu)$ and thus have gradient

$$\Gamma_{j,r,v}^{(t)} = \Gamma_{j,r,v}^{+, (t)} - \Gamma_{j,r,v}^{-, (t)} \geq \Gamma_{j,r,v}^{+, (t)} - O(\mu^{q-1})$$

Note that $\Gamma_{j,r,v}^{+, (t)}$ is either $\geq \lambda/d$ or 0 at this phase. So we can sum up the negative gradient $-O(\mu^{q-1})$ for all iterations $t \in [T_\psi^1, T_\psi^2]$ (which is fewer than $\tilde{O}(\frac{1}{\eta(d^{c_1}\mu)^{q-2}})$) to get that the feature $V_{j,r}(v) \gg \mu$. This means $\Lambda_{5,j,r} \geq 0$ when \mathcal{H}_ϕ happens and thus $\Gamma_{j,r,v}^{+, (t)} \geq 0$ for all $t \in [T_\psi^1, T_\psi^2]$. □

Lemma C.5 (feature magnitude). *Assume Inductions C.1 and C.3 holds. For any feature $\psi = (j, r, \varphi) \in \Sigma^*$, we have*

(a) when $\mathcal{C}_\psi^+(\delta)$ holds for some $\delta \in [n_G \varpi, O(\lambda/d^{c_1})]$ at $t \geq T_\psi^{2,1}$, we have $V_{j,r}^{(t)}(\phi) \geq B - O(d^{c_1} \mu), \forall \phi \in \varphi$.

(b) when $\mathcal{C}_\psi^-(\delta)$ holds for some $\delta \in [n_G\varpi, O(\lambda/d^{1+c_1})]$ at $t \geq T_\psi^{2,1}$, we have $V_{j,r}^{(t)}(\phi) \in [-B - O(\mu), \tilde{O}((\frac{d\delta}{\lambda})^{\frac{1}{q-1}})]$ for all $\phi \in \Phi_\varphi^\dagger$.

Proof. (a) If $\mathcal{C}_\psi^+(\delta)$ holds with $\delta \leq O(\lambda/d^{c_1})$, then there are two probability: $V_{j,r}^{(t)}(\phi)$ for all $\phi \in \varphi$ is as small as $O(\delta^{\frac{1}{q-1}})$ which is not possible for δ so small because of lemma C.4a; or that $V_{j,r}^{(t)}(\phi) \geq B - O(d^{c_1}\mu)$ such that $F_{5,j} \geq B$ and thus the gradient has vanished for some events.

(b) The logic from (a) can be applied here as well. Since the positive gradient is at most $O(\varpi)$ by lemma C.4b, thus whenever $\delta \gg \varpi$, the negative gradient conditioned on $\mathcal{H}_\phi, \phi \in \Phi_\varphi^\dagger$ is bounded, which gives

$$\begin{aligned} \sum_{\phi \in \varphi} \mathbb{E}[\text{logit}_{5,j} \text{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\mathcal{H}_\phi}] &\leq O(\delta) \implies (\Lambda_{5,j,r})^{q-1} \mathbb{1}_{\mathcal{H}_\phi} \leq \delta \lambda / (d \Pr(\mathcal{H}_\phi)) \quad (\text{because Fact B.2}) \\ &\implies V_{j,r}^{(t)}(\phi) \lesssim \left(\frac{n_G n_Y \lambda \delta}{d} \right)^{\frac{1}{q-1}}, \forall \phi \in \varphi \end{aligned}$$

which proves the desired result. \square

Lemma C.6 (gradient stationarity). *Let $\psi = (j, r, \varphi) \in \Sigma^*$, then for any $\delta > \varpi^{q-2}$, the followings hold:*

(a) Define $\mathcal{B}_\delta^+ := \{t \in [T_\psi^{2,1}, T_\psi^{2,3}] \mid \mathcal{C}_\psi^+(\delta) \text{ doesn't hold}\}$ then $|\mathcal{B}_\delta^+| \leq O(\frac{\log^2 d}{\eta \delta})$;

(b) Define $\mathcal{B}_\delta^- := \{t \in [T_\psi^{2,1}, T_\psi^{2,3}] \mid \mathcal{C}_\psi^-(\delta) \text{ doesn't hold}\}$ then $|\mathcal{B}_\delta^-| \leq O(\frac{\log^2 d}{\eta \delta})$;

(c) For some $t = T_\psi^{2,3} \geq T_\psi^{2,1} + O(\frac{1}{\eta n_G \varpi})$, we have $V_{j,r}^{(t)}(\phi) \geq B - O(d^{c_1}\mu)$ for all $\phi \in \varphi$ and $|V_{j,r}^{(t)}(\phi')| \leq \tilde{O}((\frac{d\varpi}{\lambda})^{\frac{1}{q-1}})$ for all $\phi' \in \Phi_\varphi^\dagger$, moreover, this will hold until $t = T_\psi^2$

Proof. The proof is very similar in spirit to the proof of Lemma B.6. We first define the following quantity:

$$\Upsilon_\psi^{(t)} = V_{j,r}^{(t)}(y) - \sum_{g \notin \varphi^1} V_{j,r}^{(t)}(g)$$

One can compute that the update of $\Upsilon_\psi^{(t)}$ is simply

$$\eta \left(\Gamma_\psi^{+, (t)} - \sum_{\varphi' \neq \varphi \in \Phi_j^*} \sum_{\phi \in \varphi'} \Gamma_{j,r,\phi}^{+, (t)} + \sum_{\phi' \notin \varphi \cup \Phi_\varphi^\dagger} \Gamma_{j,r,\phi'}^{-, (t)} \right)$$

The middle term is bounded by $O(\varpi)$ and the last term is bounded below by $-O(\varpi)$, thus we have that the update satisfy

$$\Upsilon_\psi^{(t+1)} - \Upsilon_\psi^{(t)} \geq \eta \Gamma_\psi^{+, (t)} - O(\eta \varpi)$$

Since $\Gamma_\psi^{+, (t)} \geq 0$ For any $\delta \gg \varpi$ by lemma C.4c, we know that the above bound accumulate at $t \geq T_\psi^{2,1}$ at

$$\Upsilon_\psi^{(t)} - \Upsilon_\psi^{(T_\psi^{2,1})} \geq \sum_{s \in [T_\psi^{2,1}, t] \cup \mathcal{B}_\delta^+} \eta \delta - O(\eta \varpi)$$

So for no more than $O(\frac{\log^2 d}{\eta \delta})$ iterations we should have the that $\Upsilon^{(t)} \geq \Omega(\log^2 d)$, which is impossible in our setting. This contradiction proves the result for (a). The proof of (b) is similar to (a) and the proof of lemma B.6b. (c) is also similar to lemma B.6c by using lemma C.5a-b with parameter λ/d^{c_1} and parameter $n_G \varpi$, along with lemma C.7. We do not repeat the details here. \square

Lemma C.7 (Symmetry of features). *Let $\psi = (j, r, \varphi) \in \Sigma^*$ and $t \geq T_\psi^{2,3}$, we have that if $\mathcal{F}_{\psi,2}(\delta)$ hold for some $\delta > 0$, then*

$$|V_{j,r}^{(t)}(g) - C_\alpha V_{j,r}^{(t)}(y)| \leq O(\delta)$$

For some $C_\alpha = \frac{1+n_G(n_Y-1)/n_Y}{(n_Y-1)+n_G/n_Y}$.

Proof. For a feature $\psi = (j, r, \varphi) \in \Sigma^*$, their update can be represented as:

$$\begin{aligned} V_{j,r}^{(t)}(g) - V_{j,r}^{(0)}(g) &= U_{g,y} - \sum_{y' \neq y} R_{g,y'} \\ V_{j,r}^{(t)}(y) - V_{j,r}^{(0)}(y) &= \sum_{g \in \text{Fiber}_{j,y}} U_{g,y} - \sum_{g' \notin \text{Fiber}_{j,y}} R_{g',y} \end{aligned}$$

where $U_{g,y} = \sum_{s \leq t} \eta \Gamma_{j,r,\phi}^{+, (s)}$ and $R_{g,y'}$ is

$$R_{g,y'} = \sum_{s \leq t} \eta \mathbb{E}[\text{logit}_{5,j}^{(s)} \text{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbf{1}_{\mathcal{H}_{g',y}}]$$

Therefore:

$$\sum_{g \in \text{Fiber}_{j,y}} V_{j,r}^{(t)}(g) = \sum_{g \in \text{Fiber}_{j,y}} U_{g,y} - \sum_{g \in \text{Fiber}_{j,y}} \sum_{y' \neq y} R_{g,y'}$$

Thus it is true for any $\text{Fiber}_{j,y}$ and $y \in \mathcal{Y}$ that the following holds

$$V_{j,r}^{(t)}(y) - V_{j,r}^{(0)}(y) + \sum_{g' \notin \text{Fiber}_{j,y}} R_{g',y} = \sum_{g \in \text{Fiber}_{j,y}} (V_{j,r}^{(t)}(g) - V_{j,r}^{(0)}(g)) + \sum_{g \in \text{Fiber}_{j,y}} \sum_{y' \neq y} R_{g,y'}$$

By summing both the left and right hand side with all the $(\text{Fiber}_{j,y}, y)$ pairs, we have

$$\sum_{y \in \mathcal{Y}} (V_{j,r}^{(t)}(y) - V_{j,r}^{(0)}(y)) + \sum_{y \in \mathcal{Y}} \sum_{g' \notin \text{Fiber}_{j,y}} R_{g',y} = \sum_{g \in \mathcal{G}} (V_{j,r}^{(t)}(g) - V_{j,r}^{(0)}(g)) + \sum_{g \in \mathcal{G}} \sum_{y': \tau(g \cdot y') \neq j} R_{g,y'}$$

Since $\sum_{y \in \mathcal{Y}} \sum_{g' \notin \text{Fiber}_{j,y}} R_{g',y} = \sum_{g \in \mathcal{G}} \sum_{y': \tau(g \cdot y') \neq j} R_{g,y'}$, we have

$$\begin{aligned} \sum_{y \in \mathcal{Y}} (V_{j,r}^{(t)}(y) - V_{j,r}^{(0)}(y)) &= \sum_{g \in \mathcal{G}} (V_{j,r}^{(t)}(g) - V_{j,r}^{(0)}(g)) \\ \implies \sum_{y \in \mathcal{Y}} V_{j,r}^{(t)}(y) &= \sum_{g \in \mathcal{G}} V_{j,r}^{(t)}(g) \pm O(\mu) \end{aligned}$$

Moreover, we have assumed

$$\begin{aligned} |V_{j,r}^{(t)}(g) + V_{j,r}^{(t)}(y')| &= O(\delta), \quad \forall y' \neq y, g \in \text{Fiber}_{j,y}, \\ |V_{j,r}^{(t)}(y) + V_{j,r}^{(t)}(g')| &= O(\delta), \quad \forall g' \notin \text{Fiber}_{j,y} \end{aligned}$$

Thus

$$(1 + \frac{|\mathcal{G}|}{|\mathcal{Y}|}(|\mathcal{Y}| - 1)) V_{j,r}^{(t)}(y) = (\frac{|\mathcal{Y}|(|\mathcal{Y}| - 1)}{|\mathcal{G}|} + 1) \sum_{g \in \text{Fiber}_{j,y}} V_{j,r}^{(t)}(g) \pm O(n_G \delta)$$

Therefore choosing whatever $g \in \text{Fiber}_{j,t}$, we have

$$(1 + \frac{|\mathcal{G}|}{|\mathcal{Y}|}(|\mathcal{Y}| - 1))V_{j,r}^{(t)}(y) = (\frac{|\mathcal{Y}|^2}{|\mathcal{G}|} + 1)\frac{|\mathcal{G}|}{|\mathcal{Y}|}V_{j,r}^{(t)}(g) \pm O(n_G\delta_2) = (|\mathcal{Y}| - 1 + \frac{|\mathcal{G}|}{|\mathcal{Y}|})V_{j,r}^{(t)}(g) \pm O(n_G\delta)$$

Now by dividing the right-hand side by the factor on $V_{j,r}^{(t)}(y)$ on the LHS, we have the desired result. \square

C.3.2 Proof of Induction

Lemma C.8 (Arrival times). *Assuming Inductions C.1 and C.3 for $\psi = (j, r, \varphi) \in \Sigma^*$, we have*

$$(a) \quad T_{\psi}^{2,1} - T_{\psi}^1 \leq \tilde{O}(\frac{1}{\eta(d^{c_1}\sigma_0)^{q-2}});$$

$$(b) \quad T_{\psi}^{2,2} - T_{\psi}^{2,1} \leq O(\frac{d^{\frac{1}{2}+c_1}}{\eta});$$

$$(c) \quad T_{\psi}^{2,3} \leq \tilde{O}(\frac{1}{\eta\varpi});$$

$$(d) \quad T_{\psi}^2 \leq \tilde{O}(\frac{1}{\eta(d^{2c_1}\mu)^{q-2}}).$$

Proof. (a) By lemma C.4a, after phase I, within $\tilde{O}(1/(\eta(d^{c_1}\sigma_0)^{q-2}))$ we have $V_{j,r}^{(t)}(\phi) \geq \frac{1}{2} \log d$ for all $\phi \in \varphi$.

(b) On $[T^1, T^{2,1}]$, lemma C.6a gives the rate $O(d^{\frac{1}{2}+c_1}/\eta)$ time for the positive grad to arrive at $1 - \mathbf{logit}_{5,j} \leq \frac{1}{\sqrt{d}}$ conditioned on \mathcal{H}_{φ} .

(c) This is again provided by lemma C.6c by using $\delta = n_G\varpi$, combined with lemma C.5b for the feature shape guarantees.

(d) This is simply by the definition of T_{ψ}^2 and that $T_{\psi}^{2,3} \ll \frac{1}{\eta(d^{2c_1}\mu)^{q-2}}$. \square

Proof of Inductions C.1 and C.3. Fix $\psi = (j, r, \varphi_{j,y})$ and consider $t \in [T_{\psi}^1, T_{\psi}^2]$, then

- Induction C.1a: This is guaranteed by lemma C.4a-b.
- Induction C.1b: Because the time used is small between T_{ψ}^1 and T_{ψ}^2 by lemma C.8, we have that T_{ψ}^1 is still behind when this happens.
- Induction C.1c: This will be proven in phase III, we do not use any properties that violate this hypothesis.
- Induction C.1d: This is due to Induction C.3a-b and lemma C.6a with parameter $\delta = 1/\sqrt{d}$.
- Induction C.3: For any $\phi \notin \varphi$, we have that they are either desirable features $\phi \in \varphi_{j,y'}$ for some different $y' \geq y \in \mathcal{Y}$, or that they are the wrong features for the class j , i.e., $\phi \in \varphi_{j',y}$ for all $y \in \mathcal{Y}$. The feature $\psi' = (j, r, \varphi_{j,y'})$ in the former case was out-competed by $\psi = (j, r, \phi)$ so their growth is bounded by $\tilde{O}(\mu)$ in the first two stage and negative in the end due to the feature shape $\mathcal{F}_{\psi}(d^{c_1}\mu, (\frac{d\varpi}{\lambda})^{\frac{1}{q-1}})$ at the end as shown in lemma C.6c. Combination ϕ in the latter case will be smaller as well due to lemma C.6c.

The time bounds follow from lemma C.8. This proves the induction in phase II for any feature $\psi \in \Sigma^*$. \square

C.4 Phase III: Convergence

We present the induction hypothesis in this phase.

Induction C.4 (Phase III, final). *Let $\psi = (j, r, \varphi) \in \Sigma^*$, for $t \in [T_\psi^2, T_1]$, the following holds:*

- (a) *At $t \in [T_\psi^2, T_{1,1}]$, we have $\mathcal{F}_\psi(\delta_1, \delta_2)$ holds with $\delta_1 = d^{c_1}\mu$, $\delta_2 = \tilde{O}((\frac{d\varpi}{\lambda})^{\frac{1}{q-1}})$;*
- (b) *At $t \in [T_{1,1}, T_1]$, we have $\mathcal{F}_\psi(\delta_1, \delta_2)$ holds with $\delta_1 = d^{c_1}\mu$, $\delta_2 = \tilde{O}((\varpi)^{\frac{1}{q-1}})$.*

We obtain the shape of logits at convergence.

Lemma C.9 (Logit shape at convergence). *Assuming Induction C.4. Let $\psi_{n_y^2} \in \Sigma^*$ be the last of Σ^* , then at $t \in [T_{\psi_{n_y^2}}^{2,3}, T_1]$, the followings hold:*

- (a) *For $\phi \in \Phi_j^*$, $\text{logit}_{5,j}^{(t)} \geq 1 - O(\lambda)$ conditioned on \mathcal{H}_ϕ .*
- (b) *For $\phi \notin \Phi_j^*$, $\text{logit}_{5,j}^{(t)} \leq O(\lambda/d)$ conditioned on \mathcal{H}_ϕ .*

Proof. The proof is very similar to that of lemma B.9, we do not repeat here. \square

Lemma C.10 (Gradient bounds, phase III). *Assuming Inductions C.1 and C.4 holds, and let $\psi = (j, r, \varphi) \in \Sigma^*$, then at $t \in [T_{1,1}, T_1]$, the followings hold*

- (a) $\Gamma_\psi^{(t)} \geq \Omega(\lambda)$ when $V_\psi^{(t)} \leq B - d^{c_1}\mu$;
- (b) $|\Gamma_{j,r,v}^{-(t)}| \leq O(\frac{\lambda\varpi}{dn_y})$ for any $v \in \mathcal{G} \cup \mathcal{Y}$;
- (c) For $\phi \in \cup_{y \in \mathcal{Y}} \varphi_{j,y}$, we have

$$\Gamma_{j,r,\phi}^{+,(t)} = \begin{cases} -\Theta(\lambda\varpi/n_y^2), & \text{if } V_{j,r,\phi'}^{(t)} \in [-B + 2\mu, -\varpi] \\ [-\Theta(\lambda\varpi/n_y^2), -\Theta(\lambda\varpi/dn_y^2)] \cup \{0\}, & \text{if } V_{j,r,\phi'}^{(t)} \in [-B - 2\mu, -B + 2\mu] \end{cases}$$

Proof. The proof of (a) is simple as $\lambda \gg \varpi$ which is the upper bound of all other gradient terms. (b) is because of lemma C.9b and that Induction C.4b holds which guaranteed suitable feature shape. (c) is simply a combination of (b) and the lemma C.9a, and Induction C.4b. \square

C.4.1 Proof of Induction in Phase III and theorem C.1

Proof of Induction C.4 and theorem C.1. The proof is similar to the proof of Induction B.4, thus we leave out some details here.

- Induction C.4a: We know from Induction C.3 and lemmas C.6 and C.8 that Induction C.4a holds at $t = T_\psi^2$. The feature shape $\mathcal{F}_\psi(d^{c_1}\mu, \tilde{O}((\frac{d\varpi}{\lambda})^{\frac{1}{q-1}}))$ guaranteed a starting point for reusing the argument in lemma C.6c to guarantee that the feature stays at the same feature shape.
- Induction C.4b: After $t = T_{1,1}$, which is $t = T_{\psi_{n_y^2}}^2$ for the last $\psi \in \Sigma^*$. We shall prove that the feature shape $\mathcal{F}_\psi(d^{c_1}\mu, \tilde{O}((\varpi)^{\frac{1}{q-1}}))$ holds for every $\psi \in \Sigma^*$. Since (a) holds for all $\psi \in \Sigma^*$, we have that at $T_{\psi_{n_y^2}}^{2,3}$ that the feature shape in (a) holds for all $\psi = (j, r, \varphi) \in \Sigma^*$. So by lemma C.10b and similar argument in lemma B.6c, we can prove that the feature $|V_{j,r}^{(t)}(\phi)| \leq \tilde{O}(\varpi^{\frac{1}{q-1}})$ for any $\phi \in \Phi_\varphi^\dagger$, which is $\mathcal{F}_{\psi,2}(\tilde{O}(\varpi^{\frac{1}{q-1}}))$. $\mathcal{F}_{\psi,1}(d^{c_1}\mu)$ is guaranteed by lemma C.10a and $\mathcal{F}_{\psi,3}(\tilde{O}(\varpi^{\frac{1}{q-1}}))$ is guaranteed by $\mathcal{F}_{\psi,1}, \mathcal{F}_{\psi,2}$ and lemma C.7. $\mathcal{F}_{\psi,4}$ is simple as any feature $\psi \in \Sigma_\psi^{\dagger,3}$ lost the competition and are bounded by $\tilde{O}(\mu)$. Moreover, once each $\varphi_{j,y}, j \in \tau(\mathcal{Y}), y \in \mathcal{Y}$ is learned. the feature in $\Sigma_\psi^{\dagger,3}$ has too small gradient bounded by $O(\frac{\lambda}{d}\mu^{q-1})$ which will not have sufficient growth before $t = T_1$.

Since all feature $\psi \in \Sigma^*$ have shape $\mathcal{F}_\psi(\delta_1, \delta_2)$ for $\delta_1 = d^{c_1}\mu$ and $\delta_2 = \tilde{O}(\varpi^{\frac{1}{q-1}})$ at $t = T_1$, we have proven theorem C.1. \square

D Learning the Attention Layer: Simply Transitive Case

In this section, we consider the case where the group operations form a simply transitive group. According to Assumption 4.1, we assume that for any $y_1, y_2 \in \mathcal{Y}$, there exists a unique $g \in \mathcal{G}$ such that $g \cdot y_1 = y_2$. Without loss of generality, we let $\mathcal{Y} = \{0, 1, \dots, n_y - 1\}$, where $n_y \in [\Omega(\log \log d), \log d]$.

We focus on updating only \mathbf{Q} , while keeping \mathbf{W} fixed. Combined with the attention structure specified in Assumption 3.5, it suffices to consider the updates to the blocks $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ only. We consider the contribution to the gradient from the position $i = 5$ on task \mathcal{T}^2 ; specifically, the relevant loss function is given by $\sum_{\ell=1}^2 \text{Loss}_5^{2,\ell}$. As \mathbf{W} remains fixed in this section, we omit the superscript (t) in \mathbf{W} and in all related notations that depend solely on \mathbf{W} (e.g., $V_{j,r}$) for notational simplicity.

D.1 Gradient Computations

Notations for gradient expressions. We first introduce some notations for the gradients of the attention layer. For $1 \leq \ell \leq L$, given $\mathbf{Z}^{L,\ell-1}$ and $\mathbf{k} \in \mathcal{I}^{L,\ell-1}$, define

$$\Xi_{\ell,i,\mathbf{k}}^L(\mathbf{Z}^{L,\ell-1}) \triangleq \sum_{j \in [d]} \mathcal{E}_{i,j}(\mathbf{Z}^{L,\ell-1}) \sum_{r \in [m]} \text{sReLU}'(\Lambda_{i,j,r}(\mathbf{Z}^{L,\ell-1})) \langle \mathbf{W}_{i,j,r}, \mathbf{Z}_{\mathbf{k}} \rangle, \quad i \in [5]. \quad (41)$$

For simplicity of notation, we will henceforth omit the dependence on $\mathbf{Z}^{L,\ell-1}$ in the notation of $\Xi_{\ell,i,\mathbf{k}}^L$ when it is clear from the context.

Fact D.1 (Gradients of \mathbf{Q}). For any $p, q \in [5]$, we have

$$\begin{aligned} -\nabla_{\mathbf{Q}_{p,q}} \text{Loss}^L &= \sum_{\ell=1}^L \sum_{i \in [5]} -\nabla_{\mathbf{Q}_{p,q}} \text{Loss}_i^{L,\ell}, \quad \text{where} \\ -\nabla_{\mathbf{Q}_{p,q}} \text{Loss}_i^{L,\ell} &= \\ \mathbb{E} \left[\sum_{\mathbf{k} \in \mathcal{I}^{L,\ell-1}} \mathbf{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}} \cdot \left(\Xi_{\ell,i,\mathbf{k}}^L - \sum_{\mathbf{k}' \in \mathcal{I}^{L,\ell-1}} \mathbf{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}'} \Xi_{\ell,i,\mathbf{k}'}^L \right) \mathbf{Z}_{\text{ans},\ell-1,p} \mathbf{Z}_{\mathbf{k},q}^\top \right]. \end{aligned}$$

Lemma D.1 (Gradients of $\mathbf{Q}_{4,3}$). Given $s \in \tau(\mathcal{X})$, for the diagonal entry $[\mathbf{Q}_{4,3}]_{s,s}$ of the block $\mathbf{Q}_{4,3}$, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,1} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,0}) \sum_{r \in [m]} \text{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},1} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_0)} \right], \\ \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r \in [m]} \text{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},2} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_1)} \right]. \end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,1} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,0}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},2} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_0), s'=\tau(x_1)} \right], \\ \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},1} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_1), s'=\tau(x_0)} \right]. \end{aligned}$$

Proof. For $\ell = 1$, due to Fact D.1, the diagonal entry $[\mathbf{Q}_{4,3}]_{s,s}$ with $s \in \tau(\mathcal{X})$, the expected gradient contribution from $-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,1}$ takes the form

$$\mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \left(\Xi_{\ell,5,\text{pred},1}^2 - \sum_{\mathbf{k}' \in \mathcal{T}^{2,\ell-1}} \mathbf{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}'} \Xi_{\ell,5,\mathbf{k}'}^2 \right) \cdot \mathbb{1}_{s=\tau(x_0)} \right],$$

which is nonzero in expectation only when $s = \tau(x_0)$. Therefore, combined the definition of Ξ in (41) we have:

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,1} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \left(\Xi_{\ell,5,\text{pred},1}^2 - \sum_{\mathbf{k}' \in \mathcal{T}^{2,\ell-1}} \mathbf{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}'} \Xi_{\ell,5,\mathbf{k}'}^2 \right) \mathbb{1}_{s=\tau(x_0)} \right] \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j} \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},1} \rangle - \sum_{\mathbf{k}' \in \mathcal{T}^{2,\ell-1}} \mathbf{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}'} \langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathbf{k}'} \rangle \right) \right) \mathbb{1}_{s=\tau(x_0)} \right] \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j} \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},1} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_0)} \right]. \end{aligned}$$

Other quantities are computed similarly, and thus is omitted here. \square

Lemma D.2 (Gradients of $\mathbf{Q}_{4,4}$). *Given $s \in \tau(\mathcal{X})$, for the diagonal entry $[\mathbf{Q}_{4,4}]_{s,s}$ of the block $\mathbf{Q}_{4,4}$, we have*

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,1} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,0}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{ans},0} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_0)} \right], \\ \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{ans},1} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_1)} \right]. \end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,4}]_{s,s'}$ with $s \neq s'$, we have $[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,1}]_{s,s'} = 0$, and

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{ans},0} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_1), s'=\tau(x_0)} \right]. \end{aligned}$$

Proof. The analysis is similar to that of Lemma D.1, and we omit the details here. \square

D.2 Some Useful Bounds for Gradients

In this subsection, we establish several useful bounds on the gradients of the attention layer, leveraging the feature structure of the MLP layer learned during stage 1. These bounds will be instrumental for the subsequent analysis.

As established in Lemma B.7 and Theorem B.1, at the end of stage 1, the network exhibits the following activation properties:

- **Sparse activations:** For each $j \in \tau(\mathcal{Y})$, and any feature $(g, y) \in \mathfrak{F}_j$, there exists a unique *activated* neuron $r \in [m]$ such that, under the event $g_1 = g$ and $y_0 = y$, the following holds:

$$\begin{aligned} \Lambda_{5,j,r}^{(T_1)} &\geq B - O(\delta), \quad |V_{j,r}^{(T_1)}(g) - V_{j,r}^{(T_1)}(y)| \leq O(\delta) \\ \Lambda_{5,j,r'}^{(T_1)} &\leq O(\delta^{q-1}) \quad \text{for all } r' \neq r. \end{aligned}$$

- **Cancellation of incorrect features:** let $r \in [m]$ be the activated neuron associated with $(g, y) \in \mathfrak{F}_j$. Then for any $g' \neq g \in \mathcal{G}$ and any $y' \in \mathcal{Y}$, we have:

$$\begin{aligned} |V_{j,r}^{(T_1)}(g) + V_{j,r}^{(T_1)}(y')| &\leq O(\delta), \\ |V_{j,r}^{(T_1)}(g') + V_{j,r}^{(T_1)}(y)| &\leq O(\delta), \end{aligned}$$

In the analysis of FFN layer, we have a pair $\delta = (\delta_1, \delta_2)$; in the expressions above, some terms should use δ_1 and others δ_2 . For brevity, we write $\delta := \max\{\delta_1, \delta_2\}$ and bound all such terms by δ . Since both δ_1 and δ_2 are small, this simplification does not affect our analysis.

Notations for activated neurons. Since in the simply transitive case, the feature sets are disjoint across indices—i.e., $\mathfrak{F}_j \cap \mathfrak{F}_{j'} = \emptyset$ for all $j \neq j'$ —we denote the activated neuron corresponding to $(g, y) \in \mathfrak{F}_j$ by $r_{g,y}$. Moreover, let

$$\mathfrak{A} \triangleq \cup_{j \in \tau(\mathcal{Y})} \mathfrak{A}_j, \quad \text{where } \mathfrak{A}_j \triangleq \{r \mid \exists (g, y) \in \mathfrak{F}_j, r = r_{g,y}\}.$$

In other words, \mathfrak{A} is the set of all activated neurons across all feature sets \mathfrak{F}_j for $j \in \tau(\mathcal{Y})$. Given $\mathbf{Z}^{L,\ell-1}$, letting $\widehat{\mathcal{G}}(\mathbf{Z}^{L,\ell-1}) = \cup\{g_{\ell'}\}_{\ell'=1}^L$ be the collection of all the chosen group elements in the predicate clauses. Similarly $\widehat{\mathcal{Y}} = \cup\{y_{\ell'}\}_{\ell'=0}^{\ell-1}$. Then define $\mathfrak{A}_j(\mathbf{Z}^{L,\ell-1}) = \{r_{g,y} \mid (g, y) \in \mathfrak{F}_j \wedge (g \in \widehat{\mathcal{G}}(\mathbf{Z}^{L,\ell-1}) \vee y \in \widehat{\mathcal{Y}})\}$. For simplicity, we omit the dependence on $\mathbf{Z}^{L,\ell-1}$ in the notation of $\widehat{\mathfrak{A}}_j$ when it is clear from the context. Equipped with these notations, we can summarize the above properties in the following lemmas.

Lemma D.3 (Properties of target feature magnitude). *Given $(g, y) \in \mathfrak{F}_j$ with $j \in \tau(\mathcal{Y})$, then, the following properties hold.*

$$\frac{1}{2} (V_{j,r_{g,y}}(g) + V_{j,r_{g,y}}(y)) \geq B - O(\delta), \quad |V_{j,r_{g,y}}(g) - V_{j,r_{g,y}}(y)| \leq O(\delta), \quad (42)$$

$$|V_{j,r_{g,y}}(g) + V_{j,r_{g,y}}(y')| \leq O(\delta), V_{j,r_{g,y}}(y') < 0 \quad \text{for all } y' \neq y, \quad (43)$$

$$|V_{j,r_{g \cdot y}}(g') + V_{j,r_{g \cdot y}}(y)| \leq O(\delta), V_{j,r_{g \cdot y}}(g') < 0 \quad \text{for all } g' \neq g. \quad (44)$$

$$|V_{j,r}(g)|, |V_{j,r}(y)| \leq O(\delta) \quad \text{for all } r \notin \mathfrak{A}_j. \quad (45)$$

Lemma D.4 (Properties of irrelevant magnitude). *If $(p, v) \notin \{2\} \times \mathcal{G} \cup \{5\} \times \mathcal{Y}$, or $j \notin \tau(\mathcal{Y})$, then for any $r \in [m]$, we have*

$$|\langle \mathbf{W}_{5,j,r,p}, e_v \rangle| \leq \tilde{O}(\sigma_0). \quad (46)$$

The above lemmas give us some direct computations of the inner products between the weight matrices and input embedding vectors.

Lemma D.5. *Let $j \in \tau(\mathcal{Y})$ and $\ell \in [2]$. Then for any $r \in [m]$, the following holds:*

$$\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},\ell} \rangle = V_{j,r}(g_\ell) \pm \tilde{O}(\sigma_0), \quad (47)$$

$$\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{ans},\ell-1} \rangle = V_{j,r}(y_{\ell-1}) \pm \tilde{O}(\sigma_0). \quad (48)$$

Moreover, for $j \notin \tau(\mathcal{Y})$ and any $\mathbf{k} \in \mathcal{I}^{2,1}$ and $r \in [m]$, we have

$$|\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathbf{k}} \rangle| = \tilde{O}(\sigma_0). \quad (49)$$

Proof. By direct computations and the definition of $\mathbf{Z}_{\text{pred},\ell}$, we have

$$\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},\ell} \rangle = \langle \mathbf{W}_{5,j,r,1}, e_{x_\ell} \rangle + \langle \mathbf{W}_{5,j,r,2}, e_{g_\ell} \rangle + \langle \mathbf{W}_{5,j,r,3}, e_{x_{\ell-1}} \rangle$$

Plug in Lemma D.4 and the definition of $V_{j,r}(g_\ell)$, we obtain (47). The proof of (48) and (49) is similar, and we omit the details here. \square

Furthermore, we can establish some characterizations of the $\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1})$ quantities, which are crucial for the following analysis.

Lemma D.6 (Characterizations of Lambda). *Given $\mathbf{Z}^{2,\ell-1}$ with $\ell \in [2]$, with an attention structure $\{\text{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}}\}_{\mathbf{k} \in \mathcal{I}^{2,\ell-1}}$,*

(a) *for $j \in \tau(\mathcal{Y})$, for activated neuron $r \in \mathfrak{A}_j$, we have*

$$\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1}) = \sum_{\ell'=1}^2 \text{Attn}_{\text{ans},\ell-1 \rightarrow \text{pred},\ell'} V_{j,r}(g_{\ell'}) + \sum_{\ell'=1}^{\ell} \text{Attn}_{\text{ans},\ell-1 \rightarrow \text{ans},\ell'-1} V_{j,r}(y_{\ell'-1}) \pm \tilde{O}(\sigma_0).$$

(b) *for $j \in \tau(\mathcal{Y})$, for any non-activated neuron $r \notin \mathfrak{A}_j$ we have*

$$|\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1})| \leq O(\delta).$$

(c) *for $j \notin \tau(\mathcal{Y})$, for any $r \in [m]$, we have*

$$|\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1})| \leq \tilde{O}(\sigma_0).$$

Proof. Recall the definition of $\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1})$ in (23), we have

$$\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1}) = \sum_{\mathbf{k} \in \mathcal{I}^{2,\ell-1}} \text{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}} \cdot \langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathbf{k}} \rangle + b_{5,j,r}.$$

Thus, the first part follows directly from (47) and (48); similarly, the second part holds by plugging (45) into (47) and (48); the last part is a direct consequence of (49) and the fact $b_{i,j,r} = \sigma_0 \log d$. \square

A direct consequence of the above lemma is the following finer characterization of the activated neurons.

Lemma D.7. *Given $j \in \tau(\mathcal{Y})$, for $r \in \mathfrak{A}_j \setminus \widehat{\mathfrak{A}}_j$, we have $\mathbf{sReLU}'(\Lambda_{5,j,r}) = 0$.*

Proof. For $j \in \tau(\mathcal{Y})$, for $r \in \mathfrak{A}_j$, by Lemma D.6, we have

$$\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1}) = \sum_{\ell'=1}^2 \mathbf{Attn}_{\text{ans},\ell-1 \rightarrow \text{pred},\ell'} V_{j,r}(g_{\ell'}) + \sum_{\ell'=1}^{\ell} \mathbf{Attn}_{\text{ans},\ell-1 \rightarrow \text{ans},\ell'-1} V_{j,r}(y_{\ell'-1}) \pm \tilde{O}(\sigma_0).$$

By (45), for $r \in \mathfrak{A}_j \setminus \widehat{\mathfrak{A}}_j$, we have $V_{j,r}(g_{\ell'}), V_{j,r}(y_{\ell'-1}) \leq -B - O(\delta)$. Hence

$$\Lambda_{5,j,r} = -\Omega(B) \pm \tilde{O}(\sigma_0) \ll -\varrho,$$

which implies $\mathbf{sReLU}'(\Lambda_{5,j,r}) = 0$. □

Now we are ready to further derive the gradients of the attention layer starting from Lemmas D.1 and D.2 and the properties established above.

Lemma D.8 (Refined expression for the gradient of $\mathbf{Q}_{4,3}$). *Given $s \in \tau(\mathcal{X})$, for the diagonal entry $[\mathbf{Q}_{4,3}]_{s,s}$ of the block $\mathbf{Q}_{4,3}$, letting $j_1 = \tau(g_1(y_0))$ and $j_2 = \tau(g_2(y_1))$, we have*

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,1} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \right. \\ &\quad \left((1 - \mathbf{logit}_{5,j_1}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_1}} \mathbf{sReLU}'(\Lambda_{5,j_1,r}) \cdot (V_{j_1,r}(g_1) - \Lambda_{5,j_1,r} \pm \tilde{O}(\sigma_0)) \right) \pm \tilde{O}(\delta^q) \right) \\ &\quad - \sum_{j \neq j_1 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot (V_{j,r}(g_1) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0)) \right) \pm \tilde{O}(\delta^q) \\ &\quad \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_0)=s} \Big]; \end{aligned}$$

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \right. \\ &\quad \left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot (V_{j_2,r}(g_2) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0)) \right) \pm \tilde{O}(\delta^q) \right) \\ &\quad - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot (V_{j,r}(g_2) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0)) \right) \pm \tilde{O}(\delta^q) \\ &\quad \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_1)=s} \Big]. \end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,1} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2} \cdot \right. \\ &\quad \left((1 - \mathbf{logit}_{5,j_1}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_1}} \mathbf{sReLU}'(\Lambda_{5,j_1,r}) \cdot (V_{j_1,r}(g_2) - \Lambda_{5,j_1,r} \pm \tilde{O}(\sigma_0)) \right) \pm \tilde{O}(\delta^q) \right) \end{aligned}$$

$$\begin{aligned}
& - \sum_{j \neq j_1 \in \tau(\mathcal{V})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \left(V_{j,r}(g_2) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \\
& \pm \sum_{j \notin \tau(\mathcal{V})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \mathbf{1}_{\tau(x_0)=s, \tau(x_1)=s'} \Big];
\end{aligned}$$

$$\begin{aligned}
\left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot \right. \\
& \left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \left(V_{j_2,r}(g_1) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \right. \\
& - \sum_{j \neq j_2 \in \tau(\mathcal{V})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \left(V_{j,r}(g_1) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \\
& \left. \left. \pm \sum_{j \notin \tau(\mathcal{V})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_1)=s, \tau(x_0)=s'} \right].
\end{aligned}$$

Proof. For $\ell = 1$,

- for the diagonal entry $[\mathbf{Q}_{4,3}]_{s,s}$ with $s \in \tau(\mathcal{X})$,

$$\begin{aligned}
\left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,1} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,0}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\
& \left. \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},1} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbf{1}_{s=\tau(x_0)} \right] \\
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \left((1 - \mathbf{logit}_{5,j_1}) \cdot \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j_1,r}) \left(V_{j_1,r}(g_1) - \Lambda_{5,j_1,r} \pm \tilde{O}(\sigma_0) \right) \right. \right. \\
& - \sum_{j \neq j_1 \in \tau(\mathcal{V})} \mathbf{logit}_{5,j} \cdot \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \left(V_{j,r}(g_1) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \\
& - \sum_{j \notin \tau(\mathcal{V})} \mathbf{logit}_{5,j} \cdot \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},1} \rangle - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \Big) \mathbf{1}_{\tau(x_0)=s} \Big] \\
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \right. \\
& \left((1 - \mathbf{logit}_{5,j_1}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_1}} \mathbf{sReLU}'(\Lambda_{5,j_1,r}) \cdot \left(V_{j_1,r}(g_1) - \Lambda_{5,j_1,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \right. \\
& - \sum_{j \neq j_1 \in \tau(\mathcal{V})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \left(V_{j,r}(g_1) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \\
& \left. \left. \pm \sum_{j \notin \tau(\mathcal{V})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_0)=s} \right],
\end{aligned}$$

where the last equality follows from Lemma D.6 and Lemma D.7.

- for the non-diagonal entry $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s' \in \tau(\mathcal{X})$, the analysis is similar unless the condition that

the gradient is non-zero only when $s = \tau(x_0)$ and $s' = \tau(x_1)$. Thus, we have

$$\begin{aligned}
\left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,1} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,0}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\
&\quad \left. \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},2} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbf{1}_{s=\tau(x_0), s'=\tau(x_1)} \right] \\
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2} \cdot \right. \\
&\quad \left((1 - \mathbf{logit}_{5,j_1}) \cdot \left(\sum_{r \in \hat{\mathcal{A}}_{j_1}} \mathbf{sReLU}'(\Lambda_{5,j_1,r}) \cdot (V_{j_1,r}(g_2) - \Lambda_{5,j_1,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \right. \\
&\quad - \sum_{j \neq j_1 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \hat{\mathcal{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot (V_{j,r}(g_2) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \\
&\quad \left. \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_0)=s, \tau(x_1)=s'} \right].
\end{aligned}$$

For $\ell = 2$,

- for the diagonal entry $[\mathbf{Q}_{4,3}]_{s,s}$ with $s \in \tau(\mathcal{X})$,

$$\begin{aligned}
\left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\
&\quad \left. \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},2} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbf{1}_{s=\tau(x_1)} \right] \\
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \left((1 - \mathbf{logit}_{5,j_2}) \cdot \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) (V_{j_2,r}(g_2) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0)) \right. \right. \\
&\quad - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) (V_{j,r}(g_2) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0)) \\
&\quad - \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},2} \rangle - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \left. \right) \mathbf{1}_{\tau(x_1)=s} \right] \\
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \right. \\
&\quad \left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \hat{\mathcal{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot (V_{j_2,r}(g_2) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \right. \\
&\quad - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \hat{\mathcal{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot (V_{j,r}(g_2) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \\
&\quad \left. \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_1)=s} \right],
\end{aligned}$$

where the last equality follows from Lemma D.6 and Lemma D.7.

- for the non-diagonal entry $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s' \in \tau(\mathcal{X})$, the analysis is similar unless the condition that

the gradient is non-zero only when $s = \tau(x_0)$ and $s' = \tau(x_1)$. Thus, we have

$$\begin{aligned}
\left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\
&\quad \left. \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},1} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbf{1}_{s=\tau(x_1), s'=\tau(x_0)} \right] \\
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot \right. \\
&\quad \left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \left(V_{j_2,r}(g_1) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \right. \\
&\quad - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \left(V_{j,r}(g_1) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \\
&\quad \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_1)=s, \tau(x_0)=s'} \left. \right].
\end{aligned}$$

Therefore, we complete the proof. \square

Lemma D.9 (Refined expression for the gradient of $\mathbf{Q}_{4,4}$). *Given $s \in \tau(\mathcal{X})$, for the diagonal entry $[\mathbf{Q}_{4,4}]_{s,s}$ of the block $\mathbf{Q}_{4,4}$, letting $j_1 = \tau(g_1(y_0))$ and $j_2 = \tau(g_2(y_1))$, we have*

$$\begin{aligned}
\left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,1} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \right. \\
&\quad \left((1 - \mathbf{logit}_{5,j_1}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_1}} \mathbf{sReLU}'(\Lambda_{5,j_1,r}) \cdot \left(V_{j_1,r}(y_0) - \Lambda_{5,j_1,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \right. \\
&\quad - \sum_{j \neq j_1 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \left(V_{j,r}(y_0) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \\
&\quad \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_0)=s} \left. \right].
\end{aligned}$$

$$\begin{aligned}
\left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot \right. \\
&\quad \left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \left(V_{j_2,r}(y_1) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \right. \\
&\quad - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \left(V_{j,r}(y_1) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \\
&\quad \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_1)=s} \left. \right].
\end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,4}]_{s,s'}$ with $s \neq s'$, we have $[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,1}]_{s,s'} = 0$, and

$$\left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} \right]_{s,s'} = \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot \right.$$

$$\begin{aligned}
& \left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \left(V_{j_2,r}(y_0) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \right. \\
& - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \left(V_{j,r}(y_0) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \\
& \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbb{1}_{\tau(x_1)=s, \tau(x_0)=s'} \Big].
\end{aligned}$$

The proof is similar to Lemma D.8 and we omit it here.

Notations for gradient decompositions. We shall define some useful notations to further simplify the expressions of the gradient.

- for $\ell = 1$,

– for $[\mathbf{Q}_{4,3}]_{s,s}$ with $s \in \tau(\mathcal{X})$, we have $[-\nabla_{\mathbf{Q}_{4,3}} \mathbf{Loss}_5^{2,1}]_{s,s} = \mathcal{N}_{s,3,1,i} + \mathcal{N}_{s,3,1,ii} + \mathcal{N}_{s,3,1,iii}$, where

$$\begin{aligned}
\mathcal{N}_{s,3,1,i} = & \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot (1 - \mathbf{logit}_{5,j_1}) \cdot \right. \\
& \left. \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_1}} \mathbf{sReLU}'(\Lambda_{5,j_1,r}) \cdot \left(V_{j_1,r}(g_1) - \Lambda_{5,j_1,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_0)=s} \right], \tag{50}
\end{aligned}$$

$$\begin{aligned}
\mathcal{N}_{s,3,1,ii} = & -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \sum_{j \neq j_1 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \right. \\
& \left. \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \left(V_{j,r}(g_1) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_0)=s} \right], \tag{51}
\end{aligned}$$

$$\mathcal{N}_{s,3,1,iii} = \pm \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \mathbb{1}_{\tau(x_0)=s} \right]. \tag{52}$$

– for $[\mathbf{Q}_{4,4}]_{s,s}$ with $s \in \tau(\mathcal{X})$, we have $[-\nabla_{\mathbf{Q}_{4,4}} \mathbf{Loss}_5^{2,1}]_{s,s} = \mathcal{N}_{s,4,1,i} + \mathcal{N}_{s,4,1,ii} + \mathcal{N}_{s,4,1,iii}$, where

$$\begin{aligned}
\mathcal{N}_{s,4,1,i} = & \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot (1 - \mathbf{logit}_{5,j_1}) \cdot \right. \\
& \left. \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_1}} \mathbf{sReLU}'(\Lambda_{5,j_1,r}) \cdot \left(V_{j_1,r}(y_0) - \Lambda_{5,j_1,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_0)=s} \right], \tag{53}
\end{aligned}$$

$$\begin{aligned}
\mathcal{N}_{s,4,1,ii} = & -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \sum_{j \neq j_1 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \right. \\
& \left. \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \left(V_{j,r}(y_0) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_0)=s} \right], \tag{54}
\end{aligned}$$

$$\mathcal{N}_{s,4,1,iii} = \pm \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \mathbf{1}_{\tau(x_0)=s} \right]. \quad (55)$$

• for $\ell = 2$,

– for $[\mathbf{Q}_{4,3}]_{s,s}$ with $s \in \tau(\mathcal{X})$, we have $[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2}]_{s,s} = \mathcal{N}_{s,3,2,i} + \mathcal{N}_{s,3,2,ii} + \mathcal{N}_{s,3,2,iii}$, where

$$\begin{aligned} \mathcal{N}_{s,3,2,i} = \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot (1 - \mathbf{logit}_{5,j_2}) \cdot \right. \\ \left. \left(\sum_{r \in \hat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot (V_{j_2,r}(g_2) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\tau(x_1)=s} \right], \end{aligned} \quad (56)$$

$$\begin{aligned} \mathcal{N}_{s,3,2,ii} = -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \right. \\ \left. \left(\sum_{r \in \hat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot (V_{j,r}(g_2) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\tau(x_1)=s} \right], \end{aligned} \quad (57)$$

$$\mathcal{N}_{s,3,2,iii} = \pm \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \mathbf{1}_{\tau(x_1)=s} \right]. \quad (58)$$

– for $[\mathbf{Q}_{4,4}]_{s,s}$ with $s \in \tau(\mathcal{X})$, we have $[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2}]_{s,s} = \mathcal{N}_{s,4,2,i} + \mathcal{N}_{s,4,2,ii} + \mathcal{N}_{s,4,2,iii}$, where

$$\begin{aligned} \mathcal{N}_{s,4,2,i} = \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot (1 - \mathbf{logit}_{5,j_2}) \cdot \right. \\ \left. \left(\sum_{r \in \hat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot (V_{j_2,r}(y_1) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\tau(x_1)=s} \right], \end{aligned} \quad (59)$$

$$\begin{aligned} \mathcal{N}_{s,4,2,ii} = -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \right. \\ \left. \left(\sum_{r \in \hat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot (V_{j,r}(y_1) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\tau(x_1)=s} \right], \end{aligned} \quad (60)$$

$$\mathcal{N}_{s,4,2,iii} = \pm \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \mathbf{1}_{\tau(x_1)=s} \right]. \quad (61)$$

Probabilistic Events. We conclude this subsection by introducing several probabilistic events that will be used to simplify the characterization of activated neurons in the subsequent analysis.

$$\mathcal{E}_1 \triangleq \{g_1 \neq g_2\}, \quad (62)$$

$$\mathcal{E}_2 \triangleq \{g_{\ell_1}(y_{\ell'_1}) \neq g_{\ell_2}(y_{\ell'_2}), \text{ for any } (\ell_1, \ell'_1) \neq (\ell_2, \ell'_2), \text{ where } \ell_k \in [2], \ell'_k \in \{0, 1\}\}. \quad (63)$$

It is easy to see that \mathcal{E}_1 and \mathcal{E}_2 hold with high probability $1 - O(\frac{1}{\log d})$.

D.3 Stage 2.1: Initial Growth of Gap

At the beginning of stage 2, since \mathbf{Q} has not been trained for long, we have the attention score is still close to the uniform structure. Therefore, for $\ell = 1$, we have

$$\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1}) \approx \frac{1}{3}V_{j,r}(g_1) + \frac{1}{3}V_{j,r}(g_2) + \frac{1}{3}V_{j,r}(y_0) \pm \tilde{O}(\sigma_0).$$

If $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_1$,

- for $j = j_1$, for $r \in \hat{\mathfrak{A}}_{j_1}$, only $r_{g_1 \cdot y_0}$ is activated since $\Lambda_{5,j_1,r_{g_1 \cdot y_0}} \approx \frac{1}{3}B$ and $\Lambda_{5,j_1,r_{g_2 \cdot g_2^{-1}(j_1)}} \approx -\frac{1}{3}B \ll -\varrho$;
- for $j = j'_1 \triangleq \tau(g_2(y_0))$, only $r_{g_2 \cdot y_0}$ is activated since $\Lambda_{5,j'_1,r_{g_2 \cdot y_0}} \approx \frac{1}{3}B$ and $\Lambda_{5,j'_1,r_{g_1 \cdot g_2^{-1}(j'_1)}} \approx -\frac{1}{3}B \ll -\varrho$;
- for other $j \in \tau(\mathcal{Y})$, we have $\Lambda_{5,j,r} \leq -\frac{1}{3}B$ for all $r \in \hat{\mathfrak{A}}_j$, thus no activation.

Moreover, for $\ell = 2$, we have

$$\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1}) \approx \frac{1}{4}V_{j,r}(g_1) + \frac{1}{4}V_{j,r}(g_2) + \frac{1}{4}V_{j,r}(y_0) + \frac{1}{4}V_{j,r}(y_1) \pm \tilde{O}(\sigma_0).$$

If $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_2$,

- for $j \in \{\tau(g_\ell(y_{\ell'}))\}_{\ell \in [2], \ell' \in \{0,1\}}$, only the corresponding $r_{g_\ell \cdot y_{\ell'}}$ is activated in the smoothed regime since $|\Lambda_{5,j,r_{g_\ell \cdot y_{\ell'}}}| = O(\delta)$ and $\Lambda_{5,j,r} \approx -\frac{1}{2}B \ll -\varrho$ for $r \in \hat{\mathfrak{A}}_j \setminus \{r_{g_\ell \cdot y_{\ell'}}\}$;
- for other $j \in \tau(\mathcal{Y})$, we have $\Lambda_{5,j,r} \leq -\frac{1}{2}B$ for all $r \in \hat{\mathfrak{A}}_j$, thus no activation.

Here, activation means that the corresponding $\mathbf{sReLU}'(\Lambda_{5,j,r})$ is non-zero, which is crucial for the gradient computation. Based on the above observations, we can see that the gradient from $\ell = 2$ is relatively small since Λ is only activated in the smoothed regime. Thus, initially, the learning process is dominated by $\nabla_{\mathbf{Q}} \text{Loss}_5^{2,1}$. Moreover, if we take a closer look at the gradient from $\ell = 1$, we have

$$\begin{aligned} \mathcal{N}_{s,3,1,ii} &\approx -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \mathbf{logit}_{5,j'_1} \cdot \right. \\ &\quad \left. \left(\mathbf{sReLU}'(\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}) \cdot \left(V_{j'_1,r_{g_2 \cdot y_0}}(g_1) - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\tau(x_0)=s} \right] \geq 0, \\ \mathcal{N}_{s,4,1,ii} &\approx -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \mathbf{logit}_{5,j'_1} \cdot \right. \\ &\quad \left. \left(\mathbf{sReLU}'(\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}) \cdot \left(V_{j'_1,r_{g_2 \cdot y_0}}(y_0) - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \right] \leq 0, \end{aligned}$$

since $V_{j'_1,r_{g_2 \cdot y_0}}(y_0) \geq \Omega(B)$ while $V_{j'_1,r_{g_2 \cdot y_0}}(g_1) \leq -\Omega(B)$. Thus, $[\mathbf{Q}_{4,3}]_{s,s}$ will have a significant positive gradient while $[\mathbf{Q}_{4,4}]_{s,s}$ will have a negative counterpart. This will lead to the growth of the gap between $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$.

We formally characterize this growth behavior within this substage. At the beginning of each sub-stage, we establish an induction hypothesis that we expect to hold throughout. Subsequently, we analyze the dynamics under this hypothesis within the substage, aiming to prove its validity by the end of sub-stage. Due to the symmetry of $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$ across $s \in \tau(\mathcal{X})$, we may, without loss of generality, focus on a particular $s \in \tau(\mathcal{X})$.

Induction D.1. Given $s \in \tau(\mathcal{X})$, let $T_{2,1,s}$ denote the first time that $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$ reaches $\Omega\left(\frac{1}{\log d}\right)$. For all iterations $t < T_{2,1,s}$, we have the following holds

- (a) $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$ monotonically increases;
- (b) $|[\mathbf{Q}_{4,4}^{(t)}]_{s,s}| \leq [\mathbf{Q}_{4,3}^{(t)}]_{s,s}$ and $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} - [\mathbf{Q}_{4,4}^{(t)}]_{s,s} = \Theta([\mathbf{Q}_{4,3}^{(t)}]_{s,s})$;
- (c) for $p \in \{3, 4\}$, for $s' \in \tau(\mathcal{X}) \neq s$, $|[\mathbf{Q}_{4,p}^{(t)}]_{s,s'}| \leq O\left(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,s}}{d}\right)$.

D.3.1 Attention and Logit Preliminaries

We first introduce several properties of the attention scores and logits if Induction D.1 holds.

Lemma D.10. If Induction D.1 holds for all iterations $< t$, given input $\mathbf{Z}^{2,\ell-1}$, then we have

1. for $\ell = 1$,

- (a) $\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \in \left[\frac{1}{3}, \frac{1}{3} + O\left(\frac{1}{\log d}\right)\right]$;
- (b) $\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)}, \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \leq \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)}$;
- (c) $|\mathbf{Attn}_{\text{ans},0 \rightarrow \mathbf{k}}^{(t)} - \mathbf{Attn}_{\text{ans},0 \rightarrow \mathbf{k}'}^{(t)}| \leq O\left(\frac{1}{\log d}\right)$ for $\mathbf{k} \neq \mathbf{k}' \in \mathcal{I}^{(2,0)}$.

2. for $\ell = 2$,

- (a) $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \in \left[\frac{1}{4}, \frac{1}{4} + O\left(\frac{1}{\log d}\right)\right]$;
- (b) $\mathbf{Attn}_{\text{ans},1 \rightarrow \mathbf{k}}^{(t)} \leq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}$ for $\mathbf{k} \neq (\text{pred}, 2)$;
- (c) for $\mathbf{k} \in \{(\text{ans}, 0), (\text{pred}, 1)\}$, $\mathbf{k}' \in \{(\text{ans}, 1), (\text{pred}, 2)\}$,

$$|\mathbf{Attn}_{\text{ans},1 \rightarrow \mathbf{k}}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \mathbf{k}'}^{(t)}| \leq O\left(\frac{1}{\log d}\right);$$

- (d) $|\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)}| \leq O\left(\frac{1}{d}\right)$.

Proof. For $\ell = 1$, given $\mathbf{Z}^{2,\ell-1}$, according to Assumption 3.5, we have

$$\begin{aligned} \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} &= \frac{e^{[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0), \tau(x_0)}}}{e^{[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0), \tau(x_0)}} + e^{[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0), \tau(x_1)}} + e^{[\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_0), \tau(x_0)}}} \\ &= \frac{1}{1 + e^{[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0), \tau(x_1)} - [\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0), \tau(x_0)}} + e^{[\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_0), \tau(x_0)} - [\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0), \tau(x_0)}}}. \end{aligned}$$

Thus, by Induction D.1,

$$-O\left(\frac{1}{\log d}\right) \leq [\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0), \tau(x_1)} - [\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0), \tau(x_0)}, [\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_0), \tau(x_0)} - [\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0), \tau(x_0)} \leq 0,$$

which implies that $0 \leq \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \leq \frac{1}{3} + O\left(\frac{1}{\log d}\right)$. (b) is straightforward since

$$[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0), \tau(x_1)}, [\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_0), \tau(x_0)} \leq [\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0), \tau(x_0)}.$$

(c) is a direct consequence of (a) and (b).

For $\ell = 2$, given $\mathbf{Z}^{2,\ell-1}$, (a)- (c) are very similar to the above analysis, and then for (d), we have

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)}$$

$$\begin{aligned}
&= \frac{e^{[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_1), \tau(x_0)}} - e^{[\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_1), \tau(x_0)}}}{e^{[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_1), \tau(x_0)}} + e^{[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_1), \tau(x_1)}} + e^{[\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_1), \tau(x_0)}} + e^{[\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_1), \tau(x_1)}}} \\
&\stackrel{(i)}{\leq} O\left([\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_1), \tau(x_0)} - [\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_1), \tau(x_0)}\right) \stackrel{(ii)}{\leq} O\left(\frac{1}{d}\right),
\end{aligned}$$

where (i) is due to the fact that $|e^x - e^y| \leq O(|x - y|)$ when x, y are small, and (ii) is due to Induction [D.1](#) (b). \square

Lemma D.11. *If Induction [D.1](#) holds for all iterations $< t$, given input $\mathbf{Z}^{2, \ell-1}$, then we have*

1. for $\ell = 1$, if $\mathbf{Z}^{2, \ell-1} \in \mathcal{E}_1$, then

- (a) for $j = j_1$, $\Lambda_{5, j_1, r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_{j_1} \setminus \{r_{g_1 \cdot y_0}\}$;
- (b) for $j = j'_1 \triangleq \tau(g_2(y_0))$, $\Lambda_{5, j'_1, r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_{j'_1} \setminus \{r_{g_2 \cdot y_0}\}$;
- (c) for other $j \in \tau(\mathcal{Y})$, r is not activated for all $r \in \widehat{\mathfrak{A}}_j$, i.e., $\Lambda_{5, j, r}^{(t)} \ll -\varrho$.

2. $\ell = 2$, if $\mathbf{Z}^{2, \ell-1} \in \mathcal{E}_2$, then

- (a) for $j \in \{\tau(g_\ell(y_{\ell'}))\}_{\ell \in [2], \ell' \in \{0, 1\}}$, only the corresponding $r_{g_\ell \cdot y_{\ell'}}$ may be activated, with $|\Lambda_{5, j, r_{g_\ell \cdot y_{\ell'}}}^{(t)}| \leq O(1)$, while all other $\Lambda_{5, j, r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_\ell \cdot y_{\ell'}}\}$;
- (b) for other $j \in \tau(\mathcal{Y})$, r is not activated for all $r \in \widehat{\mathfrak{A}}_j$, i.e., $\Lambda_{5, j, r}^{(t)} \ll -\varrho$.

Proof. We only prove (a) for $\ell = 2$ since the other cases are straightforward. $j = \tau(g_\ell(y_{\ell'}))$, by Lemma [D.6](#) we have

$$\begin{aligned}
\Lambda_{5, j, r_{g_\ell \cdot y_{\ell'}}} &= \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 1}^{(t)} V_{j, r_{g_\ell \cdot y_{\ell'}}}(g_1) + \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} V_{j, r_{g_\ell \cdot y_{\ell'}}}(g_2) \\
&\quad + \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 0}^{(t)} V_{j, r_{g_\ell \cdot y_{\ell'}}}(y_0) + \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} V_{j, r_{g_\ell \cdot y_{\ell'}}}(y_1) \pm \tilde{O}(\sigma_0).
\end{aligned}$$

Notice that since $\mathbf{Z}^{2, \ell-1} \in \mathcal{E}_2$, we have two V terms are positive and two are negative with magnitude $B \pm O(\delta)$. Therefore, $|\Lambda_{5, j, r_{g_\ell \cdot y_{\ell'}}}| \leq O\left(\frac{1}{\log d}\right) \cdot B = O(1)$. \square

Lemma D.12. *If Induction [D.1](#) holds for all iterations $< t$, given input $\mathbf{Z}^{2, 0} \in \mathcal{E}_1$, then we have*

- (a) $\Lambda_{5, j_1, r_{g_1 \cdot y_0}}^{(t)} \geq \frac{1}{3}B - O(1)$;
- (b) $-O(\delta) \leq \Lambda_{5, j_1, r_{g_1 \cdot y_0}}^{(t)} - \Lambda_{5, j'_1, r_{g_2 \cdot y_0}}^{(t)} \leq O(1)$;

Proof. By Lemma [D.6](#) we have

$$\begin{aligned}
\Lambda_{5, j_1, r_{g_1 \cdot y_0}} &= \\
&\mathbf{Attn}_{\text{ans}, 0 \rightarrow \text{pred}, 1}^{(t)} V_{j_1, r_{g_1 \cdot y_0}}(g_1) + \mathbf{Attn}_{\text{ans}, 0 \rightarrow \text{pred}, 2}^{(t)} V_{j_1, r_{g_1 \cdot y_0}}(g_2) + \mathbf{Attn}_{\text{ans}, 0 \rightarrow \text{ans}, 0}^{(t)} V_{j_1, r_{g_1 \cdot y_0}}(y_0) \pm \tilde{O}(\sigma_0).
\end{aligned}$$

By Lemma [D.10](#) and the cancellation in [\(44\)](#), we have

$$\begin{aligned}
&\mathbf{Attn}_{\text{ans}, 0 \rightarrow \text{pred}, 2}^{(t)} V_{j_1, r_{g_1 \cdot y_0}}(g_2) + \mathbf{Attn}_{\text{ans}, 0 \rightarrow \text{ans}, 0}^{(t)} V_{j_1, r_{g_1 \cdot y_0}}(y_0) \\
&\geq \left(\frac{1}{3} - O\left(\frac{1}{\log d}\right)\right) \cdot O(\delta) - O\left(\frac{1}{\log d}\right) \cdot B.
\end{aligned}$$

Putting it back, and using the fact that $\mathbf{Attn}_{\text{ans}, 0 \rightarrow \text{pred}, 1}^{(t)} V_{j_1, r_{g_1 \cdot y_0}}(g_1) \geq \frac{1}{3} \cdot (B - O(\delta))$, we have

$$\Lambda_{5, j_1, r_{g_1 \cdot y_0}} \geq \frac{1}{3} \cdot (B - O(\delta)) + \left(\frac{1}{3} - O\left(\frac{1}{\log d}\right)\right) \cdot O(\delta) - O\left(\frac{1}{\log d}\right) \cdot B \pm \tilde{O}(\sigma_0)$$

$$\geq \frac{1}{3}B - O(1).$$

Moving on to (b), we have

$$\begin{aligned} \Lambda_{5,j_1,r_{g_1 \cdot y_0}} - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}} &= \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)}(V_{j_1,r_{g_1 \cdot y_0}}(g_1) - V_{j'_1,r_{g_2 \cdot y_0}}(g_1)) \pm \tilde{O}(\sigma_0) \\ &+ \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)}(V_{j_1,r_{g_1 \cdot y_0}}(g_2) - V_{j'_1,r_{g_2 \cdot y_0}}(g_2)) + \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)}(V_{j_1,r_{g_1 \cdot y_0}}(y_0) - V_{j'_1,r_{g_2 \cdot y_0}}(y_0)) \\ &\leq \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \cdot (2B + O(\delta)) - \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)} \cdot (2B - O(\delta)) + \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot O(\delta) \\ &\leq (2B - O(\delta)) \cdot O\left(\frac{1}{\log d}\right) + O(\delta) \leq O(1). \end{aligned}$$

Similarly, we have

$$\begin{aligned} \Lambda_{5,j_1,r_{g_1 \cdot y_0}} - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}} &\geq \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \cdot (2B - O(\delta)) - \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)} \cdot (2B + O(\delta)) - \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot O(\delta) \\ &\geq -\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)} \cdot O(\delta) - \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot O(\delta) \geq -O(\delta). \end{aligned}$$

□

Lemma D.13. *If Induction D.1 holds for all iterations $< t$, given input $\mathbf{Z}^{2,\ell-1}$, then we have*

1. for $\ell = 1$, if $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_1$, $\mathbf{logit}_{5,j}^{(t)} = \Omega(1)$ for $j \in \{j_1, j'_1\}$, $1 - \mathbf{logit}_{5,j_1}^{(t)} - \mathbf{logit}_{5,j'_1}^{(t)} = \frac{1}{\text{poly}d}$.
2. for $\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_2$, $\mathbf{logit}_{5,j}^{(t)} = O(\frac{1}{d})$ for all j .

Proof. • For $\ell = 1$, by Lemma D.11 and Lemma D.6, we have

$$\begin{aligned} F_{5,j_1}^{(t)}(\mathbf{Z}^{2,\ell-1}) &= \sum_{r \in [m]} \mathbf{sReLU}(\Lambda_{5,j_1,r}^{(t)}) = \Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} + \varrho(m/q - 1) = \Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} + O\left(\frac{1}{\text{polylog}d}\right), \\ F_{5,j'_1}^{(t)}(\mathbf{Z}^{2,\ell-1}) &= \sum_{r \in [m]} \mathbf{sReLU}(\Lambda_{5,j'_1,r}^{(t)}) = \Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)} + \varrho(m/q - 1), \\ F_{5,j}^{(t)}(\mathbf{Z}^{2,\ell-1}) &= \sum_{r \in [m]} \mathbf{sReLU}(\Lambda_{5,j,r}^{(t)}) \leq O\left(\frac{1}{\text{polylog}d}\right) \text{ for } j \neq j_1, j'_1 \in \tau(\mathcal{Y}), \\ F_{5,j}^{(t)}(\mathbf{Z}^{2,\ell-1}) &\leq m \cdot \tilde{O}(\sigma_0^q) \text{ for } j \notin \tau(\mathcal{Y}). \end{aligned}$$

Putting it together, we obtain

$$\begin{aligned} \mathbf{logit}_{5,j_1}^{(t)} &= \frac{1}{1 + e^{\frac{F_{5,j'_1}^{(t)} - F_{5,j_1}^{(t)}}{m}} + \left(\sum_{j \neq j_1, j'_1 \in \tau(\mathcal{Y})} e^{\frac{F_{5,j}^{(t)}}{m}} + \sum_{j \notin \tau(\mathcal{Y})} e^{\frac{F_{5,j}^{(t)}}{m}} \right) \cdot e^{-\frac{F_{5,j_1}^{(t)}}{m}}} \\ &= \frac{1}{1 + e^{\frac{\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)} - \Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)}}{m}} + \left(O(\log d) \cdot e^{O(\frac{1}{\text{polylog}d})} + O(d) \cdot e^{\tilde{O}(m\sigma_0^q)} \right) \cdot e^{-\frac{F_{5,j_1}^{(t)}}{m}}}. \end{aligned}$$

Thus, by Lemma D.12, and the fact that $B = C_B \log d$ for some sufficiently large constant $C_B > 0$, we have $\mathbf{logit}_{5,j_1}^{(t)} = \frac{1}{1 + e^{-O(\delta)} + O(1) \cdot e^{-(C_B/3-1)\log d}} = \Omega(1)$. Similarly, we have

$$\begin{aligned} \mathbf{logit}_{5,j'_1}^{(t)} &= \frac{1}{1 + e^{-\frac{\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)}}{m} + \frac{\Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)}}{m}} + \left(O(\log d) \cdot e^{O(\frac{1}{\text{polylog}d})} + O(d) \cdot e^{\tilde{O}(m\sigma_0^q)} \right) \cdot e^{-\frac{F_{5,j'_1}^{(t)}}{m}}} \\ &= \frac{1}{1 + e^{O(1)} + O(1) \cdot e^{-(C_B/3-1)\log d}} = \Omega(1). \end{aligned}$$

From the above analysis, it is easy to see that $1 - \mathbf{logit}_{5,j_1}^{(t)} - \mathbf{logit}_{5,j'_1}^{(t)} \leq O\left(\frac{1}{e^{(C_B/3-1)\log d}}\right) = O\left(\frac{1}{\text{poly}d}\right)$.

- For $\ell = 2$, by Lemma D.11 and Lemma D.6, we have

$$\begin{aligned} F_{5,j}^{(t)}(\mathbf{Z}^{2,\ell-1}) &= \sum_{r \in [m]} \mathbf{sReLU}(\Lambda_{5,j_1,r}^{(t)}) \in [\varrho m/q, O(1) + \varrho(m/q - 1)] \text{ for } j \in \{\tau(g_\ell(y_{\ell'}))\}_{\ell \in [2], \ell' \in \{0,1\}} \\ F_{5,j}^{(t)}(\mathbf{Z}^{2,\ell-1}) &= \sum_{r \in [m]} \mathbf{sReLU}(\Lambda_{5,j,r}^{(t)}) \leq O\left(\frac{1}{\text{polylog}d}\right) \text{ for } j \in \tau(\mathcal{Y}) \setminus \{\tau(g_\ell(y_{\ell'}))\}_{\ell \in [2], \ell' \in \{0,1\}} \\ F_{5,j}^{(t)}(\mathbf{Z}^{2,\ell-1}) &\leq m \cdot \tilde{O}(\sigma_0^q) \text{ for } j \notin \tau(\mathcal{Y}). \end{aligned}$$

Therefore, for any j , we have $\mathbf{logit}_j^{(t)} = O(\frac{1}{d})$ since $F_{5,j'}^{(t)} \leq O(1)$ for all j' .

□

In the following, we illustrate the activations on the non-high probability event.

Lemma D.14. *If Induction D.1 holds for all iterations $< t$, given input $\mathbf{Z}^{2,\ell-1}$, then we have*

1. for $\ell = 1$, if $\mathbf{Z}^{2,\ell-1} \notin \mathcal{E}_1$, then

- (a) for $j = j_1$, $\widehat{\mathfrak{A}}_{j_1} = \{r_{g_1 \cdot y_0}\}$, $\Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} = B \pm O(\delta)$;
- (b) for $j \neq j_1 \in \tau(\mathcal{Y})$, assuming $j = \tau(g_1(y))$, then $\Lambda_{5,j,r_{g_1 \cdot y}}^{(t)} = \frac{1}{3}B \pm O(1)$ and $\Lambda_{5,j,r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_1 \cdot y}\}$.

2. $\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \notin \mathcal{E}_2$, then

- (a) if $g_1 = g_2 \wedge y_0 \neq y_1$,
 - i. for $j = j_2$, $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = \frac{1}{2}B \pm O(1)$ and $\Lambda_{5,j_2,r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$;
 - ii. for $j = \tau(g_2(y_0))$, $\Lambda_{5,j_2,r_{g_2 \cdot y_0}}^{(t)} = \frac{1}{2}B \pm O(1)$ and $\Lambda_{5,j,r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y_0}\}$;
 - iii. for other $j \in \tau(\mathcal{Y})$, assuming $j = \tau(g_2(y))$, $|\Lambda_{5,j,r_{g_2 \cdot y}}^{(t)}| \leq O(1)$ and $\Lambda_{5,j,r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$.
- (b) if $g_1 \neq g_2 \wedge y_0 = y_1$,
 - i. for $j = j_2$, $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = \frac{1}{2}B \pm O(1)$ and $\Lambda_{5,j_2,r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$;
 - ii. for $j = \tau(g_1(y_1))$, $\Lambda_{5,j,r_{g_1 \cdot y_1}}^{(t)} = \frac{1}{2}B \pm O(1)$ and $\Lambda_{5,j,r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_1 \cdot y_1}\}$;
 - iii. for other $j \in \tau(\mathcal{Y})$, assuming $j = \tau(g(y_1))$, $|\Lambda_{5,j,r_{g \cdot y_1}}^{(t)}| \leq O(1)$ and $\Lambda_{5,j,r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g \cdot y_1}\}$.
- (c) if $g_1 = g_2 \wedge y_0 = y_1$,
 - i. for $j = j_2$, $\widehat{\mathfrak{A}}_{j_2} = \{r_{g_2 \cdot y_1}\}$, $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = B \pm O(\delta)$;
 - ii. for $j \neq j_2 \in \tau(\mathcal{Y})$, $|\Lambda_{5,j,r}^{(t)}| \leq O(1)$ for all $r \in \widehat{\mathfrak{A}}_j$.
- (d) $g_1 \neq g_2 \wedge y_0 \neq y_1 \wedge (g_1(y_0) = g_2(y_1) \vee g_2(y_0) = g_1(y_1))$, $|\Lambda_{5,j,r}^{(t)}| \leq O(1)$ for all $r \in \widehat{\mathfrak{A}}_j$.

With the characterization of activated neurons, we can derive the following logits for the non-high probability event.

Lemma D.15. *If Induction D.1 holds for all iterations $< t$, given input $\mathbf{Z}^{2,\ell-1}$, then we have*

1. for $\ell = 1$, if $\mathbf{Z}^{2,\ell-1} \notin \mathcal{E}_1$, then $1 - \mathbf{logit}_{5,j_1}^{(t)} = O\left(\frac{1}{\text{poly}d}\right)$.
2. $\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \notin \mathcal{E}_2$, then

- (a) if $g_1 = g_2 \wedge y_0 \neq y_1$, $\mathbf{logit}_{5,j}^{(t)} = \Omega(1)$ for $j \in \{j_2, \tau(g_2(y_0))\}$, $1 - \mathbf{logit}_{5,j_2}^{(t)} - \mathbf{logit}_{5,\tau(g_2(y_0))}^{(t)} = \frac{1}{\text{poly}d}$.
- (b) if $g_1 \neq g_2 \wedge y_0 = y_1$, $\mathbf{logit}_{5,j}^{(t)} = \Omega(1)$ for $j \in \{j_2, \tau(g_1(y_1))\}$, $1 - \mathbf{logit}_{5,j_2}^{(t)} - \mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} = \frac{1}{\text{poly}d}$.
- (c) if $g_1 = g_2 \wedge y_0 = y_1$, $1 - \mathbf{logit}_{5,j_2}^{(t)} = O\left(\frac{1}{\text{poly}d}\right)$.
- (d) $g_1 \neq g_2 \wedge y_0 \neq y_1 \wedge \left(g_1(y_0) = g_2(y_1) \vee g_2(y_0) = g_1(y_1)\right)$, $\mathbf{logit}_{5,j_2}^{(t)} = O\left(\frac{1}{d}\right)$ for all j .

D.3.2 Gradient Lemma

Lemma D.16. If Induction D.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$, we have

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \Theta\left(\frac{\log d}{d}\right).$$

Proof. By gradient decompositions, we have

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \sum_{\ell \in [2]} \sum_{\kappa \in \{i, ii, iii\}} \mathcal{N}_{s,3,\ell,\kappa}^{(t)}.$$

By Lemma D.13 and Lemma D.15, it is straightforward to see that $|\mathcal{N}_{s,3,1,iii}^{(t)}|, |\mathcal{N}_{s,3,2,iii}^{(t)}| = O\left(\frac{1}{\text{poly}d}\right)$, and thus we can focus on other terms.

By Lemma D.11 and Lemma D.14, we have

$$\begin{aligned} \mathcal{N}_{s,3,1,i}^{(t)} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \cdot (1 - \mathbf{logit}_{5,j_1}^{(t)}) \cdot \right. \\ &\quad \left. \left(\left(V_{j_1, r_{g_1 \cdot y_0}}(g_1) - \Lambda_{5,j_1, r_{g_1 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \\ &+ \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \cdot (1 - \mathbf{logit}_{5,j_1}^{(t)}) \cdot \right. \\ &\quad \left. \left(\left(V_{j_1, r_{g_1 \cdot y_0}}(g_1) - \Lambda_{5,j_1, r_{g_1 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1^c} \right] \\ &\stackrel{(a)}{=} \Theta\left(\frac{1}{d}\right) \cdot \Omega(1) \cdot \Theta(B) + \Theta\left(\frac{1}{d}\right) \cdot O\left(\frac{1}{\text{poly}d}\right) \cdot \Theta(B) \cdot O\left(\frac{1}{\log d}\right) \\ &= \Theta\left(\frac{\log d}{d}\right), \end{aligned}$$

where (a) follows from Lemma D.12, Lemma D.13, Lemma D.14 and Lemma D.15, and the fact that $\tau(x_0) = s$ holds with probability $\frac{1}{d}$. $\mathcal{N}_{s,3,2,i}^{(t)}$ can be upper bounded similarly.

Moving to $\mathcal{N}_{s,3,1,ii}^{(t)}$, noticing that $V_{j'_1, r_{g_2 \cdot y_0}}(g_1) = -B + O(\delta)$ on \mathcal{E}_1 , we have

$$\begin{aligned} \mathcal{N}_{s,3,1,ii}^{(t)} &= -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \cdot \mathbf{logit}_{5,j'_1}^{(t)} \cdot \right. \\ &\quad \left. \left(\left(V_{j'_1, r_{g_2 \cdot y_0}}(g_1) - \Lambda_{5,j'_1, r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \\ &- \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \cdot \sum_{y \neq y_0 \in \mathcal{Y}} \mathbf{logit}_{5,\tau(g_1(y))}^{(t)} \cdot \right. \\ &\quad \left. \left(\left(V_{j'_1, r_{g_2 \cdot y_0}}(g_1) - \Lambda_{5,j'_1, r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \end{aligned}$$

$$\begin{aligned}
& \left(\left(V_{\tau(g_1(y)), r_{g_1 \cdot y}}(g_1) - \Lambda_{5, \tau(g_1(y)), r_{g_1 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1^c} \Big] \\
&= \Theta\left(\frac{1}{d}\right) \cdot \Omega(1) \cdot \Theta(B) - \Theta\left(\frac{1}{d}\right) \cdot O\left(\frac{1}{\text{poly}d}\right) \cdot \Theta(B) \cdot O\left(\frac{1}{\log d}\right) \\
&= \Theta\left(\frac{\log d}{d}\right).
\end{aligned}$$

For $\mathcal{N}_{s,3,2,ii}^{(t)}$, we only need to control the negative gradient, since the positive part can be easily upper bounded by $O\left(\frac{\log d}{d}\right)$.

$$\begin{aligned}
\mathcal{N}_{s,3,2,ii}^{(t)} &\geq -\Theta\left(\frac{1}{d}\right) \cdot O\left(\frac{1}{d}\right) \cdot \Theta(B) - \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \sum_{y \neq y_1 \in \mathcal{Y}} \mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \cdot \right. \\
&\quad \left. \left(\left(V_{\tau(g_2(y)), r_{g_2 \cdot y}}(g_2) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s, g_1=g_2, y_0 \neq y_1\}} \right] \\
&\stackrel{(a)}{\geq} -\Theta\left(\frac{1}{d}\right) \cdot O\left(\frac{1}{d}\right) \cdot \Theta(B) - \Theta\left(\frac{1}{d}\right) \cdot O(1) \cdot \Theta(B) \cdot O\left(\frac{1}{\log d}\right) \\
&\geq -O\left(\frac{1}{d}\right),
\end{aligned}$$

where (a) is due to the fact that $\mathbf{logit}_{5,\tau(g_2(y_0))}^{(t)} = \Omega(1)$ and $\mathbf{logit}_{5,\tau(g_2(y))}^{(t)} = O\left(\frac{1}{d}\right)$ for other $y \neq y_1, y_0$. Putting everything together, we complete the proof. \square

Lemma D.17 (Negative gradient). *If Induction D.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, we have*

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \geq -O\left(\frac{1}{\log d}\right) \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s}.$$

Proof. By gradient decompositions, we have

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \sum_{\ell \in [2]} \sum_{\kappa \in \{i, ii, iii\}} \mathcal{N}_{s,4,\ell,\kappa}^{(t)}.$$

Similarly as Lemma D.16, $|\mathcal{N}_{s,4,1,iii}^{(t)}|, |\mathcal{N}_{s,4,2,iii}^{(t)}| = O\left(\frac{1}{\text{poly}d}\right)$, and thus we can focus on other terms. By Lemma D.11, Lemma D.12, and Lemma D.14, we have

$$\begin{aligned}
\mathcal{N}_{s,4,1,i}^{(t)} + \mathcal{N}_{s,4,1,ii}^{(t)} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot (1 - \mathbf{logit}_{5,j_1}^{(t)}) \cdot \right. \\
&\quad \left. \left(\left(V_{j_1, r_{g_1 \cdot y_0}}(y_0) - \Lambda_{5,j_1, r_{g_1 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \tag{64}
\end{aligned}$$

$$\begin{aligned}
&+ \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot (1 - \mathbf{logit}_{5,j_1}^{(t)}) \cdot \right. \\
&\quad \left. \left(\left(V_{j_1, r_{g_1 \cdot y_0}}(y_0) - \Lambda_{5,j_1, r_{g_1 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1^c} \right] \tag{65} \\
&- \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot \mathbf{logit}_{5,j_1'}^{(t)} \cdot \right.
\end{aligned}$$

$$\left(\left(V_{j'_1, r_{g_2 \cdot y_0}}(y_0) - \Lambda_{5, j'_1, r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \quad (66)$$

$$- \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 0 \rightarrow \text{ans}, 0}^{(t)} \cdot \sum_{g \neq g_1 \in \mathcal{G}} \mathbf{logit}_{5, \tau(g(y_0))}^{(t)} \cdot \left(\left(V_{\tau(g(y_0)), r_{g \cdot y_0}}(y_0) - \Lambda_{5, \tau(g(y_0)), r_{g \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1^c} \right]. \quad (67)$$

First, considering the event $\{\tau(x_0) = s\} \cap \mathcal{E}_1$, we have

$$\begin{aligned} & (64) + (66) \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 0 \rightarrow \text{ans}, 0}^{(t)} \cdot \left((1 - \mathbf{logit}_{5, j_1}^{(t)}) \cdot \left(V_{j_1, r_{g_1 \cdot y_1}}(y_0) - \Lambda_{5, j_1, r_{g_1 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \pm \tilde{O}(\delta^q) \right) \right. \right. \\ & \quad \left. \left. - \mathbf{logit}_{5, j'_1}^{(t)} \cdot \left(V_{j'_1, r_{g_2 \cdot y_0}}(y_0) - \Lambda_{5, j'_1, r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \pm \tilde{O}(\delta^q) \right) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \\ & \stackrel{(a)}{=} \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 0 \rightarrow \text{ans}, 0}^{(t)} \cdot \left((1 - \mathbf{logit}_{5, j_1}^{(t)}) \cdot \left(V_{j_1, r_{g_1 \cdot y_1}}(y_0) - \Lambda_{5, j_1, r_{g_1 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \pm \tilde{O}(\delta^q) \right) \right. \right. \\ & \quad \left. \left. - (1 - \mathbf{logit}_{5, j_1}^{(t)} - \frac{1}{\text{poly}d}) \cdot \left(V_{j'_1, r_{g_2 \cdot y_0}}(y_0) - \Lambda_{5, j'_1, r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \pm \tilde{O}(\delta^q) \right) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \\ & \geq -\mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 0 \rightarrow \text{ans}, 0}^{(t)} \cdot \left((1 - \mathbf{logit}_{5, j_1}^{(t)}) \cdot \right. \right. \\ & \quad \left. \left(V_{j_1, r_{g_1 \cdot y_1}}(y_0) - V_{j'_1, r_{g_2 \cdot y_0}}(y_0) + \Lambda_{5, j'_1, r_{g_2 \cdot y_0}}^{(t)} - \Lambda_{5, j_1, r_{g_1 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \pm \tilde{O}(\delta^q) \right) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \\ & \stackrel{(b)}{\geq} -\mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 0 \rightarrow \text{ans}, 0}^{(t)} \cdot \left((1 - \mathbf{logit}_{5, j_1}^{(t)}) \cdot \left(O(\delta) + O(1) \pm \tilde{O}(\sigma_0) \pm \tilde{O}(\delta^q) \right) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \\ & \geq -O\left(\frac{\mathcal{N}_{s, 3, 1, i}^{(t)}}{\log d}\right), \end{aligned}$$

where (a) follows from Lemma D.13; (b) follows from Lemma D.12 and Lemma D.3.

Notice that (65) ≥ 0 since $V_{j_1, r_{g_1 \cdot y_0}}(y_0) - \Lambda_{5, j_1, r_{g_1 \cdot y_0}}^{(t)} \geq \Omega(B)$, thus we just need to consider the possible negative gradient from (67). By Lemma D.15, we have $\sum_{g \neq g_1 \in \mathcal{G}} \mathbf{logit}_{5, \tau(g(y_0))}^{(t)} \leq O(\frac{1}{\text{poly}d})$ on $\{\tau(x_0) = s\} \cap \mathcal{E}_1^c$, and hence (67) $\ll \mathcal{N}_{s, 3, 1, i}^{(t)}$.

Moving to the gradient from $\ell = 2$, it is straightforward to see that $\mathcal{N}_{s, 4, 2, i}^{(t)}$ is non-negative, and thus we can focus on the possible negative gradient from $\mathcal{N}_{s, 4, 2, ii}^{(t)}$.

$$\begin{aligned} \mathcal{N}_{s, 4, 2, ii}^{(t)} &= -\mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5, j}^{(t)} \cdot \right. \\ & \quad \left(\sum_{r \in \hat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5, j, r}^{(t)}) \cdot \left(V_{j, r}(y_1) - \Lambda_{5, j, r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \mathcal{E}_2} \right] \\ & \quad - \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5, j}^{(t)} \cdot \right. \end{aligned}$$

$$\begin{aligned}
& \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \cdot \left(V_{j,r}(y_1) - \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\{\tau(x_1)=s\} \cap \mathcal{E}_2^c} \Big] \\
& \geq -\Theta\left(\frac{1}{d}\right) \cdot O\left(\frac{1}{d}\right) \cdot \Theta(B) - \Theta\left(\frac{1}{d}\right) \cdot O(1) \cdot \Theta(B) \cdot O\left(\frac{1}{\log d}\right) \geq -O\left(\frac{\mathcal{N}_{s,3,1,i}^{(t)}}{\log d}\right).
\end{aligned}$$

Putting everything together, and combining with the fact that $\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \Theta(\mathcal{N}_{s,3,1,i}^{(t)})$ from Lemma D.16, we complete the proof. \square

Lemma D.18 (Growth of gap). *If Induction D.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, we have*

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} + \sum_{\ell=1}^2 \left[\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \geq \Omega\left(\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \right).$$

Proof. By gradient decompositions in (50)-(61), we have

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} + \sum_{\ell=1}^2 \left[\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \sum_{\ell \in [2]} \sum_{\kappa \in \{i, ii, iii\}} \mathcal{N}_{s,3,\ell,\kappa}^{(t)} - \mathcal{N}_{s,4,\ell,\kappa}^{(t)}.$$

Due to Lemma D.13 and Lemma D.15, $|\mathcal{N}_{s,p,\ell,iii}^{(t)}| = O(\frac{1}{\text{poly}d})$ for $p \in \{3, 4\}$ and $\ell \in [2]$, we can focus on the gradient difference between $\mathbf{Q}_{4,3}^{(t)}$ and $\mathbf{Q}_{4,4}^{(t)}$ contributed by other terms.

By Lemma D.10, we have

$$\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \leq \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)}, \quad \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \leq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}.$$

Hence, for $\ell \in [2]$, we have

$$\mathcal{N}_{s,3,\ell,i}^{(t)} - \mathcal{N}_{s,4,\ell,i}^{(t)} \geq -O(\delta) \cdot O\left(\frac{1}{d}\right) \cdot \Theta(1) \geq -O(\delta/d).$$

$$\begin{aligned}
& \mathcal{N}_{s,3,1,ii}^{(t)} - \mathcal{N}_{s,4,1,ii}^{(t)} \\
& = -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \cdot \mathbf{logit}_{5,j_1'}^{(t)} \cdot \left(\left(V_{j_1',r_{g_2 \cdot y_0}}(g_1) - \Lambda_{5,j_1',r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \\
& + \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot \mathbf{logit}_{5,j_1'}^{(t)} \cdot \left(\left(V_{j_1',r_{g_2 \cdot y_0}}(y_0) - \Lambda_{5,j_1',r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \\
& - \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \cdot \sum_{y \neq y_0 \in \mathcal{Y}} \mathbf{logit}_{5,\tau(g_1(y))}^{(t)} \cdot \left(\left(V_{\tau(g_1(y)),r_{g_1 \cdot y}}(g_1) - \Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1^c} \right] \\
& + \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot \sum_{y \neq y_0 \in \mathcal{Y}} \mathbf{logit}_{5,\tau(g_1(y))}^{(t)} \cdot \left(\left(V_{\tau(g_1(y)),r_{g_1 \cdot y}}(y_0) - \Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1^c} \right]
\end{aligned}$$

$$\begin{aligned}
& \left(\left(V_{\tau(g_1(y)), r_{g_1 \cdot y}}(y_0) - \Lambda_{5, \tau(g_1(y)), r_{g_1 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1^c} \Big] \\
& \stackrel{(a)}{\geq} \Theta\left(\frac{1}{d}\right) \cdot \Omega(1) \cdot \Theta(B) - \Theta\left(\frac{1}{d}\right) \cdot O\left(\frac{1}{\text{poly}d}\right) \cdot O(B) \cdot O\left(\frac{1}{\log d}\right) \geq \Omega\left(\frac{\log d}{d}\right),
\end{aligned}$$

where (a) follows from Lemma D.3 that $V_{j'_1, r_{g_2 \cdot y_0}}(g_1) - \Lambda_{5, j'_1, r_{g_2 \cdot y_0}}^{(t)} \leq -\Omega(B)$, and $V_{j'_1, r_{g_2 \cdot y_0}}(y_0) - \Lambda_{5, j'_1, r_{g_2 \cdot y_0}}^{(t)} \geq \Omega(B)$.

$$\begin{aligned}
& \mathcal{N}_{s,3,2,ii}^{(t)} - \mathcal{N}_{s,4,2,ii}^{(t)} \\
& = -\mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)} \cdot \right. \\
& \quad \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \cdot \left(V_{j,r}(g_2) - \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \mathcal{E}_2} \Big] \\
& + \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)} \cdot \right. \\
& \quad \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \cdot \left(V_{j,r}(y_1) - \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \mathcal{E}_2} \Big] \\
& - \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)} \cdot \right. \\
& \quad \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \cdot \left(V_{j,r}(g_2) - \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \mathcal{E}_2^c} \Big] \\
& + \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)} \cdot \right. \\
& \quad \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \cdot \left(V_{j,r}(y_1) - \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \mathcal{E}_2^c} \Big] \\
& \stackrel{(a)}{\geq} -\Theta\left(\frac{1}{d}\right) \cdot O\left(\frac{1}{d}\right) \cdot \Theta(B) - \Theta\left(\frac{1}{d}\right) \cdot O(1) \cdot \Theta(B) \cdot O\left(\frac{1}{\log d}\right) \geq -O\left(\frac{1}{d}\right),
\end{aligned}$$

where (a) follows from Lemma D.11 and Lemma D.13, which together imply that on the event \mathcal{E}_2 , only a constant number of neurons are activated and $\mathbf{logit}_{5,j}^{(t)} \leq O\left(\frac{1}{d}\right)$ for all $j \neq j_2$; and from Lemma D.15, which implies that on the complement event \mathcal{E}_2^c , occurring with probability at most $O\left(\frac{1}{\log d}\right)$, there exists at most one $j \neq j_2$ such that $\mathbf{logit}_{5,j}^{(t)} \geq \Omega(1)$ while $\mathbf{logit}_{5,j}^{(t)} \leq O\left(\frac{1}{\text{poly}d}\right)$ for other $j \in \tau(\mathcal{Y})$.

Putting it all together, we finish the proof. \square

Lemma D.19. *If Induction D.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} \geq \Omega\left(\frac{\rho}{\log d}\right)$, we have*

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \geq \Omega \left(\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \right).$$

Proof. Notice that by Lemma D.17, when $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} \geq \Omega(\frac{\varrho}{\log d})$, we have $[\mathbf{Q}_{4,4}^{(t)}]_{s,s} \geq -O(\frac{\varrho}{\log^2 d})$. Hence

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \geq \Omega\left(\frac{\varrho}{\log d}\right),$$

which implies $\Lambda_{5,j_2,r_{g_2,y_1}}^{(t)}(\mathbf{Z}^{2,1}) \geq \Omega(\frac{\varrho}{\log d}) \cdot B \pm O(\delta) \geq \varrho$ already lies in the linear regime for $\mathbf{Z}^{2,1} \in \mathcal{E}_2$. Then we have

$$\begin{aligned} \mathcal{N}_{s,4,2,i}^{(t)} &\geq \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left((1 - \text{logit}_{5,j_2}^{(t)}) \cdot \right. \right. \\ &\quad \left. \left(V_{j_2,r_{g_2,y_1}} - \Lambda_{5,j_2,r_{g_2,y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \mathcal{E}_2} \right] \\ &\geq \Omega(1) \cdot \Theta(B) \cdot \Theta\left(\frac{1}{d}\right) \geq \Omega\left(\frac{\log d}{d}\right). \end{aligned}$$

Moreover, from Lemma D.17, the magnitude of negative gradient from other \mathcal{N} terms can be upper bounded by $O\left(\frac{\mathcal{N}_{s,3,1,i}^{(t)}}{\log d}\right)$. Therefore, combining with the fact that $\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \Theta(\mathcal{N}_{s,3,1,i}^{(t)})$ from Lemma D.16, we complete the proof. \square

Lemma D.20. *If Induction D.1 holds for all iterations $< t$, given $s \neq s' \in \tau(\mathcal{X})$, for $p \in \{3, 4\}$, we have*

$$\left| \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s'} \right| \leq O\left(\frac{\log d}{d^2}\right) = O\left(\frac{1}{d}\right) \cdot \left| \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \right|.$$

Proof. The proof follows directly by combining the expressions from Lemma D.8, Lemma D.9, and Lemma D.16, along with the fact that the event $\{\tau(x_0) = s, \tau(x_1) = s'\}$ occurs with probability $O(\frac{1}{d^2})$. \square

D.3.3 At the End of Stage 2.1

Putting gradient lemmas together, we can directly prove that Induction D.1 holds for all iterations t until the end of stage 2.1, where we can conclude the following:

Lemma D.21 (End of stage 2.1). *Given $s \in \tau(\mathcal{X})$, Induction D.1 holds for all iterations $t < T_{2,1,s} = O\left(\frac{d}{\eta \log^2 d}\right)$, then at the end of stage 2.1, we have*

- (a) $[\mathbf{Q}_{4,p}^{(t)}]_{s,s} = \Omega(\frac{1}{\log d})$ for $p \in \{3, 4\}$;
- (b) $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} - [\mathbf{Q}_{4,4}^{(t)}]_{s,s} \geq \Omega\left(\frac{1}{\log d}\right)$;
- (c) $\left| [\mathbf{Q}_{4,p}^{(t)}]_{s,s'} \right| \leq O\left(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,s}}{d}\right)$ for $s' \in \tau(\mathcal{X}) \neq s$ for $p \in \{3, 4\}$; otherwise, $[\mathbf{Q}_{4,p}^{(t)}]_{s,s'} = 0$.

D.4 Stage 2.2: Continual Growth of Diagonal Entries

In Stage 2.2, the diagonal entries $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$ continue to grow until they reach a certain threshold. The analysis in this stage parallels that of Stage 2.1, but our focus now shifts to the gradients contributed by $\ell = 2$, as the logit at $\ell = 1$ is already near-optimal and thus contributes negligibly to the growth of $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$.

Induction D.2. *Given $s \in \tau(\mathcal{X})$, let $T_{2,2,s}$ denote the first time that $[\mathbf{Q}_{4,3}]_{s,s}$ reaches 0.0001. For all iterations $T_{2,1,s} \leq t < T_{2,2,s}$, we have the following holds*

- (a) $[\mathbf{Q}_{4,3}]_{s,s}, [\mathbf{Q}_{4,4}]_{s,s} \leq O(1)$ monotonically increases;

$$(b) [\mathbf{Q}_{4,3}^{(t)}]_{s,s} - [\mathbf{Q}_{4,4}^{(t)}]_{s,s} \in \left[\Omega\left(\frac{1}{\log d}\right), O(1) \right];$$

$$(c) \text{ for } (p, q) \in \{(4, 3), (4, 4)\}, \left| [\mathbf{Q}_{p,q}^{(t)}]_{s,s'} \right| \leq O\left(\frac{[\mathbf{Q}_{p,q}^{(t)}]_{s,s}}{d}\right) \text{ for } s' \in \tau(\mathcal{X}) \neq s; \text{ other } [\mathbf{Q}_{p,q}^{(t)}]_{s,s'} = 0.$$

Throughout the following analysis, instead of \mathcal{E}_2 defined in (63), we consider a renewed event $\tilde{\mathcal{E}}_2$ for $\ell = 2$:

$$\tilde{\mathcal{E}}_2 \triangleq \{g_1 \neq g_2 \wedge y_0 \neq y_1\}. \quad (68)$$

D.4.1 Attention and Logit Preliminaries

The proof in this part proceeds analogously to the arguments in Appendix D.3.1, with the induction hypothesis from Induction D.2 incorporated. Hence, we omit the details here.

Lemma D.22. *If Induction D.2 holds for all iterations $\in [T_{2,1,s}, t)$, given input $\mathbf{Z}^{2,\ell-1}$, then we have*

1. for $\ell = 1$,

$$(a) \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \in \left[\frac{1}{3} + \Omega\left(\frac{1}{\log d}\right), \frac{1}{3} + c_1 \right], \text{ where } c_1 > 0 \text{ is a small constant};$$

$$(b) \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)} \in \left[\frac{1}{3} - c_2, \frac{1}{3} - \Omega\left(\frac{1}{\log d}\right) \right], \text{ where } c_2 > 0 \text{ is a small constant};$$

$$(c) \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)} + \Omega\left(\frac{1}{\log d}\right) \leq \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)};$$

$$(d) \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \in \left[\Omega\left(\frac{1}{\log d}\right), c_3 \right].$$

2. for $\ell = 2$,

$$(a) \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \in \left[\frac{1}{4} + \Omega\left(\frac{1}{\log d}\right), \frac{1}{4} + c_4 \right], \text{ where } c_4 > 0 \text{ is a small constant};$$

$$(b) \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)}, \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \in \left[\frac{1}{4} - c_5, \frac{1}{4} - \Omega\left(\frac{1}{\log d}\right) \right], \text{ moreover, } |\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)}| \leq O\left(\frac{1}{d}\right), \text{ where } c_5 > 0 \text{ is a small constant};$$

$$(c) \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)}, \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \leq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \Omega\left(\frac{1}{\log d}\right);$$

$$(d) \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \in \left[\Omega\left(\frac{1}{\log d}\right), c_6 \right], \text{ where } c_6 > 0 \text{ is a small constant.}$$

Notice that the constant $c_1 - c_6$ depends on the threshold 0.0001 in Induction D.2. We choose the threshold 0.0001 small enough to ensure $2C_B(c_1 + c_2) < 1 - c_6C_B$ and $1 - 4c_5C_B > 0$.

Lemma D.23. *If Induction D.2 holds for all iterations $\in [T_{2,1,s}, t)$, given input $\mathbf{Z}^{2,\ell-1}$, then we have*

1. for $\ell = 1$, if $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_1$, then

$$(a) \text{ for } j = j_1, \Lambda_{5,j_1,r}^{(t)} \ll -\varrho \text{ for } r \in \hat{\mathfrak{A}}_{j_1} \setminus \{r_{g_1 \cdot y_0}\};$$

$$(b) \text{ for } j = j'_1 \triangleq \tau(g_2(y_0)), \Lambda_{5,j'_1,r}^{(t)} \ll -\varrho \text{ for } r \in \hat{\mathfrak{A}}_{j'_1} \setminus \{r_{g_2 \cdot y_0}\};$$

$$(c) \text{ for other } j \in \tau(\mathcal{Y}), r \text{ is not activated for all } r \in \hat{\mathfrak{A}}_j, \text{ i.e., } \Lambda_{5,j,r}^{(t)} \ll -\varrho.$$

2. $\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \in \tilde{\mathcal{E}}_2$, then

$$(a) \text{ for } j = j_2, \Lambda_{5,j_2,r}^{(t)} \ll -\varrho \text{ for } r \in \hat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\};$$

$$(b) \text{ for } j = j'_2 \triangleq \tau(g_2(y_0)), \Lambda_{5,j'_2,r}^{(t)} \ll -\varrho \text{ for } r \in \hat{\mathfrak{A}}_{j'_2} \setminus \{r_{g_2 \cdot y_0}\};$$

$$(c) \text{ for other } j \in \tau(\mathcal{Y}), r \text{ is not activated for all } r \in \hat{\mathfrak{A}}_j, \text{ i.e., } \Lambda_{5,j,r}^{(t)} \ll -\varrho.$$

Lemma D.24. If Induction D.2 holds for all iterations $\in [T_{2,1,s}, t)$, given input $\mathbf{Z}^{2,\ell}$, then we have

1. $\ell = 1$, for $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_1$,

$$(a) \Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} \geq \left(\frac{1}{3} + \Omega\left(\frac{1}{\log d}\right)\right)B;$$

$$(b) \Omega(1) \leq \Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)} \leq 2(c_1 + c_2)B.$$

2. $\ell = 2$, for $\mathbf{Z}^{2,\ell-1} \in \tilde{\mathcal{E}}_2$,

$$(a) \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \in [\Omega(1), 4c_5B];$$

$$(b) \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} \in [\Omega(1), c_6B] \text{ and } \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} \geq \Omega(1).$$

Lemma D.25. If Induction D.2 holds for all iterations $\in [T_{2,1,s}, t)$, given input $\mathbf{Z}^{2,\ell-1}$, then we have

1. for $\ell = 1$, if $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_1$, $\mathbf{logit}_{5,j'_1}^{(t)} \geq \Omega\left(\frac{1}{d^{2C_B(c_1+c_2)}}\right)$;

2. for $\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \in \tilde{\mathcal{E}}_2$, $1 - \mathbf{logit}_{5,j_2}^{(t)} = \Omega(1)$, $\mathbf{logit}_{5,j'_2}^{(t)} = O\left(\frac{1}{d^{1-c_6C_B}}\right)$.

Proof. • For $\ell = 1$, we have

$$\begin{aligned} \mathbf{logit}_{5,j'_1}^{(t)} &= \frac{1}{1 + e^{\Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)}} + O(d) \cdot e^{-\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)}}} \\ &\stackrel{(a)}{\geq} \frac{1}{1 + e^{2(c_1+c_2)B} + O(d) \cdot e^{-\left(\frac{1}{3}-2c_1-2c_2\right)B}} \geq \Omega\left(\frac{1}{d^{2C_B(c_1+c_2)}}\right), \end{aligned}$$

where the inequality (a) follows from Lemma D.24 and the last inequality is due to the fact that $(c_1 + c_2)$ is some sufficiently small constant s.t., $e^{-\left(\frac{1}{3}-2c_1-2c_2\right)B} = 1/\text{poly}d$.

• For $\ell = 2$, we have

$$\begin{aligned} \mathbf{logit}_{j_2}^{(t)} &= \frac{1}{1 + e^{-\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} + \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}} + O(d) \cdot e^{-\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}}} \\ &\stackrel{(a)}{\leq} \frac{1}{1 + O(d) \cdot e^{-4c_5B}} = O\left(\frac{1}{d^{1-4c_5C_B}}\right), \end{aligned}$$

where the inequality (a) follows from Lemma D.24. Similarly, we have

$$\begin{aligned} \mathbf{logit}_{j'_2}^{(t)} &= \frac{1}{1 + e^{\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}} + O(d) \cdot e^{-\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}}} \\ &\leq \frac{1}{1 + e^{\Omega(1)} + O(d) \cdot e^{-c_6B}} = O\left(\frac{1}{d^{1-c_6C_B}}\right). \end{aligned}$$

□

Lemma D.26. If Induction D.2 holds for all iterations $\in [T_{2,1,s}, t)$, given input $\mathbf{Z}^{2,\ell-1}$, then we have

1. for $\ell = 1$, if $\mathbf{Z}^{2,\ell-1} \notin \mathcal{E}_1$, then

$$(a) \text{ for } j = j_1, \hat{\mathfrak{A}}_{j_1} = \{r_{g_1 \cdot y_0}\}, \Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} = B \pm O(\delta);$$

$$(b) \text{ for } j \neq j_1 \in \tau(\mathcal{Y}), \text{ assuming } j = \tau(g_1(y)), \text{ then } \Lambda_{5,j,r_{g_1 \cdot y}}^{(t)} \in \left[\left(\frac{1}{3} - c_2\right)B, \left(\frac{1}{3} + c_1\right)B\right] \text{ and}$$

$$\Lambda_{5,j,r}^{(t)} \ll -\varrho \text{ for } r \in \hat{\mathfrak{A}}_j \setminus \{r_{g_1 \cdot y}\}.$$

2. $\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \notin \tilde{\mathcal{E}}_2$, then

- (a) if $g_1 = g_2 \wedge y_0 \neq y_1$,
 - i. for $j = j_2$, $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \in \left[\frac{1}{2}B + \Omega(1), \left(\frac{1}{2} + 2c_5\right)B\right]$ and $\Lambda_{5,j_2,r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$;
 - ii. for $j = \tau(g_2(y_0))$, $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} - \Lambda_{5,j_2,r_{g_2 \cdot y_0}}^{(t)} \geq \Omega(1)$ and $\Lambda_{5,j,r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y_0}\}$;
 - iii. for other $j \in \tau(\mathcal{Y})$, assuming $j = \tau(g_2(y))$, $\Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} \in [\Omega(1), c_6 B]$ and $\Lambda_{5,j,r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$.
- (b) if $g_1 \neq g_2 \wedge y_0 = y_1$,
 - i. for $j = j_2$, $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \in \left[\frac{1}{2}B + \Omega(1), \left(\frac{1}{2} + 2c_5\right)B\right]$ and $\Lambda_{5,j_2,r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$;
 - ii. for $j = \tau(g_1(y_1))$, $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} - \Lambda_{5,j_2,r_{g_1 \cdot y_1}}^{(t)} \geq \Omega(1)$ and $\Lambda_{5,j,r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_1 \cdot y_1}\}$;
 - iii. for other $j \in \tau(\mathcal{Y})$, r is not activated for all $r \in \widehat{\mathfrak{A}}_j$, i.e., $\Lambda_{5,j,r}^{(t)} \ll -\varrho$.
- (c) if $g_1 = g_2 \wedge y_0 = y_1$,
 - i. for $j = j_2$, $\widehat{\mathfrak{A}}_{j_2} = \{r_{g_2 \cdot y_1}\}$, $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = B \pm O(\delta)$;
 - ii. for other $j \in \tau(\mathcal{Y})$, assuming $j = \tau(g_2(y))$, $\Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} \in [\Omega(1), c_6 B]$ and $\Lambda_{5,j,r}^{(t)} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$.

Lemma D.27. If Induction D.2 holds for all iterations $\in [T_{2,1,s}, t)$, given input $\mathbf{Z}^{2,\ell-1}$, then we have

1. for $\ell = 1$, if $\mathbf{Z}^{2,\ell-1} \notin \mathcal{E}_1$, then $1 - \mathbf{logit}_{5,j_1}^{(t)} = O\left(\frac{1}{\text{poly}d}\right)$.

2. $\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \notin \tilde{\mathcal{E}}_2$, then

- (a) if $g_1 = g_2 \wedge y_0 \neq y_1$, $\mathbf{logit}_{5,j_2}^{(t)} = \Omega(1)$, $1 - \mathbf{logit}_{5,j_2}^{(t)} - \mathbf{logit}_{5,\tau(g_2(y_0))}^{(t)} = \frac{1}{\text{poly}d}$.
- (b) if $g_1 \neq g_2 \wedge y_0 = y_1$, $\mathbf{logit}_{5,j_2}^{(t)} = \Omega(1)$, $1 - \mathbf{logit}_{5,j_2}^{(t)} - \mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} = \frac{1}{\text{poly}d}$.
- (c) if $g_1 = g_2 \wedge y_0 = y_1$, $1 - \mathbf{logit}_{5,j_2}^{(t)} = O\left(\frac{1}{\text{poly}d}\right)$.

D.4.2 Gradient Lemma

Lemma D.28. If Induction D.2 holds for all iterations $\in [T_{2,1,s}, t)$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$, we have

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \Theta\left(\frac{\log d}{d}\right).$$

Proof. The proof is similar to Lemma D.16, but we need to shift our focus to $\left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s}$. By Lemma D.23 and Lemma D.26, we have

$$\begin{aligned} \mathcal{N}_{s,3,2,i}^{(t)} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \right. \\ &\quad \left. \left(\left(V_{j_2,r_{g_2 \cdot y_1}}(g_2) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2} \right] \\ &+ \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \right. \\ &\quad \left. \left(\left(V_{j_2,r_{g_2 \cdot y_1}}(g_2) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2^c} \right] \end{aligned}$$

$$\begin{aligned}
&\stackrel{(a)}{=} \Theta\left(\frac{1}{d}\right) \cdot \Omega(1) \cdot \Theta(B) + \Theta\left(\frac{1}{d}\right) \cdot O(1) \cdot \Theta(B) \cdot O\left(\frac{1}{\log d}\right) \\
&= \Theta\left(\frac{\log d}{d}\right),
\end{aligned}$$

where the inequality (a) follows from Lemma D.25 and Lemma D.26.

Furthermore, for $\mathcal{N}_{s,3,2,ii}^{(t)}$, we have

$$\begin{aligned}
&\left| \mathcal{N}_{s,3,2,ii}^{(t)} \right| \\
&\leq \left| \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \mathbf{logit}_{5,j'_2}^{(t)} \cdot \left(\left(V_{j'_2,r_{g_2 \cdot y_0}}(g_2) - \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2} \right] \right| \\
&+ \left| \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)} \cdot \left(\sum_{r \in \mathfrak{A}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \cdot \left(V_{j,r}(g_2) - \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2^c} \right] \right| \\
&\leq \Theta\left(\frac{1}{d}\right) \cdot O\left(\frac{1}{d^{1-c_6 C_B}}\right) \cdot \Theta(B) + \Theta\left(\frac{1}{d}\right) \cdot O(1) \cdot \Theta(B) \cdot O\left(\frac{1}{\log d}\right) \leq O\left(\frac{1}{d}\right).
\end{aligned}$$

$|\mathcal{N}_{s,3,1,i}^{(t)}|$ and $|\mathcal{N}_{s,3,2,i}^{(t)}|$ can be upper bounded by $O\left(\frac{\log d}{d}\right)$ as Lemma D.16. Thus, we complete the proof. \square

Lemma D.29. *If Induction D.2 holds for all iterations $\in [T_{2,1,s}, t)$, given $s \in \tau(\mathcal{X})$, we have*

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \Theta\left(\frac{\log d}{d}\right).$$

Proof. The proof follows a similar analysis to Lemma D.19, and we thus omit the details here. \square

Lemma D.30. *If Induction D.2 holds for all iterations $\in [T_{2,1,s}, t)$, given $s \in \tau(\mathcal{X})$, we have*

$$\begin{aligned}
&\sum_{t'=T_{2,1,s}}^t \left(\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t')}} \text{Loss}_5^{2,\ell} \right]_{s,s} - \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(t')}} \text{Loss}_5^{2,\ell} \right]_{s,s} \right) \\
&\geq -O\left(\frac{1}{\log d}\right) \cdot \left(\sum_{t'=T_{2,1,s}}^t \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(t')}} \text{Loss}_5^{2,\ell} \right]_{s,s} \right).
\end{aligned}$$

Proof. Following the analogous analysis as Lemma D.18, we have

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} + \sum_{\ell=1}^2 \left[\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \sum_{\ell \in [2]} \sum_{\kappa \in \{i, ii, iii\}} \mathcal{N}_{s,3,\ell,\kappa}^{(t)} - \mathcal{N}_{s,4,\ell,\kappa}^{(t)}.$$

Meanwhile $|\mathcal{N}_{s,p,\ell,iii}^{(t)}| = O\left(\frac{1}{\text{poly}d}\right)$ for $p \in \{3, 4\}$ and $\ell \in [2]$, we can focus on the gradient difference between $\mathbf{Q}_{4,3}^{(t)}$ and $\mathbf{Q}_{4,4}^{(t)}$ contributed by other terms.

For $\ell = 1$, since $\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \geq \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)}$, and thus it is straightforward to see that $\mathcal{N}_{s,3,1,i}^{(t)} -$

$\mathcal{N}_{s,4,1,i}^{(t)} \geq 0$. Furthermore, we have

$$\begin{aligned}
& \mathcal{N}_{s,3,1,ii}^{(t)} - \mathcal{N}_{s,4,1,ii}^{(t)} \\
&= -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \cdot \mathbf{logit}_{5,j'_1}^{(t)} \cdot \right. \\
&\quad \left. \left(\left(V_{j'_1, r_{g_2 \cdot y_0}}(g_1) - \Lambda_{5,j'_1, r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \\
&+ \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot \mathbf{logit}_{5,j'_1}^{(t)} \cdot \right. \\
&\quad \left. \left(\left(V_{j'_1, r_{g_2 \cdot y_0}}(y_0) - \Lambda_{5,j'_1, r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \\
&- \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \cdot \sum_{y \neq y_0 \in \mathcal{Y}} \mathbf{logit}_{5,\tau(g_1(y))}^{(t)} \cdot \right. \\
&\quad \left. \left(\left(V_{\tau(g_1(y)), r_{g_1 \cdot y}}(g_1) - \Lambda_{5,\tau(g_1(y)), r_{g_1 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1^c} \right] \\
&+ \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot \sum_{y \neq y_0 \in \mathcal{Y}} \mathbf{logit}_{5,\tau(g_1(y))}^{(t)} \cdot \right. \\
&\quad \left. \left(\left(V_{\tau(g_1(y)), r_{g_1 \cdot y}}(y_0) - \Lambda_{5,\tau(g_1(y)), r_{g_1 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1^c} \right] \\
&\stackrel{(a)}{\geq} \Theta\left(\frac{1}{d}\right) \cdot \Omega\left(\frac{1}{d^{2C_B(c_1+c_2)}}\right) \cdot \Theta(B) - \Theta\left(\frac{1}{d}\right) \cdot O\left(\frac{1}{\text{poly}d}\right) \cdot O(B) \cdot O\left(\frac{1}{\log d}\right) \geq \Omega\left(\frac{\log d}{d^{1+2C_B(c_1+c_2)}}\right).
\end{aligned}$$

where the inequality (a) is due to Lemma D.25 and Lemma D.27.

For $\ell = 2$, since $\mathbf{Attn}_{\text{ans},\text{pred},2 \rightarrow \text{pred},1}^{(t)} \geq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}$, and thus it is straightforward to see that $\mathcal{N}_{s,3,2,i}^{(t)} - \mathcal{N}_{s,4,2,i}^{(t)} \geq 0$. Moreover, we have

$$\begin{aligned}
& \mathcal{N}_{s,3,2,ii}^{(t)} - \mathcal{N}_{s,4,2,ii}^{(t)} \\
&= -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \mathbf{logit}_{5,j'_2}^{(t)} \cdot \right. \\
&\quad \left. \left(\left(V_{j'_2, r_{g_2 \cdot y_0}}(g_2) - \Lambda_{5,j'_2, r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2} \right] \tag{69}
\end{aligned}$$

$$\begin{aligned}
&+ \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \mathbf{logit}_{5,j'_2}^{(t)} \cdot \right. \\
&\quad \left. \left(\left(V_{j'_2, r_{g_2 \cdot y_0}}(y_1) - \Lambda_{5,j'_2, r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2} \right] \tag{70}
\end{aligned}$$

$$\begin{aligned}
&- \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)} \cdot \right. \\
&\quad \left. \left(\sum_{r \in \mathfrak{A}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \cdot \left(V_{j,r}(g_2) - \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2^c} \right] \tag{71}
\end{aligned}$$

$$\begin{aligned}
& + \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)} \cdot \right. \\
& \quad \left. \left(\sum_{r \in \mathfrak{A}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \cdot \left(V_{j,r}(y_1) - \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2^c} \right]. \quad (72)
\end{aligned}$$

By Lemma D.25, we obtain that

$$(69) + (70) \geq -O\left(\frac{1}{d}\right) \cdot O\left(\frac{1}{d^{1-c_6 C_B}}\right) \cdot \Theta(B) \geq -O\left(\frac{\log d}{d^{2-c_6 C_B}}\right).$$

By Lemma D.26, for $\tilde{\mathcal{E}}_2^c$, we only need to consider the case that $g_1 = g_2 \wedge y_0 \neq y_1$, and we have

$$\begin{aligned}
& (71) + (72) \\
& \geq -\mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} \cdot \sum_{y \neq y_1 \in \mathcal{Y}} \mathbf{logit}_{5, \tau(g_2(y))}^{(t)} \cdot \right. \\
& \quad \left. \left(\left(V_{\tau(g_2(y)), r_{g_2 \cdot y}}(g_2) - \Lambda_{5, \tau(g_2(y)), r_{g_2 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \{g_1=g_2 \wedge y_0 \neq y_1\}} \right] \\
& + \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \cdot \sum_{y \neq y_1 \in \mathcal{Y}} \mathbf{logit}_{5, \tau(g_2(y))}^{(t)} \cdot \right. \\
& \quad \left. \left(\left(V_{\tau(g_2(y)), r_{g_2 \cdot y}}(y_1) - \Lambda_{5, \tau(g_2(y)), r_{g_2 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \{g_1=g_2 \wedge y_0 \neq y_1\}} \right] \\
& \geq -\Theta\left(\frac{1}{d}\right) \cdot O(1) \cdot \Theta(B) \cdot O\left(\frac{1}{\log d}\right) \geq -O\left(\frac{1}{d}\right).
\end{aligned}$$

Putting it all together, combining with the fact that c_6 and $(c_1 + c_2)$ are sufficiently small, we finish the proof. \square

Lemma D.31 (Lower bound of gap). *If Induction D.2 holds for all iterations $s \in [T_{2,1,s}, t)$, given $s \in \tau(\mathcal{X})$, we have $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} - [\mathbf{Q}_{4,4}^{(t)}]_{s,s} \geq \Omega\left(\frac{1}{\log d}\right)$.*

Proof. Letting \tilde{T} denote the first time that $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} - [\mathbf{Q}_{4,4}^{(t)}]_{s,s} \leq \frac{1}{2} \left([\mathbf{Q}_{4,3}^{(T_{2,1,s})}]_{s,s} - [\mathbf{Q}_{4,4}^{(T_{2,1,s})}]_{s,s} \right)$, which implies that

$$\sum_{t'=T_{2,1,s}}^{\tilde{T}} \left(\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t')}} \text{Loss}_5^{2,\ell} \right]_{s,s} - \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(t')}} \text{Loss}_5^{2,\ell} \right]_{s,s} \right).$$

Hence, by Lemma D.30, we have $[\mathbf{Q}_{4,3}^{(\tilde{T})}]_{s,s}$ and $[\mathbf{Q}_{4,4}^{(\tilde{T})}]_{s,s}$ reaches $\Omega(1)$. Thus, we can have a refined lower bound for (71) + (72) in Lemma D.30, and obtain:

$$\begin{aligned}
& (71) + (72) \geq -O\left(\frac{1}{\text{poly}d}\right) \\
& - \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} \cdot \mathbf{logit}_{5, \tau(g_2(y_0))}^{(t)} \cdot \right. \\
& \quad \left. \left(\left(V_{\tau(g_2(y_0)), r_{g_2 \cdot y_0}}(g_2) - \Lambda_{5, \tau(g_2(y_0)), r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \{g_1=g_2 \wedge y_0 \neq y_1\}} \right] \\
& + \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \cdot \mathbf{logit}_{5, \tau(g_2(y_0))}^{(t)} \cdot \right.
\end{aligned}$$

$$\begin{aligned} & \left(\left(V_{\tau(g_2(y_0)), r_{g_2 \cdot y_0}}(y_1) - \Lambda_{5, \tau(g_2(y_0)), r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \{g_1=g_2 \wedge y_0 \neq y_1\}} \\ & \stackrel{(a)}{\geq} -O\left(\frac{1}{\text{poly}d}\right) - \Theta\left(\frac{1}{d}\right) \cdot O\left(\frac{1}{d^{\Omega(1)}}\right) \cdot \Theta(B) \cdot O\left(\frac{1}{\log d}\right) \geq -O\left(\frac{1}{d^{1+\Omega(1)}}\right), \end{aligned}$$

where (a) follows from the fact that on the event $\{g_1 = g_2 \wedge y_0 \neq y_1\}$, we have $\Lambda_{5, j_2, r_{g_2 \cdot y_1}}^{(t)} - \Lambda_{5, j_2, r_{g_2 \cdot y_0}}^{(t)} \geq \Omega(B)$ once $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$ and $[\mathbf{Q}_{4,4}^{(t)}]_{s,s}$ reach constant magnitude, and consequently, the logit satisfies $\mathbf{logit}_{5, \tau(g_2(y_0))}^{(t)} \leq O\left(\frac{1}{d^{\Omega(1)}}\right)$.

Therefore,

$$\sum_{t'=\tilde{T}+1}^t \left(\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}^{(t')}_{4,3}} \text{Loss}_5^{2,\ell} \right]_{s,s} - \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}^{(t')}_{4,4}} \text{Loss}_5^{2,\ell} \right]_{s,s} \right) \geq -O\left(\frac{1}{\log d \cdot d^{\Omega(1)}}\right) \cdot O(1),$$

which means that $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} - [\mathbf{Q}_{4,4}^{(t)}]_{s,s} \geq \Omega\left(\frac{1}{\log d}\right) - O\left(\frac{1}{\log d \cdot d^{\Omega(1)}}\right) \geq \Omega\left(\frac{1}{\log d}\right)$. \square

Lemma D.32. *If Induction D.2 holds for all iterations $\in [T_{2,1,s}, t)$, given $s' \neq s \in \tau(\mathcal{X})$, for $p \in \{3, 4\}$, we have*

$$\left| \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s'} \right| \leq O\left(\frac{\log d}{d^2}\right) = O\left(\frac{1}{d}\right) \cdot \left| \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \right|.$$

D.4.3 At the End of Stage 2.2

Putting gradient lemmas together, we can directly prove that Induction D.2 holds for all iterations t until the end of stage 2.2, where we can conclude the following:

Lemma D.33 (End of Stage 2.2). *Given $s \in \tau(\mathcal{X})$, Induction D.2 holds for all iterations $T_{2,1,s} \leq t < T_{2,2,s} = O\left(\frac{d}{\eta \log d}\right)$, then at the end of stage 2.2, we have*

- (a) $[\mathbf{Q}_{4,p}^{(t)}]_{s,s} = \Omega(1)$ for $p \in \{3, 4\}$;
- (b) $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} - [\mathbf{Q}_{4,4}^{(t)}]_{s,s} \in \left[\Omega\left(\frac{1}{\log d}\right), O(1)\right]$;
- (c) $|[\mathbf{Q}_{4,p}^{(t)}]_{s,s'}| \leq O\left(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,s}}{d}\right)$ for $s' \in \tau(\mathcal{X}) \neq s$, and other $[\mathbf{Q}_{4,p}]_{s,s'} = 0$.

D.5 Stage 2.3: Decrease of Gap and Convergence

After rapid growth of diagonal entries in stage 2.2, we now focus on the convergence of the attention and logit matrices, and the decrease of the gap between $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$ and $[\mathbf{Q}_{4,4}^{(t)}]_{s,s}$. Recall that

$$\begin{aligned} \epsilon_{\text{attn}}^{L,\ell}(\mathbf{Z}^{L,\ell-1}) &= 1 - \mathbf{Attn}_{\text{ans}, \ell-1 \rightarrow \text{pred}, \ell}(\mathbf{Z}^{L,\ell-1}) - \mathbf{Attn}_{\text{ans}, \ell-1 \rightarrow \text{ans}, \ell-1}(\mathbf{Z}^{L,\ell-1}), \\ \Delta^{L,\ell}(\mathbf{Z}^{L,\ell-1}) &= \mathbf{Attn}_{\text{ans}, \ell-1 \rightarrow \text{pred}, \ell}(\mathbf{Z}^{L,\ell-1}) - \mathbf{Attn}_{\text{ans}, \ell-1 \rightarrow \text{ans}, \ell-1}(\mathbf{Z}^{L,\ell-1}). \end{aligned}$$

Through stage 2.3, we will focus on the attention gap $\Delta^{L,\ell}(\mathbf{Z}^{L,\ell-1})$ instead of the gap of the attention matrices. We abbreviate $\epsilon_{\text{attn}}^{L,\ell}(\mathbf{Z}^{L,\ell-1})$ and $\Delta^{L,\ell}(\mathbf{Z}^{L,\ell-1})$ as $\epsilon_{\text{attn}}^{L,\ell}$ and $\Delta^{L,\ell}$ for simplicity.

Induction D.3. *Given $\epsilon \geq \tilde{\Omega}(\sigma_0)$, for $s \in \tau(\mathcal{X})$, let $T_{2,3,s}$ denote the first time that $\mathbb{E}[\epsilon_{\text{attn}}^{2,2} \mid \tau(x_1) = s] \leq \epsilon$. For all iterations $T_{2,2,s} \leq t < T_{2,3,s}$, we have the following holds:*

- (a) $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$ and $[\mathbf{Q}_{4,4}^{(t)}]_{s,s}$ monotonically increases $\leq \tilde{O}(1)$;

- (b) $\Delta^{2,\ell} \geq 0$ for any $\mathbf{Z}^{2,\ell}$ with $\ell \in \{1, 2\}$;
- (c) $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \leq 0.5 + \tilde{c}_1$ for some small constant $\tilde{c}_1 > 0$;
- (d) for $(p, q) \in \{(4, 3), (4, 4)\}$, $|\mathbf{Q}_{p,q}^{(t)}|_{s,s'} \leq O\left(\frac{[\mathbf{Q}_{p,q}^{(t)}]_{s,s}}{d}\right)$ for $s' \in \tau(\mathcal{X}) \neq s$; other $[\mathbf{Q}_{p,q}^{(t)}]_{s,s'} = 0$.

D.5.1 Attention and Logit Preliminaries

Lemma D.34. *If Induction D.3 holds for all iterations $\in [T_{2,2,s}, t)$, given $\mathbf{Z}^{2,\ell-1}$ then we have*

1. for $\ell = 1$,

$$(a) \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \geq \Omega(1), \text{ and } \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} > \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)};$$

2. for $\ell = 2$,

$$(a) \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \geq \Omega(1), \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)}, \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \leq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)};$$

$$(b) \left| \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \right| \leq \tilde{O}\left(\frac{1}{d}\right).$$

Moreover, given $\mathbf{Z}^{2,1}$ and corresponding $\mathbf{Z}^{2,0}$, $\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)}(\mathbf{Z}^{2,0}) \geq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}(\mathbf{Z}^{2,1})$.

Lemma D.35. *If Induction D.3 holds for all iterations $\in [T_{2,2,s}, t)$, given input $\mathbf{Z}^{2,\ell-1}$, then we have*

1. for $\ell = 1$, if $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_1$, then

$$(a) \text{ for } j = j_1, \Lambda_{5,j_1,r}^{(t)} \ll -\varrho \text{ for } r \in \widehat{\mathfrak{A}}_{j_1} \setminus \{r_{g_1 \cdot y_0}\};$$

$$(b) \text{ for } j = j'_1 \triangleq \tau(g_2(y_0)) = \tau(g_1(\tilde{y})), \Lambda_{5,j'_1,r}^{(t)} \ll -\varrho \text{ for } r \in \widehat{\mathfrak{A}}_{j'_1} \setminus \{r_{g_2 \cdot y_0}, r_{g_1 \cdot \tilde{y}}\};$$

$$(c) \text{ for other } j \in \tau(\mathcal{Y}), \text{ assuming } j = \tau(g_1(y)) \text{ for some } y \neq y_0, \text{ then } \Lambda_{5,j,r}^{(t)} \ll -\varrho \text{ for } r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_1 \cdot y}\}.$$

2. $\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \in \tilde{\mathcal{E}}_2$, then

$$(a) \text{ for } j = j_2, \Lambda_{5,j_2,r}^{(t)} \ll -\varrho \text{ for } r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\};$$

$$(b) \text{ for } j = j'_2 \triangleq \tau(g_2(y_0)), \Lambda_{5,j'_2,r}^{(t)} \ll -\varrho \text{ for } r \in \widehat{\mathfrak{A}}_{j'_2} \setminus \{r_{g_2 \cdot y_0}\};$$

$$(c) \text{ for other } j \in \tau(\mathcal{Y}), \text{ if } j = \tau(g_2(y)) \text{ for some } y \in \mathcal{Y}, \text{ then } \Lambda_{5,j,r}^{(t)} \ll -\varrho \text{ for } r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}.$$

Lemma D.36. *If Induction D.3 holds for all iterations $\in [T_{2,2,s}, t)$, given input $\mathbf{Z}^{2,\ell-1}$, then we have*

1. $\ell = 1$, for $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_1$,

$$(a) \Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} = (1 - 2\epsilon_{\text{attn}}^{2,1})B \pm O(\delta) \geq \left(\frac{1}{3} + c_1\right)B;$$

$$(b) \Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)} = 2(\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)})B \pm O(\delta) \geq 2(c_1 + c_2)B;$$

$$(c) \text{ for } y \neq y_0, \Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)} = \left(2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} - 1\right)B \pm O(\delta), \text{ which is only activated if } \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} > \frac{1}{2}$$

2. $\ell = 2$, for $\mathbf{Z}^{2,\ell-1} \in \tilde{\mathcal{E}}_2$,

$$(a) \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = (1 - 2\epsilon_{\text{attn}}^{2,2})B \pm O(\delta) \geq 4c_5B;$$

$$(b) \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} = \Delta^{2,2} \cdot B \pm O(\delta), \text{ and}$$

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} = 2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)})B \pm O(\delta);$$

(c) for $y \neq y_0, y_1$, $\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} = \left(2\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - 1\right)B \pm O(\delta)$, which is only activated if $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} > \frac{1}{2}$.

Lemma D.37. If Induction D.3 holds for all iterations $\in [T_{2,2,s}, t)$, given input $\mathbf{Z}^{2,\ell-1}$, then we have

1. for $\ell = 1$, if $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_1$,

$$1 - \mathbf{logit}_{5,j_1}^{(t)} = \Theta(1) \cdot \mathbf{logit}_{5,j_1'}^{(t)} = \Theta\left(\frac{1}{d^{2(\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)})C_B}}\right).$$

2. for $\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \in \tilde{\mathcal{E}}_2$,

$$\mathbf{logit}_{5,j_2}^{(t)} = \Theta\left(\frac{1}{d^{2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)})C_B + d^{1-\Delta^{2,2}C_B}}}\right),$$

moreover,

$$1 - \mathbf{logit}_{5,j_2}^{(t)} \geq \min\left\{\Omega(1), \Omega\left(\frac{1}{d^{C_B \cdot (1-2\epsilon_{\text{attn}}^{2,2})-1}}\right)\right\}.$$

Lemma D.38. If Induction D.3 holds for all iterations $\in [T_{2,2,s}, t)$, given input $\mathbf{Z}^{2,\ell-1}$, then we have

1. for $\ell = 1$, if $\mathbf{Z}^{2,\ell-1} \notin \mathcal{E}_1$, then

(a) for $j = j_1$, $\hat{\mathfrak{A}}_{j_1} = \{r_{g_1 \cdot y_0}\}$, $\Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} = B \pm O(\delta)$;

(b) for $j \neq j_1 \in \tau(\mathcal{Y})$, assuming $j = \tau(g_1(y))$ for $y \neq y_0$, then $\Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} - \Lambda_{5,j,r_{g_1 \cdot y}}^{(t)} \geq 2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)}$.
 $B \geq \Omega(B)$ and $\Lambda_{5,j,r}^{(t)} \ll -\varrho$ for $r \in \hat{\mathfrak{A}}_j \setminus \{r_{g_1 \cdot y}\}$.

2. $\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \notin \tilde{\mathcal{E}}_2$, then

(a) if $g_1 = g_2 \wedge y_0 \neq y_1$,

- i. for $j = j_2$, $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = (1 - \epsilon_{\text{attn}}^{2,2}) \cdot B \pm O(\delta)$ and $\Lambda_{5,j_2,r}^{(t)} \ll -\varrho$ for $r \in \hat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$;
- ii. for $j = \tau(g_2(y_0))$, $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} - \Lambda_{5,j,r_{g_2 \cdot y_0}}^{(t)} = 2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)})B \pm O(\delta)$ and $\Lambda_{5,j,r}^{(t)} \ll -\varrho$ for $r \in \hat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y_0}\}$;
- iii. for other $j \in \tau(\mathcal{Y})$, assuming $j = \tau(g_2(y))$ for some $y \neq y_0, y_1$, $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} - \Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} = 2\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}B \pm O(\delta)$ and $\Lambda_{5,j,r}^{(t)} \ll -\varrho$ for $r \in \hat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$.

(b) if $g_1 \neq g_2 \wedge y_0 = y_1$,

- i. for $j = j_2$, $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = (1 - \epsilon_{\text{attn}}^{2,2}) \cdot B \pm O(\delta)$ and $\Lambda_{5,j_2,r}^{(t)} \ll -\varrho$ for $r \in \hat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$;
- ii. for $j = \tau(g_1(y_1)) = \tau(g_2(\tilde{y}))$,

$$\begin{aligned} \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} - \Lambda_{5,j,r_{g_1 \cdot y_1}}^{(t)} &= 2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)})B \pm O(\delta) \\ \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} - \Lambda_{5,j,r_{g_2 \cdot \tilde{y}}}^{(t)} &= 2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)})B \pm O(\delta), \end{aligned}$$

where $r_{g_2 \cdot \tilde{y}}$ is only activated if $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} > \frac{1}{2}$. $\Lambda_{5,j,r}^{(t)} \ll -\varrho$ for $r \in \hat{\mathfrak{A}}_j \setminus \{r_{g_1 \cdot y_1}, r_{g_2 \cdot \tilde{y}}\}$;

iii. for other $j \in \tau(\mathcal{Y})$, assuming $j = \tau(g_2(y))$ for $y \neq y_1, \tilde{y}$, then $r_{g_2 \cdot y}$ is only activated if $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} > \frac{1}{2}$ and

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} - \Lambda_{5,j_2,r_{g_2 \cdot y}}^{(t)} = 2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)})B \pm O(\delta).$$

$\Lambda_{5,j,r}^{(t)} \ll -\varrho$ for $r \in \hat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$.

- (c) if $g_1 = g_2 \wedge y_0 = y_1$,
 i. for $j = j_2$, $\widehat{\mathfrak{A}}_{j_2} = \{r_{g_2 \cdot y_1}\}$, $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = B \pm O(\delta)$;
 ii. for other $j \in \tau(\mathcal{Y})$, assuming $j = \tau(g_2(y))$,

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} - \Lambda_{5,j_2,r_{g_2 \cdot y}}^{(t)} = 2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)})B \pm O(\delta).$$

$$\text{and } \Lambda_{5,j,r}^{(t)} \ll -\varrho \text{ for } r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}.$$

Lemma D.39. If Induction D.3 holds for all iterations $\in [T_{2,2,s}, t)$, given input $\mathbf{Z}^{2,\ell-1}$, then we have

1. for $\ell = 1$, if $\mathbf{Z}^{2,\ell-1} \notin \mathcal{E}_1$, then $1 - \mathbf{logit}_{5,j_1}^{(t)} = O\left(\frac{1}{d^{2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} C_B}}\right) = O\left(\frac{1}{\text{poly}d}\right)$.
2. $\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \notin \widetilde{\mathcal{E}}_2$, then
 - (a) if $g_1 = g_2 \wedge y_0 \neq y_1$, $\mathbf{logit}_{5,\tau(g_2(y_0))}^{(t)} = O\left(\frac{1}{d^2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)}) C_B}\right)$.
 - (b) if $g_1 \neq g_2 \wedge y_0 = y_1$, $\mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} = O\left(\frac{1}{d^2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)}) C_B}\right)$.
 - (c) if $g_1 = g_2 \wedge y_0 = y_1$, $1 - \mathbf{logit}_{5,j_2}^{(t)} = O\left(\frac{1}{d^2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}) C_B}\right) = \frac{1}{\text{poly}d}$.

D.5.2 Gradient Lemma

Lemma D.40. If Induction D.3 holds for all iterations $\in [T_{2,2,s}, t)$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,4}]_{s,s}$, we have

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \geq \Omega\left(\frac{\epsilon \log d}{d^{(1-2\epsilon)C_B}}\right).$$

Proof. By gradient decomposition, we have

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \sum_{\ell \in [2]} \sum_{\kappa \in \{i, ii, iii\}} \mathcal{N}_{s,4,\ell,\kappa}^{(t)}.$$

Firstly, for $\mathcal{N}_{s,4,2,i}^{(t)}$, by Lemma D.35 and Lemma D.38, we have

$$\begin{aligned} \mathcal{N}_{s,4,2,i}^{(t)} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \right. \\ &\quad \left. \left(\left(V_{j_2,r_{g_2 \cdot y_1}}(y_1) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \pm \widetilde{O}(\sigma_0) \right) \pm \widetilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \widetilde{\mathcal{E}}_2} \right] \\ &+ \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \right. \\ &\quad \left. \left(\left(V_{j_2,r_{g_2 \cdot y_1}}(y_1) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \pm \widetilde{O}(\sigma_0) \right) \pm \widetilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \widetilde{\mathcal{E}}_2^c} \right] \\ &\stackrel{(a)}{\geq} \Theta\left(\frac{1}{d}\right) \cdot \mathbb{E} \left[\min \left\{ \Omega(1), \Omega\left(\frac{1}{d^{C_B \cdot (1-2\epsilon_{\text{attn}}^{2,2}) - 1}}\right) \right\} \cdot 2\epsilon_{\text{attn}}^{2,2} \cdot B \mid \tau(x_1) = s \right] \\ &\geq \Omega\left(\frac{\epsilon \log d}{d^{(1-2\epsilon)C_B}}\right), \end{aligned} \tag{73}$$

where (a) follows from Lemma D.37. Noticing that $\mathcal{N}_{s,4,1,i}^{(t)} > 0$ and $|\mathcal{N}_{s,4,\ell,iii}^{(t)}|$ for $\ell \in [2]$ is sufficiently small, to provide a lower bound for $\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s}$, we only need to focus on the negative gradient from $\mathcal{N}_{s,4,1,ii}^{(t)}$ and $\mathcal{N}_{s,4,2,ii}^{(t)}$.

$$\begin{aligned}
& \mathcal{N}_{s,4,2,ii}^{(t)} = \\
& -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \mathbf{logit}_{5,j_2'}^{(t)} \cdot \right. \\
& \quad \left. \left(\left(V_{j_2',r_{g_2,y_0}}(y_1) - \Lambda_{5,j_2',r_{g_2,y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2} \right] \\
& -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)} \cdot \right. \\
& \quad \left. \left(\sum_{r \in \tilde{\mathcal{R}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \cdot \left(V_{j,r}(y_1) - \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2^c} \right] \\
& \geq -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} \cdot \left(\mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y_1)),r_{g_1,y_1}}^{(t)}) \cdot \right. \right. \\
& \quad \left. \left. \left(V_{\tau(g_1(y_1)),r_{g_1,y_1}}(y_1) - \Lambda_{5,\tau(g_1(y_1)),r_{g_1,y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\{\tau(x_1)=s\} \cap \{g_1 \neq g_2, y_0=y_1\}} \right]. \quad (74)
\end{aligned}$$

When $\mathbb{E} \left[C_B \cdot (1 - 2\epsilon_{\text{attn}}^{2,2}) \mid \tau(x_1) = s \right] < 1$, by Lemma D.37 and Lemma D.39, we have

$$\begin{aligned}
& \mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} \mathbb{1}_{\{\tau(x_1)=s\} \cap \{g_1 \neq g_2, y_0=y_1\}} \\
& \leq O \left(\frac{1}{d^2 (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)}) C_B} \right) (1 - \mathbf{logit}_{5,j_2}^{(t)}) \mathbb{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2}.
\end{aligned}$$

During this time, $\epsilon_{\text{attn}}^{2,2} \geq \Omega(1)$, thus we can lower bound $\mathcal{N}_{s,4,2,ii}^{(t)}$ by $\mathcal{N}_{s,4,2,ii}^{(t)} \geq -O \left(\frac{1}{d^{\Omega(1)} \cdot \log d} \right) \cdot \mathcal{N}_{s,4,2,i}^{(t)}$, which implies that the negative gradient from $\mathcal{N}_{s,4,2,ii}^{(t)}$ is dominated by the positive gradient from $\mathcal{N}_{s,4,2,i}^{(t)}$.

When $\mathbb{E} \left[C_B \cdot (1 - 2\epsilon_{\text{attn}}^{2,2}) \mid \tau(x_1) = s \right] \geq 1$, since $2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)}) C_B \geq (1 - 2\eta^{2,2}) C_B$, we obtain

$$\mathbf{logit}_{\tau(g_1(y_1))}^{(t)} \mathbb{1}_{\{\tau(x_1)=s\} \cap \{g_1 \neq g_2, y_0=y_1\}} \leq O \left(\frac{1}{d} \right) \cdot (1 - \mathbf{logit}_{j_2}^{(t)}) \mathbb{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2}. \quad (75)$$

For the event $\{g_1 \neq g_2, y_0 = y_1\}$, if $\Lambda_{5,\tau(g_1(y_1)),r_{g_1,y_1}}^{(t)}$ is still in the linear regime, we have

$$\Lambda_{5,\tau(g_1(y_1)),r_{g_1,y_1}}^{(t)} = (1 - 2\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \cdot B \pm O(\delta) \geq \varrho,$$

which implies

$$\epsilon_{\text{attn}}^{2,2} \geq 1 - 2\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \geq \Omega \left(\frac{\varrho}{B} \right). \quad (76)$$

Hence, putting (76) back to (74), and putting (75) back to (74), we can lower bound $\mathcal{N}_{s,4,2,ii}^{(t)}$ as follows $\mathcal{N}_{s,4,2,ii}^{(t)} \geq -O \left(\frac{1}{d} \right) \cdot \mathcal{N}_{s,4,2,i}^{(t)}$. If $\Lambda_{5,\tau(g_1(y_1)),r_{g_1,y_1}}^{(t)}$ falls into the smoothed regime, we can upper bound $\mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y_1)),r_{g_1,y_1}}^{(t)})$ by $O \left(\frac{\epsilon_{\text{attn}}^{2,2} B}{\varrho} \right)$, and then similarly, we can obtain $\mathcal{N}_{s,4,2,ii}^{(t)} \geq -O \left(\frac{1}{d} \right) \cdot \mathcal{N}_{s,4,2,i}^{(t)}$.

Following the analogous analysis, the negative gradient from $\mathcal{N}_{s,4,1,ii}^{(t)}$ can also be dominated by the positive gradient from $\mathcal{N}_{s,4,2,i}^{(t)}$. Hence, we complete the proof. \square

Lemma D.41. *If Induction D.3 holds for all iterations $\in [T_{2,2,s}, t)$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,p}]_{s,s'}$, $p \in \{3, 4\}$, $s' \neq s \in \tau(\mathcal{X})$, $\ell \in [2]$, we have*

$$\left| \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s'} \right| \leq O\left(\frac{1}{d}\right) \left| \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \right|.$$

D.5.3 Non-negative Gap

Lemma D.42. *If Induction D.3 holds for all iterations $\in [T_{2,2,s}, t)$, then at time t , we have $\Delta^{2,\ell} \geq 0$ for any $\mathbf{Z}^{2,\ell}$ with $\ell \in \{1, 2\}$.*

Proof. Let \tilde{T} denote the first time that $\mathbb{E}[\Delta^{2,2} \mid \tau(x_1) = s] < \alpha$, where $\alpha = \frac{1}{\text{poly}d}$.

Following the analogous analysis as Lemma D.30, we have

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,\ell} \right]_{s,s} + \sum_{\ell=1}^2 \left[\nabla_{\mathbf{Q}_{4,4}^{(\tilde{T})}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \sum_{\ell \in [2]} \sum_{\kappa \in \{i, ii, iii\}} \mathcal{N}_{s,3,\ell,\kappa}^{(\tilde{T})} - \mathcal{N}_{s,4,\ell,\kappa}^{(\tilde{T})}.$$

We can ignore the negligible difference introduced by $\mathcal{N}_{s,p,\ell,iii}^{(\tilde{T})}$ for $p \in \{3, 4\}$ and $\ell \in [2]$.

For $\ell = 1$, since $\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(\tilde{T})} \geq \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(\tilde{T})}$, and thus it is straightforward to see that $\mathcal{N}_{s,3,1,i}^{(\tilde{T})} - \mathcal{N}_{s,4,1,i}^{(\tilde{T})} \geq 0$. Furthermore, we have

$$\begin{aligned} & \mathcal{N}_{s,3,1,ii}^{(\tilde{T})} - \mathcal{N}_{s,4,1,ii}^{(\tilde{T})} \\ &= -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(\tilde{T})} \cdot \mathbf{logit}_{5,j'_1}^{(\tilde{T})} \cdot \mathbf{sReLU}'(\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)}) \right. \\ & \quad \left. \left(\left(V_{j'_1,r_{g_2 \cdot y_0}}(g_1) - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \end{aligned} \quad (77)$$

$$\begin{aligned} & + \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(\tilde{T})} \cdot \mathbf{logit}_{5,j'_1}^{(\tilde{T})} \cdot \mathbf{sReLU}'(\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(\tilde{T})}) \right. \\ & \quad \left. \left(\left(V_{j'_1,r_{g_2 \cdot y_0}}(y_0) - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1} \right] \end{aligned} \quad (78)$$

$$\begin{aligned} & - \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(\tilde{T})} \cdot \sum_{y \neq y_0 \in \mathcal{Y}} \mathbf{logit}_{5,\tau(g_1(y))}^{(\tilde{T})} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(\tilde{T})}) \right. \\ & \quad \left. \left(\left(V_{\tau(g_1(y)),r_{g_1 \cdot y}}(g_1) - \Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1^c} \right] \\ & + \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(\tilde{T})} \cdot \sum_{y \neq y_0 \in \mathcal{Y}} \mathbf{logit}_{5,\tau(g_1(y))}^{(\tilde{T})} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(\tilde{T})}) \right. \\ & \quad \left. \left(\left(V_{\tau(g_1(y)),r_{g_1 \cdot y}}(y_0) - \Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \mathcal{E}_1^c} \right]. \end{aligned}$$

By Lemma D.39, we obtain that $\mathcal{N}_{s,3,1,ii}^{(\tilde{T})} - \mathcal{N}_{s,4,1,ii}^{(\tilde{T})}$ is dominated by (77) + (78). Moreover, by Lemma D.37,

we have

$$\begin{aligned}\mathbf{logit}_{5,\tau(g_1(y))}^{(\tilde{T})}(\mathbf{Z}^{2,0}) &\geq \Omega\left(\frac{1}{d^2(\mathbf{Attn}_{\text{ans},0\rightarrow\text{pred},1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},0\rightarrow\text{pred},2}^{(\tilde{T})})^{C_B}}\right) \\ &\geq \Omega\left(\frac{1}{d^2(\mathbf{Attn}_{\text{ans},1\rightarrow\text{pred},2}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1\rightarrow\text{pred},1}^{(\tilde{T})})^{C_B}}\right).\end{aligned}\quad (79)$$

For $\ell = 2$, we have

$$\begin{aligned}\mathcal{N}_{s,3,2,i}^{(\tilde{T})} - \mathcal{N}_{s,4,2,i}^{(\tilde{T})} &\geq \mathbb{E}\left[\mathbf{Attn}_{\text{ans},1\rightarrow\text{pred},2}^{(\tilde{T})} \cdot (1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) \cdot \right. \\ &\quad \left. \left((V_{j_2,r_{g_2 \cdot y_1}}(g_2) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2} \right] \\ &\quad - \mathbb{E}\left[\mathbf{Attn}_{\text{ans},1\rightarrow\text{ans},1}^{(\tilde{T})} \cdot (1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) \cdot \right. \\ &\quad \left. \left((V_{j_2,r_{g_2 \cdot y_1}}(y_1) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2} \right] \\ &\geq \Omega(\alpha \epsilon \log d) \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) \cdot \mathbf{1}_{\tau(x_1)=s}\right],\end{aligned}\quad (80)$$

where the last inequality follows from the definition of \tilde{T} .

$$\begin{aligned}\mathcal{N}_{s,3,2,ii}^{(\tilde{T})} - \mathcal{N}_{s,4,2,ii}^{(\tilde{T})} &= -\mathbb{E}\left[\mathbf{Attn}_{\text{ans},1\rightarrow\text{pred},2}^{(\tilde{T})} \cdot \mathbf{logit}_{5,j'_2}^{(\tilde{T})} \cdot \mathbf{sReLU}'(\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(\tilde{T})}) \cdot \right. \\ &\quad \left. \left((V_{j'_2,r_{g_2 \cdot y_0}}(g_2) - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2} \right]\end{aligned}\quad (81)$$

$$\begin{aligned}&+ \mathbb{E}\left[\mathbf{Attn}_{\text{ans},1\rightarrow\text{ans},1}^{(\tilde{T})} \cdot \mathbf{logit}_{5,j'_2}^{(\tilde{T})} \cdot \mathbf{sReLU}'(\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(\tilde{T})}) \cdot \right. \\ &\quad \left. \left((V_{j'_2,r_{g_2 \cdot y_0}}(y_1) - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2} \right]\end{aligned}\quad (82)$$

$$\begin{aligned}&- \mathbb{E}\left[\mathbf{Attn}_{\text{ans},1\rightarrow\text{pred},2}^{(\tilde{T})} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(\tilde{T})} \cdot \right. \\ &\quad \left. \left(\sum_{r \in \tilde{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(\tilde{T})}) \cdot (V_{j,r}(g_2) - \Lambda_{5,j,r}^{(\tilde{T})} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2^c} \right]\end{aligned}\quad (83)$$

$$\begin{aligned}&+ \mathbb{E}\left[\mathbf{Attn}_{\text{ans},1\rightarrow\text{ans},1}^{(\tilde{T})} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(\tilde{T})} \cdot \right. \\ &\quad \left. \left(\sum_{r \in \tilde{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(\tilde{T})}) \cdot (V_{j,r}(y_1) - \Lambda_{5,j,r}^{(\tilde{T})} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2^c} \right].\end{aligned}\quad (84)$$

Notice that by Lemma D.36 and the definition of \tilde{T} , we have

$$|(81) + (82)| \leq O((\alpha \log d)^{q-1}) \cdot \mathbb{E} \left[\left(1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \right) \cdot \mathbf{1}_{\tau(x_1)=s} \right],$$

which is dominated by (80) due to the choice of α .

Furthermore, by Lemma D.38, for $\tilde{\mathcal{E}}_2^c$, we only need to focus on the output $\mathbf{logit}_{5,\tau(g_2(y_0))}^{(t)}$ from the case $g_1 = g_2 \wedge y_0 \neq y_1$, and obtain

$$\begin{aligned} & (83) + (84) \\ & \geq -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \cdot \mathbf{logit}_{5,\tau(g_2(y_0))}^{(\tilde{T})} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(\tilde{T})}) \right. \\ & \quad \left. \left(\left(V_{\tau(g_2(y_0)),r_{g_2 \cdot y_0}}(g_2) - \Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \{g_1=g_2 \wedge y_0 \neq y_1\}} \right] \\ & + \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})} \cdot \mathbf{logit}_{5,\tau(g_2(y_0))}^{(\tilde{T})} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(\tilde{T})}) \right. \\ & \quad \left. \left(\left(V_{\tau(g_2(y_0)),r_{g_2 \cdot y_0}}(y_1) - \Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \{g_1=g_2 \wedge y_0 \neq y_1\}} \right]. \end{aligned}$$

By Lemma D.39, we have

$$\mathbf{logit}_{5,\tau(g_2(y_0))}^{(\tilde{T})}(\mathbf{Z}^{2,1}) = O\left(\frac{1}{d^2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(\tilde{T})})C_B}\right).$$

Since $\mathbb{E}[\Delta^{2,2} \mid \tau(x_1) = s] < \alpha$, where α is sufficiently small, then combining with (79), we have

$$\mathbf{logit}_{5,\tau(g_2(y_0))}^{(\tilde{T})}(\mathbf{Z}^{2,1}) = O(1) \cdot \mathbf{logit}_{5,\tau(g_2(y_0))}^{(\tilde{T})}(\mathbf{Z}^{2,0}).$$

Therefore, since $\tilde{\mathcal{E}}_2^c$ happens with probability $O(\frac{1}{\log d})$, we obtain that

$$(83) + (84) \geq -O\left(\frac{1}{\log d}\right) \left(\mathcal{N}_{s,3,1,ii}^{(t)} - \mathcal{N}_{s,4,1,ii}^{(t)} \right).$$

Putting it all together, we can conclude that when $\mathbb{E}[\Delta^{2,2} \mid \tau(x_1) = s]$ reaches below α , the gap in non-decreasing direction is guaranteed, i.e.,

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,\ell} \right]_{s,s} + \sum_{\ell=1}^2 \left[\nabla_{\mathbf{Q}_{4,4}^{(\tilde{T})}} \text{Loss}_5^{2,\ell} \right]_{s,s} > 0,$$

which completes the proof. \square

D.5.4 Upper Bound for Q

Lemma D.43. *If Induction D.3 holds for all iterations $\in [T_{2,2,s}, t)$, given $s \in \tau(\mathcal{X})$, then at time t , we have $[\mathbf{Q}_{4,p}^{(t)}]_{s,s} \leq \tilde{O}(1)$ for $p \in \{3, 4\}$.*

Proof. Denote the first time that $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$ reaches $\Omega(\log^{1+c} d)$ for some small constant $c > 0$ as \tilde{T} . Then by

direct calculations, we have

$$\begin{aligned}\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)} &\leq O\left(\frac{1}{e^{\Omega(\log^{1+c} d)}}\right), \\ \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} &\leq O\left(\frac{1}{e^{\Omega(\log^{1+c} d)}}\right).\end{aligned}$$

Moreover, by Lemma D.37, we can simply bound the logits as follows:

$$\begin{aligned}1 - \mathbf{logit}_{j_1}^{(t)} &\leq O\left(\frac{1}{d^{C_B(1-2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)})-1}}\right) \leq O\left(\frac{1}{d^{C_B/2-1}}\right) \text{ for } \mathbf{Z}^{2,0} \in \mathcal{E}_1; \\ 1 - \mathbf{logit}_{j_2}^{(t)} &\leq O\left(\frac{1}{d^{C_B(1-2\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - 2\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)})-1}}\right) \\ &\quad + O\left(\frac{1}{d^{2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)})C_B}}\right) \leq O\left(\frac{1}{d^{C_B/2-1}}\right) \text{ for } \mathbf{Z}^{2,1} \in \tilde{\mathcal{E}}_2.\end{aligned}$$

Thus, by focusing on $\mathcal{N}_{s,3,1,i}$ and $\mathcal{N}_{s,4,1,i}$, we have

$$\left| \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \right| \leq \frac{1}{d} \cdot O\left(\frac{1}{d^{C_B/2-1}}\right) \cdot O\left(\frac{1}{e^{\Omega(\log^{1+c} d)}}\right) \cdot B \leq O\left(\frac{\log d}{e^{\Omega(\log^{1+c} d)} d^{C_B/2}}\right),$$

which implies

$$\mathbf{Q}_{4,3}^{(T)} \leq \mathbf{Q}_{4,3}^{(\tilde{T})} + O\left(\frac{T \log d}{e^{\Omega(\log^{1+c} d)} d^{C_B/2}}\right) \leq \mathbf{Q}_{4,3}^{(\tilde{T})} + O\left(\frac{\text{poly} d \cdot \log d}{e^{\Omega(\log^{1+c} d)} d^{C_B/2}}\right) \leq O(\log^{1+c} d).$$

□

D.5.5 Attention Upper Bound

Lemma D.44. *If Induction D.3 holds for all iterations $\in [T_{2,2,s}, t)$, given $s \in \tau(\mathcal{X})$, then at time t , for any $\mathbf{Z}^{2,1}$, we have $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \leq 0.5 + \tilde{c}_1$, where $\tilde{c}_1 > 0$ is some small constant.*

Proof. Let \tilde{T} denote the first time that $\mathbb{E}[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \mid \tau(x_1) = s]$ reaches $0.5 + \tilde{c}$. where $\tilde{c} > 0$ is some small constant s.t., $2\tilde{c} \cdot C_B > 1$. At \tilde{T} , we have $\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(\tilde{T})} \mathbf{1}_{\tau(x_0)=s} \geq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \mathbf{1}_{\tau(x_1)=s}$; moreover, $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})}$ and $\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(\tilde{T})}$ is still at the constant level.

- For $\ell = 1$, by Lemma D.35 and Lemma D.38, for $y \in \mathcal{Y} \setminus \{y_0\}$, we have only $r_{g_1 \cdot y}$ has been activated to the linear regime for the prediction $\tau(g_1(y))$. Furthermore, we obtain

$$\begin{aligned}\sum_{y \in \mathcal{Y} \setminus \{y_0\}} \mathbf{logit}_{5,\tau(g_1(y))}^{(\tilde{T})} &\geq \frac{\log d \cdot e^{2\tilde{c}C_B \log d}}{\log d \cdot e^{2\tilde{c}C_B \log d} + O(d)} \left(1 - \mathbf{logit}_{5,j_1}^{(\tilde{T})}\right) \\ &= (1 - o(1)) \cdot \left(1 - \mathbf{logit}_{5,j_1}^{(\tilde{T})}\right).\end{aligned}$$

Thus,

$$\left[-\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,1} \right]_{s,s}$$

$$\begin{aligned}
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(\tilde{T})} \cdot \left(\left(1 - \mathbf{logit}_{5,j_1}^{(\tilde{T})} \right) \cdot \left(\left(V_{j_1, r_{g_1 \cdot y_0}}(g_1) - \Lambda_{5,j_1, r_{g_1 \cdot y_0}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \right. \right. \\
&\quad - \sum_{y \in \mathcal{Y} \setminus \{y_0\}} \mathbf{logit}_{5, \tau(g_1(y))}^{(\tilde{T})} \cdot \left(\left(V_{\tau(g_1(y)), r_{g_1 \cdot y}}(g_1) - \Lambda_{5, \tau(g_1(y)), r_{g_1 \cdot y}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \\
&\quad \left. \pm \frac{1}{\text{poly}d} \left(1 - \mathbf{logit}_{5,j_1}^{(\tilde{T})} \right) \cdot \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_0)=s} \Bigg] \\
&\leq \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(\tilde{T})} \cdot \left(\left(1 - \mathbf{logit}_{5,j_1}^{(\tilde{T})} \right) \cdot \left(\left(2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(\tilde{T})} \cdot B \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \right. \right. \\
&\quad - \sum_{y \in \mathcal{Y} \setminus \{y_0\}} \mathbf{logit}_{5, \tau(g_1(y))}^{(\tilde{T})} \cdot \left(\left(2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(\tilde{T})} \cdot B \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \\
&\quad \left. \pm \frac{1}{\text{poly}d} \left(1 - \mathbf{logit}_{5,j_1}^{(\tilde{T})} \right) \cdot \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_0)=s} \Bigg] < 0.
\end{aligned}$$

- For $\ell = 2$, for $\mathbf{Z}^{2,1} \in \tilde{\mathcal{E}}_2 \cup \{g_1 = g_2, y_0 \neq y_1 x\}$, we have

$$\begin{aligned}
\sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5, \tau(g_2(y))}^{(\tilde{T})} &\geq \frac{\log d \cdot e^{2\tilde{C}_B \log d}}{\log d \cdot e^{2\tilde{C}_B \log d} + O(d)} \left(1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} - \mathbf{logit}_{5,j'_2}^{(\tilde{T})} \right) \\
&= (1 - o(1)) \cdot \left(1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} - \mathbf{logit}_{5,j'_2}^{(\tilde{T})} \right).
\end{aligned}$$

Otherwise, we have

$$\sum_{y \in \mathcal{Y} \setminus \{y_1\}} \mathbf{logit}_{5, \tau(g_2(y))}^{(\tilde{T})} = (1 - o(1)) \cdot \left(1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} - \mathbf{logit}_{5,j'_2}^{(\tilde{T})} \right).$$

Therefore, similar to $\ell = 1$, we obtain $\left[-\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,2} \right]_{s,s} < 0$.

Combing the above two cases, we have $\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,\ell} \right]_{s,s} < 0$. It is also clear from Lemma D.40 that $\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(\tilde{T})}} \text{Loss}_5^{2,\ell} \right]_{s,s} \geq 0$. Hence $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})}$ cannot further grow once it reaches $0.5 + \tilde{c}$. \square

D.5.6 Decreasing Gap at the End of Convergence

Let \tilde{T} denote the first time that $\mathbb{E}[\epsilon_{\text{attn}}^{2,2} \mid \tau(x_1) = s] \leq 3\epsilon$, if $\mathbb{E}[\Delta^{2,2} \mid \tau(x_1) = s] \leq O\left(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d}\right)$, then we can let $T^* = \tilde{T}$ and stop the training. Otherwise, we have $\mathbb{E}[\Delta^{2,2} \mid \tau(x_1) = s] \geq \Omega\left(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d}\right)$. Following the similar argument as in Lemma D.40, we have the gradient contribution from $\ell = 1$ is dominated by the gradient contribution from $\ell = 2$. Thus, we focus on $\ell = 2$, and obtain

$$\begin{aligned}
&\left[-\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,2} \right]_{s,s} \\
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \cdot \left(\left(1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \right) \cdot \left(\left(V_{j_2, r_{g_2 \cdot y_1}}(g_2) - \Lambda_{5,j_2, r_{g_2 \cdot y_1}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \right. \right. \\
&\quad \left. \pm \frac{1}{\text{poly}d} \left(1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \right) \cdot \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\{\tau(x_0)=s\}} \Bigg] \\
&- \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \cdot \left(\sum_{y \in \mathcal{Y} \setminus \{y_1\}} \mathbf{logit}_{5, \tau(g_2(y))}^{(\tilde{T})} \cdot \mathbf{sReLU}'(\Lambda_{5, \tau(g_2(y)), r_{g_2 \cdot y}}^{(\tilde{T})}) \right) \right]
\end{aligned}$$

$$\begin{aligned}
& \cdot \left(\left(V_{\tau(g_2(y)), r_{g_2 \cdot y}}(g_2) - \Lambda_{5, \tau(g_2(y)), r_{g_2 \cdot y}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \tilde{\mathcal{E}}_2} \Big] \\
& - \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5, j}^{(t)} \cdot \right. \\
& \quad \left(\sum_{r \in \tilde{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5, j, r}^{(t)}) \cdot \left(V_{j, r}(g_2) - \Lambda_{5, j, r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2^c} \Big] \\
& \leq \mathbb{E} \left[\left(1 - \mathbf{logit}_{5, j_2}^{(t)} \right) \cdot O(\epsilon \log d) \mathbf{1}_{\tau(x_1)=s} \right].
\end{aligned}$$

Turning to $\mathbf{Q}_{4,4}$, we have

$$\begin{aligned}
& \left[-\nabla_{\mathbf{Q}_{4,4}^{(\tilde{T})}} \text{Loss}_5^{2,2} \right]_{s,s} \\
& = \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(\tilde{T})} \cdot \left(\left(1 - \mathbf{logit}_{5, j_2}^{(\tilde{T})} \right) \cdot \left(\left(V_{j_2, r_{g_2 \cdot y_1}}(y_1) - \Lambda_{5, j_2, r_{g_2 \cdot y_1}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \right. \right. \\
& \quad \left. \left. \pm \frac{1}{\text{poly}d} \left(1 - \mathbf{logit}_{5, j_2}^{(\tilde{T})} \right) \cdot \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\{\tau(x_0)=s\}} \right] \\
& - \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(\tilde{T})} \cdot \left(\sum_{y \in \mathcal{Y} \setminus \{y_1\}} \mathbf{logit}_{5, \tau(g_2(y))}^{(\tilde{T})} \cdot \mathbf{sReLU}'(\Lambda_{5, \tau(g_2(y)), r_{g_2 \cdot y}}^{(\tilde{T})}) \right. \right. \\
& \quad \left. \left. \cdot \left(\left(V_{\tau(g_2(y)), r_{g_2 \cdot y}}(y_1) - \Lambda_{5, \tau(g_2(y)), r_{g_2 \cdot y}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \tilde{\mathcal{E}}_2} \right] \tag{85}
\end{aligned}$$

$$\begin{aligned}
& - \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5, j}^{(t)} \cdot \right. \\
& \quad \left(\sum_{r \in \tilde{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5, j, r}^{(t)}) \cdot \left(V_{j, r}(y_1) - \Lambda_{5, j, r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \tilde{\mathcal{E}}_2^c} \Big]. \tag{86}
\end{aligned}$$

For (85), we have

$$\begin{aligned}
(85) & \geq \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(\tilde{T})} \cdot \left(-\mathbf{logit}_{5, j_2}^{(\tilde{T})} \cdot \right. \right. \\
& \quad \left. \min \left\{ \Omega(\log d), \Omega \left(\left((\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(\tilde{T})}) \log d \right)^{q-1} \log d \right) \right\} \right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \tilde{\mathcal{E}}_2} \Big] \\
& \geq \mathbb{E} \left[\left(1 - \mathbf{logit}_{5, j_2}^{(\tilde{T})} \right) \cdot \Omega \left(\frac{1}{d} \right) \cdot \right. \\
& \quad \left. \min \left\{ \Omega(\log d), \Omega \left(\left((\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(\tilde{T})}) \log d \right)^{q-1} \log d \right) \right\} \mathbf{1}_{\{\tau(x_0)=s\} \cap \tilde{\mathcal{E}}_2} \right].
\end{aligned}$$

Moreover, for (86), we have

$$(86) \geq -\mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(\tilde{T})} \cdot \mathbf{logit}_{5, \tau(g_1(y_1))}^{(\tilde{T})} \cdot \mathbf{sReLU}'(\Lambda_{5, \tau(g_1(y_1)), r_{g_1 \cdot y_1}}^{(\tilde{T})}) \right]$$

$$\begin{aligned} & \left(\left(V_{\tau(g_1(y_1)), r_{g_1 \cdot y_1}}(y_1) - \Lambda_{5, \tau(g_1(y_1)), r_{g_1 \cdot y_1}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\{\tau(x_1)=s\} \cap \{g_1 \neq g_2 \wedge y_0=y_1\}} \\ & \geq -\mathbb{E} \left[\left(1 - \mathbf{logit}_{5, j_2}^{(\tilde{T})} \right) \cdot O\left(\frac{1}{d}\right) \cdot O\left(\frac{1}{\log d}\right) \cdot O\left((\epsilon \log d)^{q-1}\right) \mathbf{1}_{\{\tau(x_0)=s\} \cap \tilde{\mathcal{E}}_2} \right], \end{aligned}$$

where the last inequality holds since $\Lambda_{5, \tau(g_1(y_1)), r_{g_1 \cdot y_1}}^{(\tilde{T})} \leq \left(\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(\tilde{T})} \right) B \leq O(\epsilon \log d)$.

Since $\mathbb{E}[\Delta^{2,2} \mid \tau(x_1) = s] \geq \Omega\left(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d}\right)$ we have

$$(85) \geq d^{0.01} \cdot \left| \left[-\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,2} \right]_{s,s} \right|,$$

and thus (85) \gg (86). Thus, for $\epsilon \ll \frac{1}{d}$, if the attention gap does not decrease to the level of $O\left(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d}\right)$, $[\mathbf{Q}_{4,4}]_{s,s}$ will start to dominantly increase while $[\mathbf{Q}_{4,3}]_{s,s}$ will not change too much. On the other hand, if the gap of attention holds, then $[\mathbf{Q}_{4,3}]_{s,s} \geq [\mathbf{Q}_{4,4}]_{s,s}$, we have

$$\begin{aligned} \epsilon_{\text{attn}}^{2,2} &= 1 - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} = \frac{O(1)}{O(1) + e^{[\mathbf{Q}_{4,4}^{(t)}]_{s,s}} + e^{[\mathbf{Q}_{4,3}^{(t)}]_{s,s}}} \\ &\geq \frac{O(1)}{O(1) + 2e^{[\mathbf{Q}_{4,3}^{(\tilde{T})}]_{s,s}}} \geq \frac{1}{2} \cdot 3\epsilon > \epsilon. \end{aligned}$$

This implies that, we can find some time between \tilde{T} and $T_{2,3,s}$, s.t., the gap will decrease to $O\left(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d}\right)$. We denote this time as T^* and stop the training.

D.5.7 At the End of the Training

Putting everything together, we have that at the end of the training, we have

Lemma D.45. *Given $s \in \tau(\mathcal{X})$, at $T^* = \tilde{O}\left(\frac{d^{(1-2\epsilon)C_B}}{\eta\epsilon}\right)$, if $\tilde{\Omega}(\sigma_0) \ll \epsilon \leq O(\frac{1}{d^{1.01}})$, we have*

- (a) *Attention convergence: $\epsilon_{\text{attn}}^{2,\ell} \leq O(\epsilon)$ for $\ell \in [2]$, and $\mathbb{E}[\Delta^{2,2} \mid \tau(x_1) = s] \leq O\left(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d}\right)$;*
- (b) *$\left[\mathbf{Q}_{4,p}^{(T_{2,3,s})}\right]_{s,s'} \geq \Omega(\log d)$ for $p \in \{3,4\}$ if $s = s' \in \tau(\mathcal{X})$, otherwise, $\left[\mathbf{Q}_{4,p}^{(T_{2,3,s})}\right]_{s,s'} \leq \tilde{O}(\frac{1}{d})$;*
- (c) *Loss convergence: $\sum_{\ell=1}^2 \text{Loss}_5^{2,\ell} \leq \frac{1}{\text{poly}d}$.*

D.6 Proof of Main Theorem

Theorem D.1 (Restatement of Theorem 4.1). *Under Assumptions 3.1, 3.2, 3.3, 3.4, 3.5 and 4.1, for some constant $0 < c^* < 1$, $n_y < m \ll \log^2 d$, the transformer model $F^{(T_1+T_2)}$ obtained by Algorithm 1 with learning rate $\eta = \frac{1}{\text{poly}(d)}$, and stage 1 and 2 iteration $T_1 = \tilde{O}\left(\frac{d}{\eta(\sigma_0)^{q-2}}\right)$, $T_2 = \tilde{O}\left(\frac{\text{poly}(d)}{\eta\sigma_0}\right)$ satisfies*

$$\text{Acc}_L\left(F^{(T_1+T_2)}\right) \geq 1 - \frac{1}{\text{poly}(d)}, \text{ for every } L \leq O(d^{c^*}), \quad (87)$$

i.e., $F^{(T_1+T_2)}$, which is trained for task \mathcal{T}^1 and \mathcal{T}^2 , generalizes to solve the tasks $\mathcal{T}^\ell, \ell \leq L$.

Proof. By Lemma D.45, at the end of Stage 2 training, we have $\left[\mathbf{Q}_{4,p}^{(T_{2,3,s})}\right]_{s,s} \geq \Omega(\log d)$ for all $p \in \{3,4\}$ and $s \in \tau(\mathcal{X})$.

Therefore, for task \mathcal{T}^L with $L \leq O(d^{c^*})$, where $0 < c^* < 1$ is a constant depending on the value of ϵ in Lemma D.45, we obtain

$$\epsilon_{\text{attn}}^{L,\ell} \leq \frac{O(1) \cdot L}{O(1) \cdot L + e^{\left[\mathbf{Q}_{4,3}^{(T_{2,3,s})}\right]_{s,s}} + e^{\left[\mathbf{Q}_{4,4}^{(T_{2,3,s})}\right]_{s,s}}} = o(1).$$

Moreover, we have

$$\Delta^{L,\ell} \leq \Delta^{2,1} = \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(T_{2,3,s})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(T_{2,3,s})} \leq o(1).$$

These together guarantee that

$$1 - \mathbf{logit}_{5,\tau(g_{\ell+1}(y_\ell))}(F^{(T^*)}, \mathbf{Z}^{(L,\ell)}) \leq \frac{O(1) \cdot d + e^{o(1)}}{O(1) \cdot d + e^{o(1)} + e^{\Omega(\log d)}} \leq \frac{1}{\text{poly}(d)},$$

which implies

$$\text{Acc}_L(F^{(T^*)}) \geq 1 - \frac{1}{\text{poly}(d)}.$$

□

E Learning the Attention Layer: Symmetry Case for Short-Length

In this part, we consider the scenario where the group operations form a symmetry group. Specifically, following Assumption 5.1, we assume that $\mathcal{Y} = \{0, 1, \dots, n_y - 1\}$, and let \mathcal{G} be the symmetry group for \mathcal{Y} , with order $|\mathcal{G}| = n_y! = \Theta(\text{polylog } d) \gg \frac{1}{\epsilon}$, where $n_y = \Theta\left(\frac{\log \log d}{\log \log \log d}\right)$. Similar to the simply transitive case, we restrict our analysis to updating only \mathbf{Q} , specifically the blocks $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$.

Throughout this section, we focus on the simple task \mathcal{T}^2 and analyze gradient descent updates with respect to the per-token loss $\text{Loss}_5^{2,2}$. Given $s \in \tau(\mathcal{X})$, let $\text{Loss}_{5,s}^{2,2} = -\mathbb{E}\left[\log p_F(\mathbf{Z}_{\text{ans},2,5} \mid \mathbf{Z}^{2,1}) \mid \tau(x_1) = s\right]$. Due to the symmetry of $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$ across $s \in \tau(\mathcal{X})$, we may, without loss of generality, focus on a particular $s \in \tau(\mathcal{X})$ and analyze the corresponding loss $\text{Loss}_{5,s}^{2,2}$ in what follows.

E.1 Gradient Computations

We start with the gradient computations for the attention layer.

Notations for gradient expressions. We first introduce some notations for the gradients of the attention layer. For $1 \leq \ell \leq L$, given $\mathbf{Z}^{L,\ell-1}$ and $\mathbf{k} \in \mathcal{I}^{L,\ell-1}$, define

$$\Xi_{\ell,i,\mathbf{k}}^L(\mathbf{Z}^{L,\ell-1}) \triangleq \sum_{j \in [d]} \mathcal{E}_{i,j}(\mathbf{Z}^{L,\ell-1}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{i,j,r}(\mathbf{Z}^{L,\ell-1})) \langle \mathbf{W}_{i,j,r}, \mathbf{Z}_{\mathbf{k}} \rangle, \quad i \in [5]. \quad (88)$$

For simplicity of notation, we will henceforth omit the dependence on $\mathbf{Z}^{L,\ell-1}$ in the notation of $\Xi_{\ell,i,\mathbf{k}}^L$ when it is clear from the context.

Fact E.1 (Gradients of \mathbf{Q}). For any $p, q \in [5]$, we have

$$\begin{aligned} -\nabla_{\mathbf{Q}_{p,q}} \text{Loss}^L &= \sum_{\ell=1}^L \sum_{i \in [5]} -\nabla_{\mathbf{Q}_{p,q}} \text{Loss}_i^{L,\ell}, \quad \text{where} \\ -\nabla_{\mathbf{Q}_{p,q}} \text{Loss}_i^{L,\ell} &= \\ &\mathbb{E} \left[\sum_{\mathbf{k} \in \mathcal{I}^{L,\ell-1}} \mathbf{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}} \cdot \left(\Xi_{\ell,i,\mathbf{k}}^L - \sum_{\mathbf{k}' \in \mathcal{I}^{L,\ell-1}} \mathbf{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}'} \Xi_{\ell,i,\mathbf{k}'}^L \right) \mathbf{Z}_{\text{ans},\ell-1,p} \mathbf{Z}_{\mathbf{k},q}^\top \right]. \end{aligned}$$

Lemma E.1 (Gradients of $\mathbf{Q}_{4,3}$). *Given $s \in \tau(\mathcal{X})$, for the diagonal entry $[\mathbf{Q}_{4,3}]_{s,s}$ of the block $\mathbf{Q}_{4,3}$, we have*

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},2} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_1)} \right]. \end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},1} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_1), s'=\tau(x_0)} \right]. \end{aligned}$$

Lemma E.2 (Gradients of $\mathbf{Q}_{4,4}$). *Given $s \in \tau(\mathcal{X})$, for the diagonal entry $[\mathbf{Q}_{4,4}]_{s,s}$ of the block $\mathbf{Q}_{4,4}$, we have*

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{ans},1} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_1)} \right]. \end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,4}]_{s,s'}$ with $s \neq s'$, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{ans},0} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_1), s'=\tau(x_0)} \right]. \end{aligned}$$

E.2 Some Useful Bounds for Gradients

In this subsection, we establish several useful bounds on the gradients of the attention layer, leveraging the feature structure of the MLP layer learned during stage 1.1. These bounds will be instrumental for the subsequent analysis.

Recall that for $j \in \tau(\mathcal{Y})$ and $y \in \mathcal{Y}$, the fiber $\text{Fiber}_{j,y}$, and the set of feature combinations for predicting $y = \tau^{-1}(j)$, are defined as

$$\text{Fiber}_{j,y} = \{g \in \mathcal{G} \mid \tau(g(y)) = j\}, \quad \mathfrak{F}_j = \{(\text{Fiber}_{j,y}, y) \in 2^{\mathcal{G}} \times \mathcal{Y}\}.$$

As established in Theorem C.1, at the end of stage 1.1, we have

- **Sparse activations:** For $j \in \tau(\mathcal{Y})$, let $\phi = (\text{Fiber}_{j,y}, y) \in \mathfrak{F}_j$, then there exists exactly one *activated* neuron $r \in [m]$ such that when $g_1 = g \in \text{Fiber}_{j,y}$, $y_0 = y$ happens:

$$\begin{aligned} \Lambda_{5,j,r}^{(T_{1,1})} &\geq B - O(\delta), & \left| C_\alpha V_{j,r}^{(t)}(y) - V_{j,r}^{(t)}(g) \right| &\leq O(\delta) \\ \Lambda_{5,j,r'}^{(T_{1,1})} &\leq O(d^{-\Omega(1)}) \quad \forall r' \neq r. \end{aligned}$$

for some $C_\alpha = \Theta(n_y)$;

- **Cancellation of incorrect features:** For $j \in \tau(\mathcal{Y})$, let $\phi = (\text{Fiber}_{j,y}, y) \in \mathfrak{F}_j$, and let the $r \in [m]$ be the activated neuron, then for any $g' \notin \text{Fiber}_{j,y}$, and any $y' \neq y \in \mathcal{Y}$, we have

$$\left| V_{j,r}^{(T_{1.1})}(g) + V_{j,r}^{(T_{1.1})}(y') \right| \leq O(\delta) \quad \text{and} \quad \left| V_{j,r}^{(T_{1.1})}(g') + V_{j,r}^{(T_{1.1})}(y) \right| \leq O(\delta).$$

In the analysis of FFN layer, we have a pair $\delta = (\delta_1, \delta_2)$; in the expressions above, some terms should use δ_1 and others δ_2 .

Notations for activated neurons. We denote by $r_{j,y}$ the unique activated neuron corresponding to $\phi = (\text{Fiber}_{j,y}, y) \in \mathfrak{F}_j$. For any $g \in \text{Fiber}_{j,y}$, we also write $r_{g,y}$ for the same neuron $r_{j,y}$. Note that $r_{g_1,y} = r_{g_2,y}$ for distinct $g_1, g_2 \in \text{Fiber}_{j,y}$. Moreover, define

$$\mathfrak{A} \triangleq \cup_{j \in \tau(\mathcal{Y})} \mathfrak{A}_j, \quad \text{where } \mathfrak{A}_j \triangleq \{r_{j,y} \mid y \in \mathcal{Y}\}.$$

In other words, \mathfrak{A} is the set of all activated neurons across all feature sets \mathfrak{F}_j for $j \in \tau(\mathcal{Y})$. Given $\mathbf{Z}^{L,\ell-1}$, letting $\widehat{\mathcal{G}}(\mathbf{Z}^{L,\ell-1}) = \cup \{g_{\ell'}\}_{\ell'=1}^L$ be the collection of all the chosen group elements in the predicate clauses.

Similarly $\widehat{\mathcal{Y}} = \cup \{y_{\ell'}\}_{\ell'=0}^{\ell-1}$. Then define $\widehat{\mathfrak{A}}_j(\mathbf{Z}^{L,\ell-1}) = \left\{ r_{g,y} \mid g \in \text{Fiber}_{j,y} \wedge \left(g \in \widehat{\mathcal{G}}(\mathbf{Z}^{L,\ell-1}) \vee y \in \widehat{\mathcal{Y}} \right) \right\}$.

For simplicity, we omit the dependence on $\mathbf{Z}^{L,\ell-1}$ in the notation of $\widehat{\mathfrak{A}}_j$ when it is clear from the context. Equipped with these notations, we can summarize the above properties in the following lemmas.

Lemma E.3 (Properties of target feature magnitude). *Given $j \in \tau(\mathcal{Y})$ and $y \in \mathcal{Y}$ then for $g \in \text{Fiber}_{j,y}$, the following properties hold.*

$$V_{j,r_{g,y}}(g) + V_{j,r_{g,y}}(y) \geq 2B - O(\delta), \quad \left| C_{\alpha} V_{j,r}^{(t)}(y) - V_{j,r}^{(t)}(g) \right| \leq O(\delta), \quad (89)$$

$$\left| V_{j,r_{g,y}}(g) + V_{j,r_{g,y}}(y') \right| \leq O(\delta), V_{j,r_{g,y}}(y') < 0 \quad \text{for all } y' \neq y, \quad (90)$$

$$\left| V_{j,r_{g,y}}(g') + V_{j,r_{g,y}}(y) \right| \leq O(\delta), V_{j,r_{g,y}}(g') < 0 \quad \text{for all } g' \notin \text{Fiber}_{j,y}. \quad (91)$$

$$\left| V_{j,r}(g) \right|, \left| V_{j,r}(y) \right| \leq O(\delta) \quad \text{for all } r \notin \mathfrak{A}_j. \quad (92)$$

Lemma E.4 (Properties of irrelevant magnitude). *If $(p, v) \notin \{2\} \times \mathcal{G} \cup \{5\} \times \mathcal{Y}$, or $j \notin \tau(\mathcal{Y})$, then for any $r \in [m]$, we have*

$$\left| \langle \mathbf{W}_{5,j,r,p}, e_v \rangle \right| \leq \widetilde{O}(\sigma_0). \quad (93)$$

The above lemmas give us some direct computations of the inner products between the weight matrices and input embedding vectors.

Lemma E.5. *Let $j \in \tau(\mathcal{Y})$ and $\ell \in [2]$. Then for any $r \in [m]$, the following holds:*

$$\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},\ell} \rangle = V_{j,r}(g_{\ell}) \pm \widetilde{O}(\sigma_0), \quad (94)$$

$$\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{ans},\ell-1} \rangle = V_{j,r}(y_{\ell-1}) \pm \widetilde{O}(\sigma_0). \quad (95)$$

Moreover, for $j \notin \tau(\mathcal{Y})$ and any $\mathbf{k} \in \mathcal{I}^{2,1}$ and $r \in [m]$, we have

$$\left| \langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathbf{k}} \rangle \right| = \widetilde{O}(\sigma_0). \quad (96)$$

Furthermore, we can establish some characterizations of the $\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1})$ quantities, which are crucial for the following analysis.

Lemma E.6 (Characterizations of Lambda). *Given $\mathbf{Z}^{2,1}$ with $\{\text{Attn}_{\text{ans},1 \rightarrow \mathbf{k}}\}_{\mathbf{k} \in \mathcal{I}^{2,1}}$, then*

(a) for $j \in \tau(\mathcal{Y})$, for activated neuron $r \in \mathfrak{A}_j$, we have

$$\Lambda_{5,j,r} = \sum_{\ell'=1}^2 \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell'} V_{j,r}(g_{\ell'}) + \sum_{\ell'=1}^2 \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},\ell'-1} V_{j,r}(y_{\ell'-1}) \pm \tilde{O}(\sigma_0).$$

(b) for $j \in \tau(\mathcal{Y})$, for any non-activated neuron $r \notin \mathfrak{A}_j$ we have

$$|\Lambda_{5,j,r}| \leq O(\delta).$$

(c) for $j \notin \tau(\mathcal{Y})$, for any $r \in [m]$, we have

$$|\Lambda_{5,j,r}| \leq \tilde{O}(\sigma_0).$$

A direct consequence of the above lemma is the following finer characterization of the activated neurons.

Lemma E.7. Given $j \in \tau(\mathcal{Y})$, for $r \in \mathfrak{A}_j \setminus \hat{\mathfrak{A}}_j$, we have $\mathbf{sReLU}'(\Lambda_{5,j,r}) = 0$.

Now we are ready to further derive the gradients of the attention layer starting from Lemmas E.1 and E.2 and the properties established above.

Lemma E.8 (Refined expression for the gradient of $\mathbf{Q}_{4,3}$). Given $s \in \tau(\mathcal{X})$, for the diagonal entry $[\mathbf{Q}_{4,3}]_{s,s}$ of the block $\mathbf{Q}_{4,3}$, letting $j_2 = \tau(g_2(y_1))$, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \right. \\ &\quad \left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \hat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot (V_{j_2,r}(g_2) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \right. \\ &\quad \left. - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \hat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot (V_{j,r}(g_2) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \right. \\ &\quad \left. \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_1)=s} \right]. \end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot \right. \\ &\quad \left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \hat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot (V_{j_2,r}(g_1) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \right. \\ &\quad \left. - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \hat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot (V_{j,r}(g_1) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \right. \\ &\quad \left. \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_1)=s, \tau(x_0)=s'} \right]. \end{aligned}$$

Lemma E.9 (Refined expression for the gradient of $\mathbf{Q}_{4,4}$). Given $s \in \tau(\mathcal{X})$, for the diagonal entry $[\mathbf{Q}_{4,4}]_{s,s}$ of the block $\mathbf{Q}_{4,4}$, letting $j_2 = \tau(g_2(y_1))$, we have

$$\left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} \right]_{s,s} = \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot \right.$$

$$\begin{aligned}
& \left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \left(V_{j_2,r}(y_1) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \right. \\
& - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \left(V_{j,r}(y_1) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \\
& \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_1)=s}.
\end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\begin{aligned}
\left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot \right. \\
& \left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \left(V_{j_2,r}(y_0) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \right. \\
& - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \left(V_{j,r}(y_0) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \\
& \left. \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_1)=s, \tau(x_0)=s'} \right].
\end{aligned}$$

Following the above calculations, we can further obtain the gradient summation of $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ as follows:

Lemma E.10 (Gradient sum of $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$). *Given $s \in \tau(\mathcal{X})$, letting $j_2 = \tau(g_2(y_1))$, we have*

$$\begin{aligned}
& \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} \right]_{s,s} \\
&= \mathbb{E} \left[\left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \right. \right. \right. \\
& \quad \left(-\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot V_{j_2,r}(y_0) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot V_{j_2,r}(g_1) \right. \\
& \quad \left. \left. + (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \\
& + \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \\
& \quad \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot V_{j,r}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot V_{j,r}(g_1) \right. \\
& \quad \left. \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \\
& \left. \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_1)=s} \right].
\end{aligned}$$

Notations for gradient decompositions. We shall define some useful notations to further simplify the expressions of gradient.

Lemma E.11. *For any $s \in \tau(\mathcal{X})$, we define the following notations for the gradient decompositions:*

1. for $[\mathbf{Q}_{4,3}]_{s,s}$ we have $[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2}]_{s,s} = \mathbb{E}[\mathcal{N}_{s,3,2,i} + \mathcal{N}_{s,3,2,ii} + \mathcal{N}_{s,3,2,iii}]$, where

$$\mathcal{N}_{s,3,2,i} = \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot (1 - \mathbf{logit}_{5,j_2}). \quad (97)$$

$$\left(\sum_{r \in \widehat{\mathfrak{A}}_{j_2}} \text{sReLU}'(\Lambda_{5,j_2,r}) \cdot \left(V_{j_2,r}(g_2) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s},$$

$$\begin{aligned} \mathcal{N}_{s,3,2,ii} = & -\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \\ & \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \text{sReLU}'(\Lambda_{5,j,r}) \cdot \left(V_{j,r}(g_2) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s}, \end{aligned} \quad (98)$$

$$\mathcal{N}_{s,3,2,iii} = \pm \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \mathbb{1}_{\tau(x_1)=s}. \quad (99)$$

2. for $[\mathbf{Q}_{4,4}]_{s,s}$, we have $[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2}]_{s,s} = \mathbb{E}[\mathcal{N}_{s,4,2,i} + \mathcal{N}_{s,4,2,ii} + \mathcal{N}_{s,4,2,iii}]$, where

$$\begin{aligned} \mathcal{N}_{s,4,2,i} = & \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot (1 - \mathbf{logit}_{5,j_2}) \cdot \\ & \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_2}} \text{sReLU}'(\Lambda_{5,j_2,r}) \cdot \left(V_{j_2,r}(y_1) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s}, \end{aligned} \quad (100)$$

$$\begin{aligned} \mathcal{N}_{s,4,2,ii} = & -\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \\ & \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \text{sReLU}'(\Lambda_{5,j,r}) \cdot \left(V_{j,r}(y_1) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s}, \end{aligned} \quad (101)$$

$$\mathcal{N}_{s,4,2,iii} = \pm \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \mathbb{1}_{\tau(x_1)=s}. \quad (102)$$

3. for the summation of $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$, we have $[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2}]_{s,s} + [-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2}]_{s,s} = \mathbb{E}[\mathcal{N}_{s,2,i} + \mathcal{N}_{s,2,ii} + \mathcal{N}_{s,2,iii}]$, where

$$\begin{aligned} \mathcal{N}_{s,2,i} = & (1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_2}} \text{sReLU}'(\Lambda_{5,j_2,r}) \cdot \right. \\ & \left(-\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot V_{j_2,r}(y_0) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot V_{j_2,r}(g_1) \right. \\ & \left. \left. + (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s} \\ \mathcal{N}_{s,2,ii} = & \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \text{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \\ & \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot V_{j,r}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot V_{j,r}(g_1) \right. \\ & \left. \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s} \\ \mathcal{N}_{s,2,iii} = & \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \tilde{O}(\sigma_0^q) \mathbb{1}_{\tau(x_1)=s}. \end{aligned}$$

Probabilistic events. We conclude this subsection by introducing several probabilistic events that will be used to simplify the characterization of activated neurons in the subsequent analysis. We first define some

events that may contribute non-trivially to the gradient of $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$:

$$\mathcal{E}_1 = \left\{ y_0 \neq y_1, g_1(y_{\ell-1}) \neq g_2(y_{\ell-1}), \text{ for all } \ell \in [2] \right\}, \quad (103)$$

$$\mathcal{E}_2 = \left\{ g_1, g_2 \in \text{Fiber}_{\tau(g_2(y_0)), y_0}, g_1(y_1) \neq g_2(y_1) \right\}, \quad (104)$$

$$\mathcal{E}_3 = \left\{ y_0 = y_1, g_1(y_1) \neq g_2(y_1) \right\}, \quad (105)$$

$$\mathcal{E}_4 = \left\{ y_0 \neq y_1, g_1(y_{\ell-1}) = g_2(y_{\ell-1}), \text{ for all } \ell \in [2] \right\}. \quad (106)$$

We first observe that event \mathcal{E}_1 occurs with high probability $1 - O(\frac{1}{n_c})$, and serves as the primary regime of interest. In contrast, events \mathcal{E}_2 and \mathcal{E}_3 each occur with probability $\Theta(1/n_y)$ and correspond to instances of initial prediction ambiguity, where certain incorrect classes may exhibit disproportionately large logits. Similarly, \mathcal{E}_4 corresponds to instances of initial prediction ambiguity, while occurs with probability $\Theta(1/n_y^2)$. Furthermore, we define the following events, which yield negligible gradient contributions to $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$, as they do not lead to significant confusion among the incorrect predictions:

$$\mathcal{E}_5 = \left\{ y_0 \neq y_1, g_1(y_1) = g_2(y_1), g_1(y_0) \neq g_2(y_0) \right\}, \quad (107)$$

$$\mathcal{E}_6 = \left\{ y_0 = y_1, g_1(y_1) = g_2(y_1) \right\}. \quad (108)$$

Here, \mathcal{E}_5 occurs with probability $\Theta(1/n_y)$, and \mathcal{E}_6 occurs with probability $\Theta(1/n_c^2)$. Together, the events $\cup_{i \in [6]} \mathcal{E}_i$ form a partition of the entire sample space.

E.3 Stage 1.2.1: Initial Growth of Q

We define the following notations:

$$\begin{aligned} \epsilon_{\text{attn}}^{L,\ell}(\mathbf{Z}^{L,\ell-1}) &= 1 - \mathbf{Attn}_{\text{ans},\ell-1 \rightarrow \text{pred},\ell}(\mathbf{Z}^{L,\ell-1}) - \mathbf{Attn}_{\text{ans},\ell-1 \rightarrow \text{ans},\ell-1}(\mathbf{Z}^{L,\ell-1}), \\ \Delta^{L,\ell}(\mathbf{Z}^{L,\ell-1}) &= \left| \mathbf{Attn}_{\text{ans},\ell-1 \rightarrow \text{pred},\ell}(\mathbf{Z}^{L,\ell-1}) - \mathbf{Attn}_{\text{ans},\ell-1 \rightarrow \text{ans},\ell-1}(\mathbf{Z}^{L,\ell-1}) \right|. \end{aligned}$$

We abbreviate $\epsilon_{\text{attn}}^{L,\ell}(\mathbf{Z}^{L,\ell-1})$ and $\Delta^{L,\ell}(\mathbf{Z}^{L,\ell-1})$ as $\epsilon_{\text{attn}}^{L,\ell}$ and $\Delta^{L,\ell}$ for simplicity. Since we only focus on the input $\mathbf{Z}^{2,1}$ in the following analysis, we will omit the notation related to the length, i.e., we use ϵ_{attn} and Δ when the context is clear.

Induction E.1. *Given $s \in \tau(\mathcal{X})$, let $T_{1,2,1,s}$ denote the first time that $\mathbb{E}[\epsilon_{\text{attn}} \mid \tau(x_1) = s] \leq 0.4$. For all iterations $t \leq T_{1,2,1,s}$, we have the following holds*

- (a) $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} + [\mathbf{Q}_{4,4}^{(t)}]_{s,s} \leq O(1)$ monotonically increases;
- (b) for $p \in \{3, 4\}$, for $s' \in \tau(\mathcal{X}) \neq s$, $\left| [\mathbf{Q}_{4,p}^{(t)}]_{s,s'} \right| \leq O\left(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,s'}}{d}\right)$; otherwise $[\mathbf{Q}_{4,p}^{(t)}]_{s,s'} = 0$;
- (c) for any sample $\mathbf{Z}^{2,1}$, we have $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \geq -O\left(\frac{\log \log d}{\log d}\right)$;
- (d) for any sample $\mathbf{Z}^{2,1}$, we have $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \leq c_1$ for some small constant $c_1 > 0$.

E.3.1 Attention and Lambda Preliminaries

Lemma E.12. *If Induction E.1 holds for all iterations $< t$, then we have*

1. $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \in [0.4 \pm \tilde{O}(\frac{1}{d}), 0.5]$;

$$2. |\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)}| \leq \tilde{O}\left(\frac{1}{d}\right).$$

Lemma E.13. *If Induction E.1 holds for all iterations $< t$, then given $\mathbf{Z}^{2,1} \in \mathcal{E}_1$,*

1. *for the prediction j_2 , we have*

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) \cdot 2B + \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \right) \cdot \frac{2B}{n_y} + O(\delta).$$

2. *for the prediction $j'_2 = \tau(g_2(y_0))$, we have*

$$\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta).$$

3. *for the prediction $\tau(g_1(y_0))$, we have*

$$\Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \right) \cdot \frac{2B}{n_y} + O(\delta).$$

furthermore, we have $\Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)} \leq \Lambda_{5,\tau(g_2(y_1)),r_{g_2 \cdot y_1}}^{(t)}$.

4. *for the prediction $\tau(g_1(y_1))$, we have*

$$\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \right) \cdot \frac{2B}{n_y} + \tilde{O}\left(\frac{B}{d}\right) + O(\delta).$$

5. *for other $j \in \tau(\mathcal{Y})$, if there are j and y , s.t., $g_1, g_2 \in \text{Fiber}_{j,y}$ (notice that $y \neq y_0, y_1$ for \mathcal{E}_1), then for such a j , $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$ cannot be activated, moreover, we have*

$$\Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

else, none of $r \in \widehat{\mathfrak{A}}_j$ can be activated.

Notice that for \mathcal{E}_1 , the neurons mentioned in 1-4 should be different neurons, while the predictions in 1 and 3, or 2 and 4 can be the same in some cases, e.g., $g_1(y_1) = g_2(y_0)$. In all cases, except for the neurons mentioned above, i.e., $\cup_{\ell, \ell' \in [2]} \{r_{g_\ell \cdot y_{\ell'-1}}\}$ (which may not be activated), all other neurons $r \in \cup_{\ell, \ell' \in [2]} \left(\widehat{\mathfrak{A}}_{\tau(g_\ell \cdot y_{\ell'-1})} \setminus \{r_{g_\ell \cdot y_{\ell'-1}}\} \right)$ cannot be activated.

Lemma E.14. *If Induction E.1 holds for all iterations $< t$, then given $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, we have*

1. *for the prediction j_2 , $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$ cannot be activated, moreover, we have*

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) \cdot 2B + \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \right) \cdot \frac{2B}{n_y} + O(\delta).$$

2. *for the prediction $j'_2 = \tau(g_2(y_0))$, $r \in \widehat{\mathfrak{A}}_{j'_2} \setminus \{r_{g_2 \cdot y_0}\}$ (notice that in this case $r_{g_2 \cdot y_0} = r_{g_1 \cdot y_0}$) cannot be activated, moreover, we have*

$$\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + O(\delta).$$

3. for the prediction $\tau(g_1(y_1))$, $r \in \widehat{\mathfrak{A}}_{\tau(g_1(y_1))} \setminus \{r_{g_1 \cdot y_1}\}$ cannot be activated, moreover, we have

$$\Lambda_{5, \tau(g_1(y_1)), r_{g_1 \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} \right) \cdot \frac{2B}{n_y} + O(\delta).$$

4. for other $j \in \tau(\mathcal{Y})$, if there are j and y , s.t., $g_1, g_2 \in \text{Fiber}_{j, y}$ (notice that $y \neq y_0, y_1$ for \mathcal{E}_2), then for such a j , $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$ cannot be activated, moreover, we have

$$\Lambda_{5, j, r_{g_2 \cdot y}}^{(t)} = \left(\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

else, none of $r \in \widehat{\mathfrak{A}}_j$ can be activated.

Lemma E.15. If Induction E.1 holds for all iterations $< t$, then given $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, we have

1. for the prediction j_2 , $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$ cannot be activated, moreover, we have

$$\Lambda_{5, j_2, r_{g_2 \cdot y_1}}^{(t)} = \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} \cdot 2B \pm O(\delta).$$

2. for the prediction $\tau(g_1(y_1))$, $r \in \widehat{\mathfrak{A}}_{\tau(g_1(y_1))} \setminus \{r_{g_1 \cdot y_1}\}$ cannot be activated, moreover, we have

$$\Lambda_{5, \tau(g_1(y_1)), r_{g_1 \cdot y_1}}^{(t)} = \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 1}^{(t)} \cdot 2B \pm O(\delta).$$

3. for other $j = g(y_1)$, where $g_1, g_2 \notin \text{Fiber}_{j, y_1}$, we have

$$\Lambda_{5, j, r_{g \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} \right) \cdot \frac{2B}{n_y} + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

moreover, if there exists y , s.t., $g_1, g_2 \in \text{Fiber}_{j, y}$ (notice that $y \neq y_1$ for \mathcal{E}_3), we have

$$\Lambda_{5, j, r_{g_2 \cdot y}}^{(t)} = \left(\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

besides, other $r \in \widehat{\mathfrak{A}}_j$ cannot be activated.

Lemma E.16. If Induction E.1 holds for all iterations $< t$, then given $\mathbf{Z}^{2,1} \in \mathcal{E}_4$, we have

1. for the prediction j_2 , $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$ cannot be activated, moreover, we have

$$\Lambda_{5, j_2, r_{g_2 \cdot y_1}}^{(t)} = \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} \cdot 2B + O(\delta).$$

2. for the prediction $\tau(g_2(y_0))$, $r \in \widehat{\mathfrak{A}}_{\tau(g_2(y_0))} \setminus \{r_{g_2 \cdot y_0}\}$ cannot be activated, moreover, we have

$$\Lambda_{5, j_2, r_{g_2 \cdot y_0}}^{(t)} = \left(\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} + \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 1}^{(t)} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \right) \cdot 2B + O(\delta).$$

3. for other $j \in \tau(\mathcal{Y})$, if there exist j and y , s.t., $g_1, g_2 \in \text{Fiber}_{j, y}$, then for such a j , $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$ cannot be activated, moreover, we have

$$\Lambda_{5, j, r_{g_2 \cdot y}}^{(t)} = \left(\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

else, none of $r \in \widehat{\mathfrak{A}}_j$ can be activated.

Lemma E.17. If Induction E.1 holds for all iterations $< t$, then given $\mathbf{Z}^{2,1} \in \mathcal{E}_5$, we have

1. for the prediction j_2 , $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$ cannot be activated, moreover, we have

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot 2B \pm O(\delta).$$

2. for the prediction $\tau(g_2(y_0))$, $r \in \widehat{\mathfrak{A}}_{\tau(g_2(y_0))} \setminus \{r_{g_2 \cdot y_0}\}$ cannot be activated, moreover, we have

$$\Lambda_{5,j_2,r_{g_2 \cdot y_0}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta).$$

3. for the prediction $\tau(g_1(y_0))$, $r \in \widehat{\mathfrak{A}}_{\tau(g_1(y_0))} \setminus \{r_{g_1 \cdot y_0}\}$ cannot be activated, moreover, we have

$$\begin{aligned} \Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)} &= \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B \\ &+ \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \right) \cdot \frac{2B}{n_y} + O(\delta). \end{aligned}$$

4. for other $j \in \tau(\mathcal{Y})$, if there are j and y , s.t., $g_1, g_2 \in \text{Fiber}_{j,y}$, then for such a j , $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$ cannot be activated, moreover, we have

$$\Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

else, none of $r \in \widehat{\mathfrak{A}}_j$ can be activated.

Lemma E.18. If Induction E.1 holds for all iterations $< t$, then given $\mathbf{Z}^{2,1} \in \mathcal{E}_6$, we have

1. for the prediction j_2 , $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$ cannot be activated, moreover, we have

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \right) \cdot 2B \pm O(\delta).$$

2. for other $j = g(y_1)$, where $g_2 \notin \text{Fiber}_{j,y_1}$, we have

$$\Lambda_{5,j,r_{g \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \right) \cdot \frac{2B}{n_y} + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

moreover, if there exists $y \neq y_1$, s.t., $g_1, g_2 \in \text{Fiber}_{j,y}$, we have

$$\Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

besides, other $r \in \widehat{\mathfrak{A}}_j$ cannot be activated.

E.3.2 Gradient Lemma

Lemma E.19. If Induction E.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, we have

$$\begin{aligned} &\left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} \\ &\geq \min \left\{ \Omega\left(\frac{1}{d \cdot n_y}\right), \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_1 \right] \right\} > 0. \end{aligned}$$

Proof. By Lemma E.10 and Lemma E.11, we have $\left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} \right]_{s,s} = \mathbb{E}[\mathcal{N}_{s,2,i} +$

$\mathcal{N}_{s,2,ii} + \mathcal{N}_{s,2,iii}$. Based on Lemma E.13-Lemma E.18, we can first directly bound the term $\mathcal{N}_{s,2,iii}^{(t)}$ as follows:

$$\mathbb{E} \left[\mathcal{N}_{s,2,iii}^{(t)} \right] \leq \tilde{O}(\sigma_0^q) = \frac{1}{\text{poly}d}.$$

In the following discussion, we focus on $\mathcal{N}_{s,2,i}^{(t)}$ and $\mathcal{N}_{s,2,ii}^{(t)}$, and consider two regimes for the gradient lower bound:

(i) $\mathbb{E}[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \mid \tau(x_1) = s] \leq \frac{\rho}{B}$, and (ii) $\mathbb{E}[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \mid \tau(x_1) = s] \geq \frac{\rho}{B}$. The proof strategy is to analyze $\mathcal{N}_{s,2,i}^{(t)}$ and $\mathcal{N}_{s,2,ii}^{(t)}$ under different event conditions in two regimes.

1. For regime (i),

(a) For $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, by Induction E.1 and Lemma E.13, we have an immediate logit upper bound for $j \in \tau(\mathcal{Y})$: $\mathbf{logit}_{5,j}^{(t)} \leq O(\frac{1}{d})$. Then by Lemma E.10, we have

$$\begin{aligned} & \mathbb{E} \left[\mathcal{N}_{s,2,i}^{(t)} \mathbf{1}_{\mathcal{E}_1} \right] \\ &= \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(\sum_{r \in \mathfrak{A}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}^{(t)}) \cdot \right. \right. \\ & \quad \left. \left(- \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r}(y_0) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j_2,r}(g_1) \right. \right. \\ & \quad \left. \left. + (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j_2,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_1} \right]. \end{aligned}$$

By Lemma E.13, we can obtain

$$\begin{aligned} & \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(\mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \cdot \right. \right. \\ & \quad \left. \left(- \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(y_0) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(g_1) \right. \right. \\ & \quad \left. \left. + (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_1} \right] \\ & \geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E} \left[\mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_1 \right]. \end{aligned} \tag{109}$$

On the other hand,

$$\begin{aligned} & - \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_1 \cdot y_0}}^{(t)}) \cdot \right. \\ & \quad \left. \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r_{g_1 \cdot y_0}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j_2,r_{g_1 \cdot y_0}}(g_1) \right) \right. \\ & \quad \left. \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_1} \mathbf{1}_{g_1(y_0)=g_2(y_1)} \right] \\ & \geq -O\left(\frac{B}{d \cdot n_y}\right) \cdot \mathbb{E} \left[\mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_1 \right], \end{aligned}$$

which implies that $\mathbb{E} \left[\mathcal{N}_{s,2,i}^{(t)} \mathbf{1}_{\mathcal{E}_1} \right] \geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E} \left[\mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_1 \right] \geq 0$. Moving to $\mathcal{N}_{s,2,ii}^{(t)}$, by Lemma E.10 and Lemma E.13, we have

$$\mathbb{E} \left[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_1} \right]$$

$$\begin{aligned}
&= \mathbb{E} \left[\sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)} \cdot \left(\sum_{r \in \widehat{\mathcal{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\
&\quad \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r}(g_1) \right. \\
&\quad \left. \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)} \pm \widetilde{O}(\sigma_0) \right) \pm \widetilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_1} \Big] \\
&\geq -O\left(\frac{n_y \cdot B}{d^2}\right).
\end{aligned}$$

(b) For $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, by Induction [E.1](#) and Lemma [E.14](#), we can also obtain a crude bound of logit: $\mathbf{logit}_{5,j}^{(t)} \leq O(\frac{1}{d})$ for $j \neq j'_2$. Then $\mathcal{N}_{s,2,i}^{(t)}$ can be bounded in the same way as [\(109\)](#), and we have

$$\mathbb{E} \left[\mathcal{N}_{s,2,i}^{(t)} \mathbb{1}_{\mathcal{E}_2} \right] \geq \Omega\left(\frac{B}{d \cdot n_y}\right) \cdot \mathbb{E} \left[\mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) | \tau(x_1) = s, \mathcal{E}_2 \right] \geq 0.$$

Moving to $\mathcal{N}_{s,2,ii}^{(t)}$,

- for $j = \tau(g_1(y_1))$, we have

$$\begin{aligned}
&= \mathbb{E} \left[\mathbf{logit}_{5,j}^{(t)} \left(\mathbf{sReLU}'(\Lambda_{5,j,r_{g_1 \cdot y_1}}) \cdot \right. \right. \\
&\quad \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r_{g_1 \cdot y_1}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r_{g_1 \cdot y_1}}(g_1) \right. \\
&\quad \left. \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)} \pm \widetilde{O}(\sigma_0) \right) \pm \widetilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_2} \Big] \\
&\geq -O\left(\frac{B}{n_y^2 \cdot d^2}\right).
\end{aligned}$$

where the last inequality is due to the cancellation of the term $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r_{g_1 \cdot y_1}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r_{g_1 \cdot y_1}}(g_1)$ and the fact that

$$\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \right) \cdot \frac{2B}{n_y} + O(\delta) \leq O\left(\frac{B}{n_y}\right).$$

- for $j = \tau(g_2(y))$ if there exists $y \neq y_0, y_1$ s.t., $g_1(y) = g_2(y)$
 - when $\mathbb{E}[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} | \tau(x_1) = s] \leq 0.01$, by Induction [E.1](#) and Lemma [E.14](#), we have $\mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \ll O(\frac{1}{d}) \cdot \mathbf{logit}_{5,j'_2}^{(t)}$ and $\mathbf{logit}_{5,j'_2}^{(t)} = \Omega(1)$. Then

$$\begin{aligned}
&\left| \mathbb{E} \left[\mathbf{logit}_{5,j}^{(t)} \cdot \left(\mathbf{sReLU}'(\Lambda_{5,j,r_{g_1 \cdot y}}) \cdot \right. \right. \right. \\
&\quad \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r_{g_1 \cdot y}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r_{g_1 \cdot y}}(g_1) \right. \\
&\quad \left. \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r_{g_1 \cdot y}}^{(t)} \pm \widetilde{O}(\sigma_0) \right) \pm \widetilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_2} \Big] \Big| \\
&\leq O\left(\frac{B}{n_y \cdot d^2}\right).
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&\mathbb{E} \left[\mathbf{logit}_{5,j'_2}^{(t)} \cdot \left(\mathbf{sReLU}'(\Lambda_{5,j'_2,r_{g_1 \cdot y_0}}) \cdot \right. \right. \\
&\quad \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j'_2,r_{g_1 \cdot y_0}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j'_2,r_{g_1 \cdot y_0}}(g_1) \right. \\
&\quad \left. \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j'_2,r_{g_1 \cdot y_0}}^{(t)} \pm \widetilde{O}(\sigma_0) \right) \pm \widetilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_2} \Big]
\end{aligned}$$

$$\begin{aligned}
& - \left((1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j'_2,r_{g_1 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \\
& \quad \pm \tilde{O}(\delta^q) \Big) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_1} \Big] \\
& \geq \Omega\left(\frac{B}{n_y \cdot d}\right),
\end{aligned}$$

where the last inequality follows from the fact that

$$\begin{aligned}
& \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j'_2,r_{g_1 \cdot y_0}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j'_2,r_{g_1 \cdot y_0}}(g_1) \\
& - \left((1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j'_2,r_{g_1 \cdot y_0}}^{(t)} \right) \\
& = \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot \left(2B - 2\Lambda_{5,j'_2,r_{g_1 \cdot y_0}}^{(t)} \right) \pm O(\delta) \geq \Omega(B).
\end{aligned}$$

– when $\mathbb{E}[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \mid \tau(x_1) = s] \geq 0.01$, by Lemma E.14, we have $\Lambda_{5,j,r_{g_1 \cdot y}}^{(t)}$ cannot be activated. Furthermore,

$$\begin{aligned}
& \mathbb{E} \left[\mathbf{logit}_{5,j'_2}^{(t)} \cdot \left(\mathbf{sReLU}'(\Lambda_{5,j'_2,r_{g_1 \cdot y_0}}^{(t)}) \cdot \right. \right. \\
& \quad \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j'_2,r_{g_1 \cdot y_0}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j'_2,r_{g_1 \cdot y_0}}(g_1) \right. \\
& \quad \left. \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j'_2,r_{g_1 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_1} \Big] \\
& \geq 0.
\end{aligned}$$

Putting the above discussion together, we have

$$\mathbb{E} \left[\mathcal{N}_{s,2,ii}^{(t)} \mathbb{1}_{\mathcal{E}_2} \right] \geq -O\left(\frac{B}{n_y^2 \cdot d^2}\right).$$

(c) For $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, by Induction E.1 and Lemma E.15, we can first have some facts of logit:

$$\begin{aligned}
1 - \mathbf{logit}_{5,j'_2}^{(t)} &= \Omega(1), \quad \mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} = \Omega(1) \\
\mathbf{logit}_{5,j}^{(t)} &\leq \frac{1}{\text{poly}d} \text{ for other } j.
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
& \mathbb{E} \left[\mathcal{N}_{s,2,i}^{(t)} \mathbb{1}_{\mathcal{E}_3} \right] \\
& = \mathbb{E} \left[(1 - \mathbf{logit}_{5,j'_2}^{(t)}) \cdot \left(- \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j'_2,r_{g_2 \cdot y_1}}(y_1) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j'_2,r_{g_2 \cdot y_1}}(g_1) \right. \right. \\
& \quad \left. \left. + (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j'_2,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_3} \right] \\
& \geq \Omega\left(\frac{B}{n_y \cdot d}\right).
\end{aligned}$$

Moving to $\mathcal{N}_{s,2,ii}^{(t)}$, we have

$$\mathbb{E} \left[\mathcal{N}_{s,2,ii}^{(t)} \mathbb{1}_{\mathcal{E}_3} \right]$$

$$\begin{aligned}
&\geq \mathbb{E} \left[\mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} \cdot \right. \\
&\quad \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{\tau(g_1(y_1)),r_{g_1 \cdot y_1}}(y_1) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{\tau(g_1(y_1)),r}(g_1) \right. \\
&\quad \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right. \\
&\quad \left. \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_3} \left. \right] - O\left(\frac{n_y \cdot B}{d \cdot \text{poly}d}\right) \\
&\geq \Omega\left(\frac{B}{n_y \cdot d}\right) - O\left(\frac{B}{d \cdot \text{poly}d}\right) = \Omega\left(\frac{B}{n_y \cdot d}\right).
\end{aligned}$$

- (d) For $\mathbf{Z}^{2,1} \in \mathcal{E}_4$, by comparing Lemma E.16 with Lemma E.14, we can directly bound $\mathcal{N}_{s,2,ii}^{(t)}$ in the same way as \mathcal{E}_2 , where the only difference is that we do not need to consider $\tau(g_1(y_1))$ for \mathcal{E}_4 . Thus $\mathbb{E} \left[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_4} \right] \geq 0$. Moreover, it is clear that

$$\begin{aligned}
&\mathbb{E} \left[\mathcal{N}_{s,2,i}^{(t)} \mathbf{1}_{\mathcal{E}_4} \right] \\
&= \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot (\mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \cdot \right. \\
&\quad \left(- \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(y_0) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(g_1) \right. \\
&\quad \left. + (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \left. \right) \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_4} \left. \right] \\
&\geq 0.
\end{aligned}$$

where the last inequality is due to the cancellation of $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(g_1)$, and the fact that

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot 2B + O(\delta) \geq \Omega(B).$$

- (e) For $\mathbf{Z}^{2,1} \in \mathcal{E}_5 \cup \mathcal{E}_6$, by Induction E.1, Lemma E.17 and Lemma E.18, we can derive the following logit condition: $1 - \mathbf{logit}_{5,j_2}^{(t)}, \mathbf{logit}_{5,j}^{(t)} \leq \frac{1}{\text{poly}d}$ for $j \neq j_2$. Hence, we can simply bound $\mathcal{N}_{s,2,i}^{(t)}$ and $\mathcal{N}_{s,2,ii}^{(t)}$ as follows:

$$\left| \mathbb{E} \left[\mathcal{N}_{s,2,i}^{(t)} \mathbf{1}_{\mathcal{E}_m} \right] \right|, \left| \mathbb{E} \left[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_m} \right] \right| \leq O\left(\frac{B}{dn_y} \cdot \frac{1}{\text{poly}d}\right) \text{ for } m \in \{5, 6\}.$$

Putting it all together, we have for the regime (1),

$$\left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} = \sum_{\kappa \in \{i, ii, iii\}} \sum_{i \in [6]} \mathbb{E} \left[\mathcal{N}_{s,2,\kappa}^{(t)} \mathbf{1}_{\mathcal{E}_i} \right] \geq \Omega\left(\frac{B}{d \cdot n_y}\right).$$

2. In regime (ii), the analysis follows a structure analogous to that of regime (i). The primary distinction lies in the fact that, under the main event \mathcal{E}_1 , the term $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}$ is guaranteed to remain within the linear regime, thereby serving as the dominant component driving the overall gradient.

- (a) For $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, by Induction E.1 and Lemma E.13, we have

$$\mathbb{E} \left[\mathcal{N}_{s,2,i}^{(t)} \mathbf{1}_{\mathcal{E}_1} \right]$$

$$\begin{aligned}
&= \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}^{(t)}) \cdot \right. \right. \\
&\quad \left(- \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r}(y_0) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j_2,r}(g_1) \right. \\
&\quad \left. \left. + (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j_2,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_1} \right].
\end{aligned}$$

By Lemma E.13, we can obtain

$$\begin{aligned}
&\mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(\left(- \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(y_0) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(g_1) \right. \right. \right. \\
&\quad \left. \left. + (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_1} \right] \quad (110) \\
&\geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_1 \right].
\end{aligned}$$

On the other hand,

$$\begin{aligned}
&- \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_1 \cdot y_0}}^{(t)}) \cdot \right. \\
&\quad \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r_{g_1 \cdot y_0}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j_2,r_{g_1 \cdot y_0}}(g_1) \right) \\
&\quad \left. \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_1} \mathbb{1}_{g_1(y_0)=g_2(y_1)} \right] \\
&\geq -O\left(\frac{B}{d \cdot n_y}\right) \cdot \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_1 \right],
\end{aligned}$$

which implies that $\mathbb{E} \left[\mathcal{N}_{s,2,i}^{(t)} \mathbb{1}_{\mathcal{E}_1} \right] \geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_1 \right] \geq 0$. Moving to $\mathcal{N}_{s,2,ii}^{(t)}$, by Lemma E.13, we have

- for $j = \tau(g_1(y_0))$, $r = r_{g_1 \cdot y_0}$

$$\begin{aligned}
&\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r}(g_1) \right. \\
&\quad \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \geq 0.
\end{aligned}$$

- for $j = \tau(g_1(y_1))$, $r = r_{g_1 \cdot y_1}$, due to the cancellation of $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r}(g_1)$, we have

$$\begin{aligned}
&\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r}(g_1) \right. \\
&\quad \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \\
&\geq - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)}.
\end{aligned}$$

- for $j = \tau(g_2(y_0))$, $r = r_{g_2 \cdot y_0}$

$$\begin{aligned}
&\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r}(g_1) \right. \\
&\quad \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \\
&\geq (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)}.
\end{aligned}$$

- for $j = g_1(y)$, where $\exists y \neq y_0, y_1$, s.t., $g_2(y) = g_1(y)$, we have

$$\begin{aligned} & \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r}(g_1) \right. \\ & \quad \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \\ & \geq -(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)}. \end{aligned}$$

Putting them together, and upper bound $\sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)}$ by $1 - \mathbf{logit}_{5,j_2}^{(t)}$, we have

$$\begin{aligned} & \mathbb{E} \left[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_1} \right] \\ &= \mathbb{E} \left[\sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)} \cdot \left(\sum_{r \in \mathfrak{A}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \cdot \right. \right. \\ & \quad \left. \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r}(g_1) \right. \right. \\ & \quad \left. \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_1} \right] \\ & \geq -\mathbb{E} \left[\left(1 - \mathbf{logit}_{5,j_2}^{(t)} \right) \cdot (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \right. \\ & \quad \cdot \max_{y \neq y_0, y_1} \left\{ \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}, \Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)} \right\} \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_1} \right], \end{aligned}$$

where the last inequality is due to the fact that $\Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)} = \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \pm \tilde{O}(\frac{B}{d \cdot n_y})$ for $y \neq y_0, y_1$ s.t., $g_1(y) = g_2(y)$. Notice that

$$\begin{aligned} & \max_{y \neq y_0, y_1} \left\{ \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}, \Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)} \right\} \leq \\ & \max \left\{ \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}, \Theta\left(\frac{1}{n_y}\right) \right\} 2B, \end{aligned}$$

while

$$\begin{aligned} (110) & \geq \mathbb{E} \left[\left(1 - \mathbf{logit}_{5,j_2}^{(t)} \right) \cdot \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot 2B \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_1} \right] \\ &= \frac{1}{2} \mathbb{E} \left[\left(1 - \mathbf{logit}_{5,j_2}^{(t)} \right) \cdot (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) 2B \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_1} \right]. \end{aligned}$$

Since by Induction E.1, $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \leq c_1 \ll \frac{1}{2}$, thus we have

$$\mathbb{E} \left[\mathcal{N}_{s,2,i}^{(t)} \mathbf{1}_{\mathcal{E}_1} \right] + \mathbb{E} \left[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_1} \right] \geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_1 \right].$$

- (b) For $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, $\mathcal{N}_{s,2,i}^{(t)}$ can be bounded in the same way as (110), and we have

$$\mathbb{E} \left[\mathcal{N}_{s,2,i}^{(t)} \mathbf{1}_{\mathcal{E}_2} \right] \geq \Omega\left(\frac{B}{d \cdot n_y}\right) \cdot \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_2 \right] \geq 0.$$

Moving to $\mathcal{N}_{s,2,ii}^{(t)}$,

- for $j = \tau(g_1(y_1))$, we have

$$\begin{aligned}
&= \mathbb{E} \left[\mathbf{logit}_{5,j}^{(t)} \left(\mathbf{sReLU}'(\Lambda_{5,j,r_{g_1 \cdot y_1}}) \cdot \right. \right. \\
&\quad \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r_{g_1 \cdot y_1}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r_{g_1 \cdot y_1}}(g_1) \right. \\
&\quad \left. \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_2} \Big] \\
&\geq -O\left(\frac{B}{n_y^2 \cdot d}\right) \cdot \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_2 \right].
\end{aligned} \tag{111}$$

where the last inequality is due to the cancellation of the term $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r_{g_1 \cdot y_1}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r_{g_1 \cdot y_1}}(g_1)$ and the fact that

$$\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \right) \cdot \frac{2B}{n_y} + O(\delta) \leq O\left(\frac{B}{n_y}\right).$$

- for $j'_2 = \tau(g_2(y_0)) = \tau(g_1(y_0))$, clearly, we have

$$\begin{aligned}
&\mathbb{E} \left[\mathbf{logit}_{5,j'_2}^{(t)} \cdot \left(\mathbf{sReLU}'(\Lambda_{5,j'_2,r_{g_1 \cdot y_0}}) \cdot \right. \right. \\
&\quad \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j'_2,r_{g_1 \cdot y_0}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j'_2,r_{g_1 \cdot y_0}}(g_1) \right. \\
&\quad \left. \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j'_2,r_{g_1 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \\
&\quad \left. \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_2} \right] \geq 0,
\end{aligned}$$

where the last inequality is due to the fact that

$$\begin{aligned}
&\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j'_2,r_{g_1 \cdot y_0}}(g_1) - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \\
&\quad - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j'_2,r_{g_1 \cdot y_0}}^{(t)} \\
&\geq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \\
&\quad \cdot \left(1 - 2 \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \right) 2B \\
&\geq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \left(1 - 2(c_1 + 0.25) \right) 2B \geq 0.
\end{aligned}$$

- for $j = \tau(g_2(y))$ if there exists $y \neq y_0, y_1$ s.t., $g_1(y) = g_2(y)$
 - when $\mathbb{E}[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \mid \tau(x_1) = s] \leq 0.01$, by Induction [E.1](#) and Lemma [E.14](#), we have $\mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \ll O\left(\frac{1}{d}\right) \cdot \mathbf{logit}_{5,j'_2}^{(t)}$. Then

$$\begin{aligned}
&\left| \mathbb{E} \left[\mathbf{logit}_{5,j}^{(t)} \cdot \left(\mathbf{sReLU}'(\Lambda_{5,j,r_{g_1 \cdot y}}) \cdot \right. \right. \right. \\
&\quad \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r_{g_1 \cdot y}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r_{g_1 \cdot y}}(g_1) \right. \\
&\quad \left. \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r_{g_1 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_2} \Big] \Big| \\
&\leq O\left(\frac{1}{d}\right) \cdot (111).
\end{aligned}$$

- when $\mathbb{E}[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \mid \tau(x_1) = s] \geq 0.01$, by Lemma E.14, we have $\Lambda_{5,j,r_{g_1 \cdot y}}^{(t)}$ cannot be activated.

Putting the above discussion together, we have

$$\mathbb{E}[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_1}] + \mathbb{E}[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_2}] \geq \Omega\left(\frac{B}{d \cdot n_y}\right) \cdot \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_2\right].$$

- (c) For $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, by Induction E.1 and Lemma E.15, we can first have the following logit bound:

$$\mathbf{logit}_{5,j}^{(t)} \leq \frac{1}{\text{poly}d} \cdot \left(1 - \mathbf{logit}_{5,j_2}^{(t)}\right) \text{ for } j \neq j_2, \tau(g_1(y_1)).$$

Therefore, we have

$$\begin{aligned} & \mathbb{E}\left[\mathcal{N}_{s,2,i}^{(t)} \mathbf{1}_{\mathcal{E}_3}\right] \\ &= \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(-\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(y_1) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(g_1)\right.\right. \\ & \quad \left.\left.+ (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \pm \tilde{O}(\delta^q)\right) \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_3}\right] \\ &\geq \Omega\left(\frac{B}{n_y \cdot d}\right) \cdot \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_3\right]. \end{aligned}$$

Moving to $\mathcal{N}_{s,2,ii}^{(t)}$, we have

$$\begin{aligned} & \mathbb{E}\left[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_3}\right] \\ &\geq \mathbb{E}\left[\mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} \cdot \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{\tau(g_1(y_1)),r_{g_1 \cdot y_1}}(y_1) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{\tau(g_1(y_1)),r}(g_1)\right.\right. \\ & \quad \left.- (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \pm \tilde{O}(\delta^q)\right) \\ & \quad \left. \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_3}\right] \\ &= O\left(\frac{n_y \cdot B}{n_y \cdot d \cdot \text{poly}d}\right) \cdot \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_3\right]. \end{aligned}$$

Thus,

$$\mathbb{E}[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_3}] + \mathbb{E}[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_4}] \geq \Omega\left(\frac{B}{n_y \cdot d}\right) \cdot \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_3\right].$$

- (d) For $\mathbf{Z}^{2,1} \in \mathcal{E}_4$, we can bound the gradient in a manner similar to regime (i), and obtain

$$\mathbb{E}[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_4}] + \mathbb{E}[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_4}] \geq 0.$$

- (e) For $\mathbf{Z}^{2,1} \in \mathcal{E}_5 \cup \mathcal{E}_6$, by Induction E.1, Lemma E.17 and Lemma E.18, we can derive the following logit condition: $1 - \mathbf{logit}_{5,j_2}^{(t)}, \mathbf{logit}_{5,j}^{(t)} \leq \frac{1}{\text{poly}d} \cdot \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_1\right]$ for $j \neq j_2$. Hence,

we can simply bound $\mathcal{N}_{s,2,i}^{(t)}$ and $\mathcal{N}_{s,2,ii}^{(t)}$ as follows:

$$\begin{aligned} & \left| \mathbb{E} \left[\mathcal{N}_{s,2,i}^{(t)} \mathbb{1}_{\mathcal{E}_m} \right] \right|, \left| \mathbb{E} \left[\mathcal{N}_{s,2,ii}^{(t)} \mathbb{1}_{\mathcal{E}_m} \right] \right| \\ & \leq O\left(\frac{B}{dn_y} \cdot \frac{1}{\text{poly}d}\right) \cdot \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_1 \right], \text{ for } m \in \{5, 6\}. \end{aligned}$$

Putting everything together, we have for the regmie (ii),

$$\begin{aligned} & \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} = \sum_{\kappa \in \{i, ii, iii\}} \sum_{i \in [6]} \mathbb{E} \left[\mathcal{N}_{s,2,\kappa}^{(t)} \mathbb{1}_{\mathcal{E}_i} \right] \\ & \geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_1 \right]. \end{aligned}$$

□

Lemma E.20. *If Induction E.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,p}]_{s,s'}$, $p \in \{3, 4\}$, $s' \neq s \in \tau(\mathcal{X})$, we have*

$$\left| \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s'} \right| \leq O\left(\frac{1}{d}\right) \left| \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} \right|.$$

E.3.3 Bounded Decrease of Attention to Related Context Clause

Lemma E.21. *If Induction E.1 holds for all iterations $< t$, then for any sample $\mathbf{Z}^{2,1}$, we have*

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \geq -O\left(\frac{\log d \log d}{\log d}\right).$$

Proof. Denote the first time that $\mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} | \tau(x_1) = s \right] \leq -\Omega\left(\frac{\log d \log d}{\log d}\right)$ as \tilde{T} .

Notice that $\left| [\mathbf{Q}_{4,p}^{(t)}]_{s,s'} \right| \leq O\left(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,s}}{d}\right)$ for $p \in \{3, 4\}$, thus for any sample $\mathbf{Z}^{2,1}$ satisfying $\tau(x_1) = s$, at time \tilde{T} , we have

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \leq -\Omega\left(\frac{\log d \log d}{\log d}\right).$$

Based on the the gradient compositions from Lemma E.11, we have $\left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s} = \mathbb{E} \left[\mathcal{N}_{s,3,2,i} + \mathcal{N}_{s,3,2,ii} + \mathcal{N}_{s,3,2,iii} \right]$, and we will discuss $\mathcal{N}_{s,3,2,\kappa}$ for $\kappa \in \{i, ii, iii\}$ on different samples $\mathbf{Z}^{2,1}$. Following the similar argument as Lemma E.19, we can first directly bound the term $\mathcal{N}_{s,3,2,iii}^{(\tilde{T})}$ as follows:

$$\mathbb{E} \left[\mathcal{N}_{s,3,2,iii}^{(\tilde{T})} \right] \leq \tilde{O}(\sigma_0^q) = \frac{1}{\text{poly}d}.$$

- for $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, at time \tilde{T} , since the neurons for predicting j_2 cannot be activated, and $\mathbf{logit}_{5,j}^{(\tilde{T})} \leq O(\frac{1}{d})$ for $j \neq j_2$, thus we can naively bound the gradient on the event \mathcal{E}_1 as follows:

$$\left| \sum_{\kappa \in \{i, ii\}} \mathbb{E} [\mathcal{N}_{s,3,2,\kappa}^{(\tilde{T})} \mathbb{1}_{\mathcal{E}_1}] \right| \leq O\left(\frac{n_y B}{d^2}\right) + \tilde{O}(\delta^q).$$

- for $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, similarly, $\Lambda_{5,j_2,r}^{(t)}$ is not activated, and thus we can focus on the term $\mathcal{N}_{s,3,2,ii}$, and specifically, the prediction of $j_2' = \tau(g_2(y_0))$. By (98) and Lemma E.14, we have

$$\begin{aligned} \mathbb{E} \left[\mathcal{N}_{s,3,2,ii}^{(\tilde{T})} \mathbf{1}_{\mathcal{E}_2} \right] &\geq -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \right. \\ &\quad \cdot \left(1 + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(\tilde{T})} \right) \mid \tau(x_1) = s \Big] \\ &\quad \cdot \mathbb{E} \left[\mathbf{logit}_{5,j_2'}^{(\tilde{T})} \mid \tau(x_1) = s, \mathcal{E}_2 \right] \frac{2B}{n_y \cdot d} \pm \tilde{O}(\sigma_0^q). \end{aligned} \quad (112)$$

- for $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, by Lemma E.15, we can mainly focus on the term $\mathcal{N}_{s,3,2,i}^{(\tilde{T})}$, and the prediction of $\tau(g_1(y_1))$ in $\mathcal{N}_{s,3,2,ii}^{(\tilde{T})}$ since $\mathbf{logit}_{5,\tau(g_1(y_1))}^{(\tilde{T})} = 1 - O(\frac{1}{\log d})$. Hence, we have

$$\begin{aligned} \mathbb{E} \left[\left(\mathcal{N}_{s,3,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,3,2,i}^{(\tilde{T})} \right) \mathbf{1}_{\mathcal{E}_3} \right] &\geq \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \right. \\ &\quad \cdot \left(1 + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \right) \mid \tau(x_1) = s \Big] \cdot \left(1 - O\left(\frac{1}{\log d}\right) \right) \frac{2B}{n_y \cdot d} \pm \tilde{O}(\sigma_0^q). \end{aligned} \quad (113)$$

- for $\mathbf{Z}^{2,1} \in \mathcal{E}_4$, the negative gradient can be bounded in the same way as (112), however, the probability of \mathcal{E}_4 is order-wise smaller than \mathcal{E}_2 and \mathcal{E}_3 , which can be neglected. Moreover, for $\mathbf{Z}^{2,1} \in \mathcal{E}_5 \cup \mathcal{E}_6$, the overall gradient is also negligible since $\mathbf{logit}_{5,j_2}^{(t)}$ is very close to 1.

Putting it all together, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,2} \right]_{s,s} &\geq -\mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \right. \\ &\quad \cdot \left(1 + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(\tilde{T})} \right) \mid \tau(x_1) = s \Big] \\ &\quad \cdot \mathbb{E} \left[\mathbf{logit}_{5,j_2'}^{(\tilde{T})} \mid \tau(x_1) = s, \mathcal{E}_2 \right] \frac{B}{n_y \cdot d} \\ &\quad + \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \right. \\ &\quad \cdot \left(1 + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \right) \mid \tau(x_1) = s \Big] \cdot \left(1 - O\left(\frac{1}{\log d}\right) \right) \frac{2B}{n_y \cdot d} \pm \tilde{O}(\sigma_0^q). \end{aligned}$$

Notice that

$$\begin{aligned} &\left(1 + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \right) \\ &\quad - \left(1 + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(\tilde{T})} \right) \\ &\quad = 2\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})}. \end{aligned}$$

If $2\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})} < c$ for some small constant $c > 0$, s.t., $\frac{cB}{\log d} < 1$, then $\Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(\tilde{T})} \leq cB$. Hence $\mathbf{logit}_{5,j_2'}^{(\tilde{T})} \leq \frac{1}{d^{1-\frac{cB}{\log d}}} = o(1)$, and (112) is dominated by the positive term (113). Else, clearly, (112) can be cancelled out by the positive term (113). Therefore,

$$\left[-\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,2} \right]_{s,s} \geq \Omega\left(\frac{B}{n_y \cdot d}\right).$$

As a consequence, together with the growth of $\mathbf{Q}_{4,3}^{(\tilde{T})} + \mathbf{Q}_{4,4}^{(\tilde{T})}$, and nearly no change of $[\mathbf{Q}_{4,p}^{(t)}]_{s,s'}$ in Lemma E.20, we have $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}$ must start to decrease and cannot be larger than $O(\frac{\log \log d}{\log d})$. \square

E.3.4 Attention gap is small

Lemma E.22. *If Induction E.1 holds for all iterations $< t$, then for any sample $\mathbf{Z}^{2,1}$, we have*

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \leq c_1,$$

where $c_1 > 0$ is a small constant.

Proof. Let \tilde{T} denote the first time $\mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \mid \tau(x_1) = s \right] \geq \frac{1.0005 \log d}{2B}$. Notice that $\left| [\mathbf{Q}_{4,p}^{(t)}]_{s,s'} \right| \leq O \left(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,s}}{d} \right)$ for $p \in \{3, 4\}$, thus for any sample $\mathbf{Z}^{2,1}$ satisfying $\tau(x_1) = s$, at time \tilde{T} , we have

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})} \geq \frac{1.0005 \log d}{2B}.$$

Based on Lemma E.11, we will discuss the gradients of $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ on different samples $\mathbf{Z}^{2,1}$. Since $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})}, \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(\tilde{T})}$ is guaranteed to be activated to the linear regime for $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, we only need to focus on the main event \mathcal{E}_1 .

From Lemma E.13, at time \tilde{T} , we have

$$\mathbf{logit}_{5,j'_2}^{(\tilde{T})} = \frac{e^{\frac{1.0005 \log d}{2B} \cdot 2B}}{e^{\frac{1.0005 \log d}{2B} \cdot 2B} + O(d)} \left(1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \right) = \left(1 - O(1/d^{0.0005}) \right) \left(1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \right).$$

Then by Lemma E.11, we can obtain the gradient on the event \mathcal{E}_1 as follows:

$$\begin{aligned} & \mathbb{E} \left[\left(\mathcal{N}_{s,3,2,i}^{(\tilde{T})} + \mathcal{N}_{s,3,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,3,2,iii}^{(\tilde{T})} \right) \mathbf{1}_{\mathcal{E}_1} \right] \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \cdot \left((1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) \cdot (V_{j_2,r_{g_2 \cdot y_1}}(g_2) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0)) \right. \right. \\ & \quad \left. \left. - \left(1 - O\left(\frac{1}{d^{0.0005}}\right) \right) (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot (V_{j'_2,r_{g_2 \cdot y_0}}(g_2) - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0)) \right. \right. \\ & \quad \left. \left. \pm \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_1} \right] \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \cdot \left(\left(1 - O\left(\frac{1}{d^{0.0005}}\right) \right) (1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) \cdot (\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(\tilde{T})} - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})}) \right. \right. \\ & \quad \left. \left. + O\left(\frac{1}{d^{0.0005}}\right) (1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) \cdot (V_{5,j_2,r_{g_2 \cdot y_1},2}(g_2) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})}) \right) \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_1} \right], \end{aligned}$$

on the other hand, we have

$$\begin{aligned} & \mathbb{E} \left[\left(\mathcal{N}_{s,4,2,i}^{(\tilde{T})} + \mathcal{N}_{s,4,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,4,2,iii}^{(\tilde{T})} \right) \mathbf{1}_{\mathcal{E}_1} \right] \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \cdot \left((1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) \cdot (V_{j_2,r_{g_2 \cdot y_1}}(y_1) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0)) \right) \right] \end{aligned}$$

$$\begin{aligned}
& - \left(1 - O\left(\frac{1}{d^{0.0005}}\right) \right) (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(V_{j_2', r_{g_2 \cdot y_0}}(y_1) - \Lambda_{5,j_2, r_{g_2 \cdot y_0}}^{(\tilde{T})} \pm \tilde{O}(\sigma_0) \right) \\
& \quad \pm \tilde{O}(\sigma_0^q) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_1} \Big] \\
& = \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(\tilde{T})} \cdot \left(\left(1 - O\left(\frac{1}{d^{0.0005}}\right) \right) (1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) \right. \right. \\
& \quad \cdot \left(-V_{j_2', r_{g_2 \cdot y_0}}(y_1) + \Lambda_{5,j_2', r_{g_2 \cdot y_0}}^{(\tilde{T})} - \Lambda_{5,j_2, r_{g_2 \cdot y_1}}^{(\tilde{T})} \right) \\
& \quad \left. \left. - O\left(\frac{1}{d^{0.0005}}\right) (1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) \cdot \Lambda_{5,j_2, r_{g_2 \cdot y_1}}^{(\tilde{T})} \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_1} \right].
\end{aligned}$$

Since at time \tilde{T} , we have $\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(\tilde{T})} \leq \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(\tilde{T})} \leq \frac{1.005 \log d}{2B}$, we have $\Lambda_{5,j_2', r_{g_2 \cdot y_0}}^{(\tilde{T})} - \Lambda_{5,j_2, r_{g_2 \cdot y_1}}^{(\tilde{T})} \ll -V_{j_2', r_{g_2 \cdot y_0}}(y_1)$. Therefore,

$$\left[-\nabla_{\mathbf{Q}_{4,4}^{(\tilde{T})}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,2} \right]_{s,s} \geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E} \left[1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} | \tau(x_1) = s \right],$$

which implies that $[\mathbf{Q}_{4,4}]_{s,s}$ will grow faster than $[\mathbf{Q}_{4,3}]_{s,s}$, and thus the attention gap cannot be further increased. \square

E.3.5 At the End of Stage 1.2.1

Lemma E.23. *For all iterations $t \leq T_{1,2,1,s} = O\left(\frac{n_y d}{\eta \log d}\right) + O\left(\frac{d^{\frac{(0.1+c_1/2)B}{\log d}}}{\eta \log d}\right)$, we have Induction E.1 holds, and at time $T_{1,2,1,s}$, we have*

- (a) $[\mathbf{Q}_{4,3}^{(T_{1,2,1,s})}]_{s,s}, [\mathbf{Q}_{4,4}^{(T_{1,2,1,s})}]_{s,s} \geq \Omega(1)$;
- (b) other $|[\mathbf{Q}_{4,p}^{(T_{1,2,1,s})}]_{s,s'}|$ for $p \in \{3, 4\}$, $s' \in \tau(\mathcal{X}) \neq s$ are at most $\tilde{O}(\frac{1}{d})$.

Proof. The existence of $T_{1,2,1,s}$ can be directly obtained by using Lemmas E.19 to E.22. Furthermore, $[\mathbf{Q}_{4,4}^{(T_{1,2,1,s})}]_{s,s} \geq \Omega(1)$ can be guaranteed since Lemma E.22 implies that $\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(T_{1,2,1,s})} \geq 0.3 - \frac{c_1}{2} \pm \tilde{O}(1/d) > 0.2$, which means that $[\mathbf{Q}_{4,4}^{(T_{1,2,1,s})}]_{s,s}$ should at least grow to a constant level compared to $|[\mathbf{Q}_{4,p}^{(T_{1,2,1,s})}]_{s,s'}| = \tilde{O}(1/d)$ with $p \in \{3, 4\}$.

We will handle $[\mathbf{Q}_{4,3}^{(T_{1,2,1,s})}]_{s,s}$ by means of a proof by contradiction. Suppose that $\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(T_{1,2,1,s})} \geq 0.4 - \tilde{c}$, where $\tilde{c} = \frac{\log d}{8B}$ is a sufficiently small constant. Then denote \tilde{T} the first time that $\mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} | \tau(x_1) = s \right] \geq 0.4 - 2\tilde{c}$. Notice that $\left| [\mathbf{Q}_{4,p}^{(t)}]_{s,s'} \right| \leq O\left(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,s}}{d}\right)$ for $p \in \{3, 4\}$, thus for any sample $\mathbf{Z}^{2,1}$ satisfying $\tau(x_1) = s$, at time \tilde{T} , we have

$$\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(\tilde{T})} \geq 0.4 - 2\tilde{c} \pm \tilde{O}(1/d).$$

- If $\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 0}^{(\tilde{T})} \geq 2\varrho$, then
 - for $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, $\Lambda_{5,j_2, r_{g_2 \cdot y_1}}^{(\tilde{T})}$ has already been activated to the linear regime. Furthermore, $\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 0}^{(\tilde{T})} \leq 0.2 + 2\tilde{c} - 0.2 \leq 2\tilde{c}$, which implies $1 - \mathbf{logit}_{5,j_2}^{(1)} = 1 - o(1)$. Thus, by Lemma E.11,

we have

$$\mathbb{E} \left[\left(\mathcal{N}_{s,3,2,i}^{(\tilde{T})} + \mathcal{N}_{s,3,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,3,2,iii}^{(\tilde{T})} \right) \mathbb{1}_{\mathcal{E}_1} \right] \geq \Omega\left(\frac{B}{d}\right),$$

while

$$\mathbb{E} \left[\left(\mathcal{N}_{s,4,2,i}^{(\tilde{T})} + \mathcal{N}_{s,4,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,4,2,iii}^{(\tilde{T})} \right) \mathbb{1}_{\mathcal{E}_1} \right] \leq -\Omega\left(\frac{B}{d}\right).$$

- for $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, $\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(\tilde{T})} = (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(\tilde{T})})2B \leq (0.4 + \frac{4}{3}\tilde{c} - 0.4 + 2\tilde{c})2B = \frac{5 \log d}{6}$. Thus $\mathbf{logit}_{5,j_2}^{(1)} = O(\frac{1}{d^{1/6}})$. Hence by Lemma E.11, we have

$$\mathbb{E} \left[\left(\mathcal{N}_{s,3,2,i}^{(\tilde{T})} + \mathcal{N}_{s,3,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,3,2,iii}^{(\tilde{T})} \right) \mathbb{1}_{\mathcal{E}_1} \right] \geq -O\left(\frac{B}{d^{7/6}n_y}\right),$$

while

$$\mathbb{E} \left[\left(\mathcal{N}_{s,4,2,i}^{(\tilde{T})} + \mathcal{N}_{s,4,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,4,2,iii}^{(\tilde{T})} \right) \mathbb{1}_{\mathcal{E}_1} \right] \leq -\Omega\left(\frac{B}{n_y d}\right).$$

- $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, we can use the following naive bounds for $p \in \{3, 4\}$:

$$|\mathbb{E} \left[\left(\mathcal{N}_{s,p,2,i}^{(\tilde{T})} + \mathcal{N}_{s,p,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,p,2,iii}^{(\tilde{T})} \right) \mathbb{1}_{\mathcal{E}_1} \right]| \leq O\left(\frac{B}{dn_y}\right).$$

Putting them together, combining with the fact that the gradient contributed by $\mathcal{E}_4 \cup \mathcal{E}_5 \cup \mathcal{E}_6$ is negligible, we can conclude that

$$\left[-\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,2} \right]_{s,s} \geq \Omega\left(\frac{B}{d}\right), \left[-\nabla_{\mathbf{Q}_{4,4}^{(\tilde{T})}} \text{Loss}_5^{2,2} \right]_{s,s} \leq -\Omega\left(\frac{B}{d}\right),$$

which implies $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})}$ cannot further increase above $0.4 - 2\tilde{c}$.

- If $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(\tilde{T})} < 2\varrho$, we shift our focus to the comparison between event \mathcal{E}_2 \mathcal{E}_3 ,
 - for $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, we have

$$\mathbb{E} \left[\left(\mathcal{N}_{s,3,2,i}^{(\tilde{T})} + \mathcal{N}_{s,3,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,3,2,iii}^{(\tilde{T})} \right) \mathbb{1}_{\mathcal{E}_1} \right] \geq -\tilde{O}(\sigma_0^q),$$

while

$$\mathbb{E} \left[\left(\mathcal{N}_{s,4,2,i}^{(\tilde{T})} + \mathcal{N}_{s,4,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,4,2,iii}^{(\tilde{T})} \right) \mathbb{1}_{\mathcal{E}_1} \right] \leq \tilde{O}(\sigma_0^q).$$

- for $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, similarly as previous case, we have

$$\mathbb{E} \left[\left(\mathcal{N}_{s,3,2,i}^{(\tilde{T})} + \mathcal{N}_{s,3,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,3,2,iii}^{(\tilde{T})} \right) \mathbb{1}_{\mathcal{E}_2} \right] \geq -O\left(\frac{B}{d^{7/6}n_y}\right),$$

while

$$\mathbb{E} \left[\left(\mathcal{N}_{s,4,2,i}^{(\tilde{T})} + \mathcal{N}_{s,4,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,4,2,iii}^{(\tilde{T})} \right) \mathbb{1}_{\mathcal{E}_2} \right] \leq O\left(\frac{B}{d^{7/6}n_y}\right).$$

- $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, $\left| \Lambda_{5,j'_2,r_{g_1 \cdot y_1}}^{(\tilde{T})} - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})} \right| \leq 4\rho B = o(1)$, hence, we have

$$\mathbb{E} \left[\left(\mathcal{N}_{s,3,2,i}^{(\tilde{T})} + \mathcal{N}_{s,3,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,3,2,iii}^{(\tilde{T})} \right) \mathbb{1}_{\mathcal{E}_3} \right] \geq \Omega\left(\frac{B}{dn_y}\right),$$

while

$$\mathbb{E} \left[\left(\mathcal{N}_{s,4,2,i}^{(\tilde{T})} + \mathcal{N}_{s,4,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,4,2,iii}^{(\tilde{T})} \right) \mathbf{1}_{\mathcal{E}_3} \right] \leq O\left(\frac{B \log \log d}{dn_y \cdot \log d}\right).$$

Putting them together, again we can conclude that

$$\left[-\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,2} \right]_{s,s} \geq \Omega\left(\frac{B}{n_y d}\right), \left[-\nabla_{\mathbf{Q}_{4,4}^{(\tilde{T})}} \text{Loss}_5^{2,2} \right]_{s,s} \leq O\left(\frac{B \log \log d}{dn_y \cdot \log d}\right),$$

which implies $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})}$ cannot further increase above $0.4 - 2\tilde{c}$.

Consequently, this leads to a contradiction, and $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(T_{1,2,1,s})} < 0.4 - \tilde{c}$, where $\tilde{c} = \frac{\log d}{8B}$. Then it would follow that $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(T_{1,2,1,s})} \geq \tilde{c}$, and thus $[\mathbf{Q}_{4,3}^{(T_{1,2,1,s})}]_{s,s} \geq \Omega(1)$. \square

E.4 Stage 1.2.2: Convergence with Small Wrong Attention

Recall that $\text{Loss}_{5,s}^{2,2} = -\mathbb{E} \left[\log p_F(\mathbf{Z}_{\text{ans},2,5} | \mathbf{Z}^{(2,1)}) | \tau(x_1) = s \right]$ for $s \in \tau(\mathcal{X})$.

Induction E.2. Given $s \in \tau(\mathcal{X})$, let $T_{1,2,2,s}$ denote the first time that $\text{Loss}_{5,s}^{2,2}$ decreases below $\Theta\left(e^{(-\frac{1}{2} + 3.01c_1) \cdot 2B}\right)$. For all iterations $T_{1,2,1,s} \leq t < T_{1,2,2,s}$, we have the following holds

- (a) $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} + [\mathbf{Q}_{4,4}^{(t)}]_{s,s}$ monotonically increases;
- (b) for $p \in \{3, 4\}$, for $j \in \tau(\mathcal{X}) \neq s$, $|\mathbf{Q}_{4,p}^{(t)}]_{s,j}| \leq O\left(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,j}}{d}\right)$;
- (c) for any sample $\mathbf{Z}^{2,1}$, we have $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \leq c_1$ for some sufficiently small constant $c_1 = \frac{1.005 \log d}{B} > 0$;
- (d) for any $\mathbf{Z}^{2,1}$, we have $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \leq \min \left\{ \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - c_2, 0 \right\}$, where $c_2 = \frac{\log d}{4B} > 0$ is some sufficiently small constant.

E.4.1 Attention and Lambda Preliminaries

Lemma E.24. If Induction E.2 holds for all iterations $[T_{1,2,1,s}, t)$, then for any sample $\mathbf{Z}^{2,1}$, we have

1. $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \in \left[\frac{4}{3}c_1, 0.4 \pm \tilde{O}\left(\frac{1}{d}\right) \right]$;
2. $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \geq \Omega(1)$;
3. $|\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)}| \leq \tilde{O}\left(\frac{1}{d}\right)$.

Proof. In the following, we focus on the main events \mathcal{E}_1 , \mathcal{E}_2 , and \mathcal{E}_3 , which correspond to cases where some confused wrong predictions occur. We denote

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)}, \quad \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} = c + \tilde{O}\left(\frac{1}{d}\right).$$

- If $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \geq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}$, then
 - If $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, we have

$$\log p_F(\mathbf{Z}_{\text{ans},2,5} | \mathbf{Z}^{(2,1)})$$

$$\begin{aligned}
&\leq \Theta(1) \cdot \max \left\{ e^{\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) 2B - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) 2B}, \right. \\
&\quad \left. e^{\log d - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) 2B} \right\} \\
&\leq \Theta \left(e^{2c_1 B - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) 2B} \right) \\
&\leq \Theta \left(e^{-2 \left(\frac{1}{2} - 2c - c_1 \right) B} \right),
\end{aligned}$$

where the last inequality follows from the fact that $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \geq \frac{1}{2}(1 - 2c) = \frac{1}{2} - c$.

– If $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, we have

$$\begin{aligned}
&\log p_F(\mathbf{Z}_{\text{ans},2,5} | \mathbf{Z}^{(2,1)}) \leq \Theta(1) \\
&\cdot \max \left\{ e^{\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) B - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) 2B}, \right. \\
&\quad \left. e^{\log d - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) 2B} \right\} \\
&\leq \Theta \left(e^{2(c+c_1)B - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) 2B} \right) \\
&\leq \Theta \left(e^{-2 \left(\frac{1}{2} - 3c - c_1 \right) B} \right).
\end{aligned}$$

– If $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, we have

$$\begin{aligned}
&\log p_F(\mathbf{Z}_{\text{ans},2,5} | \mathbf{Z}^{(2,1)}) \\
&\leq \Theta(1) \cdot \max \left\{ e^{\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \right) 2B}, e^{\log d - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} 2B} \right\} \\
&\leq \Theta(1) \cdot \max \left\{ e^{-2 \left(\frac{1}{2} - c_1 - c \right) B}, e^{-2 \left(\frac{1}{2} - 2c \right) B} \right\}.
\end{aligned}$$

• If $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \leq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}$, then

– If $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, we have

$$\begin{aligned}
&\log p_F(\mathbf{Z}_{\text{ans},2,5} | \mathbf{Z}^{(2,1)}) \leq \Theta(1) \cdot e^{\log d - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) 2B} \\
&\leq \Theta \left(e^{2c_1 B - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) 2B} \right) \\
&\leq \Theta \left(e^{-2 \left(\frac{1}{2} - \frac{5}{2}c - \frac{c_2}{2} - c_1 \right) B} \right),
\end{aligned}$$

where the last inequality uses the fact that $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \leq c - c_2$, implying $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \geq \frac{1}{2} - \frac{3}{2}c + \frac{c_2}{2}$.

– If $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, we have

$$\begin{aligned}
&\log p_F(\mathbf{Z}_{\text{ans},2,5} | \mathbf{Z}^{(2,1)}) \leq \Theta(1) \cdot \\
&\max \left\{ e^{\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) 2B - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) 2B}, \right.
\end{aligned}$$

$$\begin{aligned}
& e^{\log d - (\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 0}^{(t)})2B} \Big\} \\
& \leq \Theta \left(e^{\max\{c, c_1\}2B - (\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 0}^{(t)})2B} \right) \\
& \leq \Theta \left(e^{-2(\frac{1}{2} - 2c - \max\{c, c_1\})B} \right).
\end{aligned}$$

– If $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, we have

$$\begin{aligned}
& \log p_F(\mathbf{Z}_{\text{ans}, 2, 5} | \mathbf{Z}^{(2,1)}) \\
& \leq \Theta(1) \cdot \max \left\{ e^{(\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 1}^{(t)} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)})2B}, e^{\log d - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)}2B} \right\} \\
& \leq \Theta(1) \cdot \max \left\{ e^{-2(\frac{1}{2} - c_1 - c)B}, e^{-2(\frac{1}{2} - 2c)B} \right\}.
\end{aligned}$$

Putting all cases together, if $c \leq \frac{2}{3}c_1$, we must have

$$\text{Loss}_{5,s}^{2,2} \leq \Theta \left(e^{-2(\frac{1}{2} + 3 \cdot \frac{2c_1}{3} + c_1)B} \right) = \Theta \left(e^{(-\frac{1}{2} + 3c_1)2B} \right),$$

which contradicts the definition of $T_{1,2,2,s}$. Therefore, it must hold that $c \geq \frac{2}{3}c_1$ for all $t \leq T_{1,2,2,s}$. \square

Lemma E.25. *If Induction E.2 holds for all iterations $[T_{1,2,1,s}, t)$, then given $\mathbf{Z}^{2,1} \in \mathcal{E}_1$,*

1. *for the prediction j_2 , we have*

$$\Lambda_{5,j_2, r_{g_2 \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 0}^{(t)} \right) \cdot 2B + O\left(\frac{B}{n_y}\right) + O(\delta).$$

2. *for the prediction $j'_2 = \tau(g_2(y_0))$, we have*

$$\Lambda_{5,j'_2, r_{g_2 \cdot y_0}}^{(t)} = \left(\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta).$$

3. *for the prediction $\tau(g_1(y_0))$, $r_{g_1 \cdot y_0}$ cannot be activated.*

4. *for the prediction $\tau(g_1(y_1))$, we have*

$$\Lambda_{5, \tau(g_1(y_1)), r_{g_1 \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} \right) \cdot \frac{2B}{n_y} + \tilde{O}\left(\frac{B}{d}\right) + O(\delta).$$

5. *for other $j \in \tau(\mathcal{Y})$, if there are j and y , s.t., $g_1, g_2 \in \text{Fiber}_{j,y}$ (notice that $y \neq y_0, y_1$ for \mathcal{E}_1), then for such a j , $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$ cannot be activated, moreover, we have*

$$\Lambda_{5,j, r_{g_2 \cdot y}}^{(t)} = \left(\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, 2}^{(t)} - \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

else, none of $r \in \widehat{\mathfrak{A}}_j$ can be activated.

Notice that for \mathcal{E}_1 , the neurons mentioned in 1-4 should be different neurons, while the predictions in 1 and 3, or 2 and 4 can be the same in some cases, e.g., $g_1(y_1) = g_2(y_0)$. In all cases, except for the

neurons mentioned above, i.e., $\cup_{\ell, \ell' \in [2]} \{r_{g_\ell \cdot y_{\ell' - 1}}\}$ (which may not be activated), all other neurons $r \in \cup_{\ell, \ell' \in [2]} \left(\widehat{\mathfrak{A}}_{\tau(g_\ell \cdot y_{\ell' - 1})} \setminus \{r_{g_\ell \cdot y_{\ell' - 1}}\} \right)$ cannot be activated.

Lemma E.26. *If Induction E.2 holds for all iterations $t \in [T_{1,2,1,s}, T_{1,2,2,s})$, then given $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, we have*

1. *for the prediction j_2 , $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$ cannot be activated, moreover, we have*

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{n_y}\right) + O(\delta).$$

2. *for the prediction $j'_2 = \tau(g_2(y_0))$, $r \in \widehat{\mathfrak{A}}_{j'_2} \setminus \{r_{g_2 \cdot y_0}\}$ (notice that in this case $r_{g_2 \cdot y_0} = r_{g_1 \cdot y_0}$) cannot be activated, moreover, we have*

$$\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + O(\delta).$$

3. *for the prediction $\tau(g_1(y_1))$, $r \in \widehat{\mathfrak{A}}_{\tau(g_1(y_1))} \setminus \{r_{g_1 \cdot y_1}\}$ cannot be activated, moreover, we have*

$$\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \right) \cdot \frac{2B}{n_y} + O(\delta).$$

4. *for other $j \in \tau(\mathcal{Y})$, if there are j and y , s.t., $g_1, g_2 \in \text{Fiber}_{j,y}$ (notice that $y \neq y_0, y_1$ for \mathcal{E}_2), then for such a j , $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$ cannot be activated, moreover, we have*

$$\Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

else, none of $r \in \widehat{\mathfrak{A}}_j$ can be activated.

Lemma E.27. *If Induction E.2 holds for all iterations $t \in [T_{1,2,1,s}, T_{1,2,2,s})$, then given $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, we have*

1. *for the prediction j_2 , $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$ cannot be activated, moreover, we have*

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot 2B \pm O(\delta).$$

2. *for the prediction $\tau(g_1(y_1))$, $r \in \widehat{\mathfrak{A}}_{\tau(g_1(y_1))} \setminus \{r_{g_1 \cdot y_1}\}$ cannot be activated, moreover, we have*

$$\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} = \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot 2B \pm O(\delta).$$

3. *for other $j = g(y_1)$, where $g_1, g_2 \notin \text{Fiber}_{j,y_1}$, we have*

$$\Lambda_{5,j,r_{g \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \right) \cdot \frac{2B}{n_y} + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

moreover, if there exists y , s.t., $g_1, g_2 \in \text{Fiber}_{j,y}$ (notice that $y \neq y_1$ for \mathcal{E}_3), we have

$$\Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

besides, other $r \in \widehat{\mathfrak{A}}_j$ cannot be activated.

Lemma E.28. *If Induction E.2 holds for all iterations $t \in [T_{1,2,1,s}, T_{1,2,2,s})$, then given $\mathbf{Z}^{2,1} \in \mathcal{E}_4$, we have*

1. for the prediction j_2 , $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$ cannot be activated, moreover, we have

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot 2B + O(\delta).$$

2. for the prediction $\tau(g_2(y_0))$, $r \in \widehat{\mathfrak{A}}_{\tau(g_2(y_0))} \setminus \{r_{g_2 \cdot y_0}\}$ cannot be activated, moreover, we have

$$\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + O(\delta).$$

3. for other $j \in \tau(\mathcal{Y})$, if there are j and y , s.t., $g_1, g_2 \in \text{Fiber}_{j,y}$, then for such a j , $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$ cannot be activated, moreover, we have

$$\Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

else, none of $r \in \widehat{\mathfrak{A}}_j$ can be activated.

Lemma E.29. If Induction E.2 holds for all iterations $t \in [T_{1,2,1,s}, T_{1,2,2,s})$, then given $\mathbf{Z}^{2,1} \in \mathcal{E}_5$, we have

1. for the prediction j_2 , $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$ cannot be activated, moreover, we have

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot 2B \pm O(\delta).$$

2. for the prediction $\tau(g_2(y_0))$, $r \in \widehat{\mathfrak{A}}_{\tau(g_2(y_0))} \setminus \{r_{g_2 \cdot y_0}\}$ cannot be activated, moreover, we have

$$\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta).$$

3. for the prediction $\tau(g_1(y_0))$, none of $r \in \widehat{\mathfrak{A}}_{\tau(g_2(y_0))}$ can be activated.

4. for other $j \in \tau(\mathcal{Y})$, if there are j and y , s.t., $g_1, g_2 \in \text{Fiber}_{j,y}$, then for such a j , $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$ cannot be activated, moreover, we have

$$\Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

else, none of $r \in \widehat{\mathfrak{A}}_j$ can be activated.

Lemma E.30. If Induction E.2 holds for all iterations $t \in [T_{1,2,1,s}, T_{1,2,2,s})$, then given $\mathbf{Z}^{2,1} \in \mathcal{E}_6$, we have

1. for the prediction j_2 , $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2 \cdot y_1}\}$ cannot be activated, moreover, we have

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \right) \cdot 2B \pm O(\delta).$$

2. for other $j = g(y_1)$, where $g_2 \notin \text{Fiber}_{j,y_1}$, we have $r_{g \cdot y_1}$ cannot be activated, moreover, if there exists $y \neq y_1$, s.t., $g_1, g_2 \in \text{Fiber}_{j,y}$, we have

$$\Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + \tilde{O}\left(\frac{B}{d \cdot n_y}\right) + O(\delta);$$

besides, other $r \in \widehat{\mathfrak{A}}_j$ cannot be activated.

E.4.2 Gradient Lemma

Lemma E.31. *If Induction E.2 holds for all iterations $t \in [T_{1,2,1,s}, T_{1,2,2,s})$, given $s \in \tau(\mathcal{X})$, if $\mathbb{E}[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} | \tau(x_1) = s] \leq \frac{1}{2}$ we have*

$$\begin{aligned} & \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} \\ & \geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_1\right] + \Omega\left(\frac{B}{n_y d}\right) \cdot \sum_{m=2}^3 \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_m\right]. \end{aligned}$$

Proof. The analysis follows the similar idea as Lemma E.19, while Induction E.2 ensures that $\mathbb{E}[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} | \tau(x_1) = s] \geq \Omega(1) > \frac{\rho}{B}$, which falls into the regime (ii). Thus for the main event \mathcal{E}_1 , the term $\Lambda^{(t)}_{5,j_2,r_{g_2} \cdot y_1}$ is guaranteed to remain within the linear regime. We can first directly upper bound the term $\mathcal{N}_{s,2,iii}^{(t)}$ by $\tilde{O}(\sigma_0^q)$.

(a) For $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, by Induction E.2 and Lemma E.25, we can obtain

$$\begin{aligned} & \mathbb{E}\left[\mathcal{N}_{s,2,i}^{(t)} \mathbb{1}_{\mathcal{E}_1}\right] \\ & = \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(\left(-\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r_{g_2} \cdot y_1}(y_0) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j_2,r_{g_2} \cdot y_1}(g_1) \right. \right. \right. \\ & \quad \left. \left. + (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j_2,r_{g_2} \cdot y_1}^{(t)} \pm \tilde{O}(\sigma_0) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_1} \right] \quad (114) \\ & \geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_1\right]. \end{aligned}$$

Moving to $\mathcal{N}_{s,2,ii}^{(t)}$, by Lemma E.25, we have

- for $j = \tau(g_1(y_1))$, $r = r_{g_1 \cdot y_1}$, due to the cancellation of $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r}(g_1)$, we have

$$\begin{aligned} & \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r}(g_1) \right. \\ & \quad \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \\ & \geq -(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)}. \end{aligned}$$

- for $j = \tau(g_2(y_0))$, $r = r_{g_2 \cdot y_0}$

$$\begin{aligned} & \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r}(g_1) \right. \\ & \quad \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \\ & \geq -(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)}. \end{aligned}$$

- for $j = g_1(y)$, where $\exists y \neq y_0, y_1$, s.t., $g_2(y) = g_1(y)$, we have

$$\begin{aligned} & \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r}(g_1) \right. \\ & \quad \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \\ & \geq -(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)}. \end{aligned}$$

Putting them together, and upper bound $\sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)}$ by $1 - \mathbf{logit}_{5,j_2}^{(t)}$, we have

$$\begin{aligned}
& \mathbb{E} \left[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_1} \right] \\
&= \mathbb{E} \left[\sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\
&\quad \left. \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r}(g_1) \right. \right. \\
&\quad \left. \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_1} \right] \\
&\geq -\mathbb{E} \left[\left(1 - \mathbf{logit}_{5,j_2}^{(t)} \right) \cdot (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \right. \\
&\quad \left. \cdot \max_{y \neq y_0, y_1} \left\{ \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}, \Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)} \right\} \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_1} \right],
\end{aligned}$$

where the last inequality is due to the fact that $\Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)} = \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \pm \tilde{O}(\frac{B}{d \cdot n_y})$ for $y \neq y_0, y_1$ s.t., $g_1(y) = g_2(y)$. Notice that

$$\begin{aligned}
& \max_{y \neq y_0, y_1} \left\{ \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}, \Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)} \right\} \leq \\
& \max \left\{ \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}, \Theta\left(\frac{1}{n_y}\right) \right\} 2B,
\end{aligned}$$

while

$$\begin{aligned}
(114) &\geq \mathbb{E} \left[\left(1 - \mathbf{logit}_{5,j_2}^{(t)} \right) \cdot \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot 2B \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_1} \right] \\
&= \mathbb{E} \left[\left(1 - \mathbf{logit}_{5,j_2}^{(t)} \right) \cdot (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) B \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_1} \right].
\end{aligned}$$

Since by Induction E.2, $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \leq c_1 \ll \frac{1}{2}$, thus we have

$$\mathbb{E} \left[\mathcal{N}_{s,2,i}^{(t)} \mathbf{1}_{\mathcal{E}_1} \right] + \mathbb{E} \left[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_1} \right] \geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_1 \right].$$

(b) For $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, $\mathcal{N}_{s,2,i}^{(t)}$ can be bounded analogously to (114), yielding

$$\mathbb{E} \left[\mathcal{N}_{s,2,i}^{(t)} \mathbf{1}_{\mathcal{E}_2} \right] \geq \Omega\left(\frac{B}{d \cdot n_y}\right) \cdot \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_2 \right] \geq 0.$$

Moving to $\mathcal{N}_{s,2,ii}^{(t)}$,

- For $j = \tau(g_1(y_1))$, we obtain

$$\begin{aligned}
&= \mathbb{E} \left[\mathbf{logit}_{5,j}^{(t)} \left(\mathbf{sReLU}'(\Lambda_{5,j,r_{g_1 \cdot y_1}}) \cdot \right. \right. \\
&\quad \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r_{g_1 \cdot y_1}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r_{g_1 \cdot y_1}}(g_1) \right. \\
&\quad \left. \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j,r}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbf{1}_{\tau(x_1)=s} \mathbf{1}_{\mathcal{E}_2} \right]
\end{aligned} \tag{115}$$

$$\geq -O\left(\frac{B}{n_y^2 \cdot d}\right) \cdot \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_2\right],$$

where the inequality follows from the cancellation of the term

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r_{g_1 \cdot y_1}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r_{g_1 \cdot y_1}}(g_1),$$

together with the fact that

$$\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} = \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}\right) \cdot \frac{2B}{n_y} + O(\delta) \leq O\left(\frac{B}{n_y}\right).$$

- For $j'_2 = \tau(g_2(y_0)) = \tau(g_1(y_0))$, it is clear that

$$\begin{aligned} & \mathbb{E}\left[\mathbf{logit}_{5,j'_2}^{(t)} \left(\mathbf{sReLU}'(\Lambda_{5,j'_2,r_{g_1 \cdot y_0}}^{(t)}) \cdot \right.\right. \\ & \quad \left.\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j'_2,r_{g_1 \cdot y_0}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j'_2,r_{g_1 \cdot y_0}}(g_1) \right.\right. \\ & \quad \left.\left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)})\Lambda_{5,j'_2,r_{g_1 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0)\right) \pm \tilde{O}(\delta^q)\right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_2}\Big] \\ & \geq 0, \end{aligned}$$

where the last inequality is due to the fact that

$$\begin{aligned} & \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j'_2,r_{g_1 \cdot y_0}}(g_1) - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)})\Lambda_{5,j'_2,r_{g_1 \cdot y_0}}^{(t)} \\ & \geq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \left(1 - 2\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}\right)\right) 2B \\ & \geq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \left(1 - 2(c_1 + 0.2)\right) 2B \geq 0. \end{aligned}$$

- For $j = \tau(g_2(y))$ with some $y \neq y_0, y_1$ such that $g_1(y) = g_2(y)$, we distinguish two cases:

– If

$$\mathbb{E}[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \mid \tau(x_1) = s] \leq \frac{c_1}{4},$$

then by Induction E.2 and Lemma E.26, we have we have $\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} - \Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} \geq (-\frac{c_1}{6} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)})2B \geq (-\frac{c_1}{4} + \frac{3}{4}c_1)2B = c_1B$, which implies

$$\sum_{y \neq y_0, y_1} \mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \mathbb{1}_{g_1(y)=g_2(y)} \ll \tilde{O}\left(\frac{1}{d^{1/2}}\right) \cdot \mathbf{logit}_{5,j'_2}^{(t)}.$$

Consequently,

$$\begin{aligned} & \sum_{y \neq y_0, y_1} \left| \mathbb{E}\left[\mathbf{logit}_{5,j}^{(t)} \left(\mathbf{sReLU}'(\Lambda_{5,j,r_{g_1 \cdot y}}^{(t)}) \cdot \right.\right.\right. \\ & \quad \left.\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j,r_{g_1 \cdot y}}(y_0) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j,r_{g_1 \cdot y}}(g_1) \right.\right. \\ & \quad \left.\left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)})\Lambda_{5,j,r_{g_1 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0)\right) \pm \tilde{O}(\delta^q)\right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_2} \mathbb{1}_{g_1(y)=g_2(y)} \Big] \Big| \leq O\left(\frac{1}{d^{1/2}}\right) \cdot (115). \end{aligned}$$

– If

$$\mathbb{E}[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \mid \tau(x_1) = s] \geq \frac{c_1}{4},$$

then by Lemma E.26, the activation $\Lambda_{5,j,r_{g_1 \cdot y}}^{(t)}$ cannot be triggered.

Putting the above discussion together, we have

$$\mathbb{E}[\mathcal{N}_{s,2,ii}^{(t)} \mathbb{1}_{\mathcal{E}_1}] + \mathbb{E}[\mathcal{N}_{s,2,ii}^{(t)} \mathbb{1}_{\mathcal{E}_2}] \geq \Omega\left(\frac{B}{d \cdot n_y}\right) \cdot \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_2\right].$$

(c) For $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, the gradient admits an analysis analogous to that in Lemma E.19. In particular, by Induction E.2 and Lemma E.27, we first obtain the following logit bound:

$$\mathbf{logit}_{5,j}^{(t)} \leq \frac{1}{\text{poly}d} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}), \quad j \neq j_2, \tau(g_1(y_1)).$$

Therefore,

$$\begin{aligned} & \mathbb{E}[\mathcal{N}_{s,2,i}^{(t)} \mathbb{1}_{\mathcal{E}_3}] \\ &= \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(-\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(y_1) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(g_1)\right.\right. \\ & \quad \left.\left.+ (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)})\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \pm \tilde{O}(\delta^q)\right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_3}\right] \\ &\geq \Omega\left(\frac{B}{n_y \cdot d}\right) \cdot \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_3\right]. \end{aligned}$$

Moving to $\mathcal{N}_{s,2,ii}^{(t)}$, we have

$$\begin{aligned} & \mathbb{E}[\mathcal{N}_{s,2,ii}^{(t)} \mathbb{1}_{\mathcal{E}_3}] \\ &\geq \mathbb{E}\left[\mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} \cdot \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{\tau(g_1(y_1)),r_{g_1 \cdot y_1}}(y_1) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} \cdot V_{\tau(g_1(y_1)),r(g_1)}\right.\right. \\ & \quad \left.\left.- (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)})\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \pm \tilde{O}(\delta^q)\right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_3}\right] \\ &\quad - O\left(\frac{B}{d \cdot \text{poly}d}\right) \cdot \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_3\right]. \end{aligned}$$

Combining the two bounds, we conclude

$$\mathbb{E}[\mathcal{N}_{s,2,i}^{(t)} \mathbb{1}_{\mathcal{E}_3}] + \mathbb{E}[\mathcal{N}_{s,2,ii}^{(t)} \mathbb{1}_{\mathcal{E}_3}] \geq \Omega\left(\frac{B}{n_y \cdot d}\right) \cdot \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_3\right].$$

(d) For $\mathbf{Z}^{2,1} \in \mathcal{E}_4$, in analogy with Lemma E.19, we obtain the following crude bound:

$$\mathbb{E}[\mathcal{N}_{s,2,i}^{(t)} \mathbb{1}_{\mathcal{E}_4}] + \mathbb{E}[\mathcal{N}_{s,2,ii}^{(t)} \mathbb{1}_{\mathcal{E}_4}] \geq 0.$$

(e) For $\mathbf{Z}^{2,1} \in \mathcal{E}_5 \cup \mathcal{E}_6$, by Induction E.2, Lemma E.29 and Lemma E.30, we obtain the following logit condition:

$$1 - \mathbf{logit}_{5,j_2}^{(t)}, \mathbf{logit}_{5,j}^{(t)} \leq \frac{1}{\text{poly}d} \cdot \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_1\right], \quad j \neq j_2.$$

Hence, $\mathcal{N}_{s,2,i}^{(t)}$ and $\mathcal{N}_{s,2,ii}^{(t)}$ can be bounded as

$$\begin{aligned} \left| \mathbb{E}[\mathcal{N}_{s,2,i}^{(t)} \mathbf{1}_{\mathcal{E}_m}] \right|, \quad \left| \mathbb{E}[\mathcal{N}_{s,2,ii}^{(t)} \mathbf{1}_{\mathcal{E}_m}] \right| \\ \leq O\left(\frac{B}{dn_y} \cdot \frac{1}{\text{poly}d}\right) \cdot \mathbb{E}\left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_1\right], \quad m \in \{5, 6\}. \end{aligned}$$

Moreover, for $\mathbf{Z}^{2,1} \in \mathcal{E}_6$, if

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} > \frac{1}{2},$$

then $\mathcal{N}_{s,2,i}^{(t)} = 0$. Nevertheless, due to the above order-wise bound, this observation does not affect the overall analysis.

Putting everything together, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} &= \sum_{\kappa \in \{i, ii, iii\}} \sum_{m \in [6]} \mathbb{E} \left[\mathcal{N}_{s,2,\kappa}^{(t)} \mathbf{1}_{\mathcal{E}_m} \right] \\ &\geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_1 \right] + \Omega\left(\frac{B}{n_y d}\right) \cdot \sum_{m=2}^3 \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_m \right]. \end{aligned}$$

□

Lemma E.32. *If Induction E.2 holds for all iterations $t \in [T_{1,2,1,s}, T_{1,2,2,s})$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,p}]_{s,s'}$, $p \in \{3, 4\}$, $s' \neq s \in \tau(\mathcal{X})$, we have*

$$\left| \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s'} \right| \leq O\left(\frac{1}{d}\right) \left| \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} \right|.$$

E.4.3 Attention gap is small

Lemma E.33. *If Induction E.2 holds for all iterations $t \in [T_{1,2,1,s}, T_{1,2,2,s})$, then for any sample $\mathbf{Z}^{2,1}$, we have*

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \leq c_1,$$

where $c_1 > 0$ is a small constant.

The proof follows along the same lines as Lemma E.22, and hence the details are omitted for brevity.

Lemma E.34. *If Induction E.2 holds for all iterations $t \in [T_{1,2,1,s}, T_{1,2,2,s})$, then for any sample $\mathbf{Z}^{2,1}$, we have*

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \leq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - c_2,$$

where $c_2 = \frac{\log d}{4B} > 0$ is a small constant.

Proof. Let \tilde{T} denote the first time such that

$$\mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \mid \tau(x_1) = s \right] \leq \frac{\log d}{4.02B}.$$

Since $|\mathbf{Q}_{4,p}^{(t)}|_{s,s'} \leq O\left(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,s}}{d}\right)$ for $p \in \{3, 4\}$, it follows that for any sample $\mathbf{Z}^{2,1}$ with $\tau(x_1) = s$, at time \tilde{T}

we have

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \geq \frac{\log d}{4.02B} \pm \tilde{O}(1/d).$$

By Lemma E.24, we also have $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(\tilde{T})} \geq \Omega(1) \gg \varrho$. We then consider the following cases:

- If $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, then $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})}$ is already activated into the linear regime. Moreover, since $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})} \leq \frac{\log d}{4.02B} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} < 0$, it follows that $\Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(\tilde{T})}$ cannot be activated. Thus, by Lemma E.11,

$$\mathbb{E} \left[\left(\mathcal{N}_{s,3,2,i}^{(\tilde{T})} + \mathcal{N}_{s,3,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,3,2,iii}^{(\tilde{T})} \right) \mathbf{1}_{\mathcal{E}_1} \right] \geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E} \left[1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \mid \tau(x_1) = s, \mathcal{E}_1 \right],$$

while

$$\mathbb{E} \left[\left(\mathcal{N}_{s,4,2,i}^{(\tilde{T})} + \mathcal{N}_{s,4,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,4,2,iii}^{(\tilde{T})} \right) \mathbf{1}_{\mathcal{E}_1} \right] \leq -\Omega\left(\frac{B}{dn_y}\right) \cdot \mathbb{E} \left[1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \mid \tau(x_1) = s, \mathcal{E}_1 \right].$$

- If $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, then

$$\Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(\tilde{T})} = (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(\tilde{T})})2B \leq \frac{\log d}{2.01}.$$

Hence $\mathbf{logit}_{5,j_2}^{(1)} \leq O(d^{-1.01/2.01})(1 - \mathbf{logit}_{5,j_2}^{(t)})$. By Lemma E.11,

$$\mathbb{E} \left[\left(\mathcal{N}_{s,3,2,i}^{(\tilde{T})} + \mathcal{N}_{s,3,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,3,2,iii}^{(\tilde{T})} \right) \mathbf{1}_{\mathcal{E}_2} \right] \geq \Omega\left(\frac{B}{dn_y}\right) \cdot \mathbb{E} \left[1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \mid \tau(x_1) = s, \mathcal{E}_2 \right],$$

and

$$\mathbb{E} \left[\left(\mathcal{N}_{s,4,2,i}^{(\tilde{T})} + \mathcal{N}_{s,4,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,4,2,iii}^{(\tilde{T})} \right) \mathbf{1}_{\mathcal{E}_2} \right] \leq -\Omega\left(\frac{B}{dn_y}\right) \cdot \mathbb{E} \left[1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \mid \tau(x_1) = s, \mathcal{E}_2 \right].$$

- If $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, we can apply the crude bounds

$$\mathbb{E} \left[\left(\mathcal{N}_{s,3,2,i}^{(\tilde{T})} + \mathcal{N}_{s,3,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,3,2,iii}^{(\tilde{T})} \right) \mathbf{1}_{\mathcal{E}_3} \right] \geq 0,$$

and

$$\mathbb{E} \left[\left(\mathcal{N}_{s,4,2,i}^{(\tilde{T})} + \mathcal{N}_{s,4,2,ii}^{(\tilde{T})} + \mathcal{N}_{s,4,2,iii}^{(\tilde{T})} \right) \mathbf{1}_{\mathcal{E}_3} \right] \leq 0.$$

Combining the above, and noting that the gradient contribution from $\mathcal{E}_4 \cup \mathcal{E}_5 \cup \mathcal{E}_6$ is negligible, we conclude

$$\begin{aligned} [-\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,2}]_{s,s} &\geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E} \left[1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \mid \tau(x_1) = s, \mathcal{E}_1 \right] \\ &\quad + \Omega\left(\frac{B}{dn_y}\right) \cdot \mathbb{E} \left[1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \mid \tau(x_1) = s, \mathcal{E}_2 \right], \end{aligned} \tag{116}$$

$$\begin{aligned} [-\nabla_{\mathbf{Q}_{4,4}^{(\tilde{T})}} \text{Loss}_5^{2,2}]_{s,s} &\leq -\Omega\left(\frac{B}{d}\right) \cdot \mathbb{E} \left[1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \mid \tau(x_1) = s, \mathcal{E}_1 \right] \\ &\quad - \Omega\left(\frac{B}{dn_y}\right) \cdot \mathbb{E} \left[1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \mid \tau(x_1) = s, \mathcal{E}_2 \right]. \end{aligned} \tag{117}$$

Finally, observe that

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} = \frac{e^{[\mathbf{Q}_{4,3}^{(t)}]_{s,s}} - e^{[\mathbf{Q}_{4,4}^{(t)}]_{s,s}} - e^{\tilde{O}(1/d)}}{e^{[\mathbf{Q}_{4,3}^{(t)}]_{s,s}} + e^{[\mathbf{Q}_{4,4}^{(t)}]_{s,s}} + e^{\tilde{O}(1/d)}}.$$

Hence, (116) and (117) together imply that

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}$$

must decrease at time $\tilde{T} + 1$. □

E.4.4 At the End of Training

Lemma E.35 (At the end of stage 1.2.2). *Induction E.2 holds for all iterations $T_{1,2,1,s} < t \leq T_{1,2,2,s} = O\left(\frac{\text{poly}(d)}{\eta B}\right)$. At the end of training, we have*

(a) *Attention concentration: $\mathbb{E}[\epsilon_{\text{attn}} \mid \tau(x_1) = s] \leq 2.52c_1$ for some small constant $0 < c_1 = \frac{1.005 \log d}{2B}$;*

(b) *Loss convergence: $\text{Loss}_{5,s}^{2,2} \leq e^{(-\frac{1}{2} + 3.01c_1) \cdot 2B} = \frac{1}{\text{poly}d}$.*

Proof. Lemma E.24 and Lemma E.31 guarantee the continual growth of $[\mathbf{Q}_{4,3}^{(t)}]$ and $[\mathbf{Q}_{4,4}^{(t)}]$ until the attention weight ϵ_{attn} reaches $\frac{4c_1}{3}$. Combining Lemma E.31, Lemma E.32, Lemma E.33, and Lemma E.34, we conclude that there exists a stopping time

$$T_{1,2,2,s} = O\left(\frac{\text{poly}(d)}{\eta B}\right),$$

and that Induction E.2 holds for all $t \in [T_{1,2,1,s}, T_{1,2,2,s}]$.

Next, we establish an upper bound for ϵ_{attn} at the end of training. In this stage, it suffices to focus on the main event \mathcal{E}_1 . We denote

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)}, \quad \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} = c + \tilde{O}\left(\frac{1}{d}\right).$$

- If $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \geq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}$, then for $\mathbf{Z}^{2,1} \in \mathcal{E}_1$ we have

$$\begin{aligned} \log p_F(\mathbf{Z}_{\text{ans},2,5} | \mathbf{Z}^{(2,1)}) &\geq \Theta(1) \cdot e^{\log d - (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)})2B} \\ &\geq \Theta\left(e^{-\left(\frac{1}{2} + \frac{c_1}{2} - 2c - \frac{\log d}{B}\right)2B}\right), \end{aligned}$$

where the last inequality follows from

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \leq \frac{1}{2}(1 - 2c + c_1) = \frac{1+c_1}{2} - c.$$

- If $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \leq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}$, then for $\mathbf{Z}^{2,1} \in \mathcal{E}_1$ we have

$$\begin{aligned} \log p_F(\mathbf{Z}_{\text{ans},2,5} | \mathbf{Z}^{(2,1)}) &\geq \Theta(1) \cdot e^{\log d - (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)})2B} \\ &\geq \Theta\left(e^{-\left(\frac{1}{2} - 2c - \frac{\log d}{B}\right)2B}\right), \end{aligned}$$

where the last inequality follows from the fact that $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \leq \frac{1}{2} - c$.

Putting the above cases together, we obtain

$$\frac{1}{2} + \frac{c_1}{2} - 2c - \frac{\log d}{B} \geq \frac{1}{2} - 3.01c_1,$$

which implies

$$c \leq \frac{1}{2}(3.02c_1 + 0.5c_1 - c_1) = 1.26c_1.$$

□

F Recursive Learning the Attention Layer: Symmetry Case

In this section, we analyze the recursive learning dynamics of the attention layer in the symmetry case. To this end, we recall the definitions of the greedy data annotator, the bootstrapped LEGO data distribution, and the associated self-training loss.

Definition F.1 (Greedy data annotator). The greedy language model \hat{p}_F induced by a network F is defined as

$$\hat{p}_F(Z_{\text{ans}, L'+1} \mid Z^{L, L'}) = \begin{cases} 1, & \text{if } Z_{\text{ans}, L'+1} = \operatorname{argmax}_Z p_F(Z \mid Z^{L, L'}), \\ 0, & \text{otherwise.} \end{cases} \quad (118)$$

Definition F.2 (Bootstrapped LEGO distribution). We define $\mathcal{D}_F^{L, L'}$ as the LEGO distribution in Assumption 3.2, except that the answers $Z_{\text{ans}, \ell}$, $1 \leq \ell \leq L'$ are generated recursively by sampling $Z_{\text{ans}, \ell} \sim \hat{p}_F(\cdot \mid Z^{L, \ell-1})$ from the greedy language model \hat{p}_F .

Definition F.3 (Self-training loss). Given a (fixed) model \tilde{F} and length L , the self-training next-clause-prediction loss is defined by replacing $\mathcal{D}^{L, L'}$ in Definition 3.7 with the bootstrapped distribution $\mathcal{D}_F^{L, L'}$ from Definition F.2:

$$\text{Loss}_{\tilde{F}}^{L, L'}(F) \triangleq \mathbb{E}_{Z^{L, L'} \sim \mathcal{D}_F^{L, L'}} \left[-\log p_F(Z_{\text{ans}, L', i} \mid Z^{L, L'-1}) \right]. \quad (119)$$

Similarly, the per-token loss is defined by

$$\text{Loss}_{\tilde{F}, i}^{L, L'} = \mathbb{E}_{Z^{L, L'} \sim \mathcal{D}_F^{L, L'}} \left[-\log p_{F_i}(Z_{\text{ans}, L', i} \mid Z^{L, L'-1}) \right], \quad i \in [5].$$

At stage $k \geq 2$, let $F^{(T_{k-1})}$ denote the model obtained from the previous stage, trained on the task $\mathcal{T}^{L_{k-1}}$ with $L_{k-1} = 2^{k-1}$. Using the greedy annotator $\hat{p}_{F^{(T_{k-1})}}$, we construct the bootstrapped LEGO distribution $\mathcal{D}_{F^{(T_{k-1})}}^{L_k, L'_k}$ with total length $L_k = 2^k$. The corresponding self-training loss $\text{Loss}_{F^{(T_{k-1})}}^{L_k, L'_k}$ is then defined as in Definition F.3. In this stage, we focus on training via gradient descent on $\text{Loss}_{F^{(T_{k-1})}, i}^{L_k, L'_k}$ with sequence length $L' = 2$ and target token $i = 5$, initialized from the previous-stage model $F^{(T_{k-1})}$.

For notational simplicity, throughout the discussion we drop the subscript of the expectation operator \mathbb{E} when \mathbf{Z} is sampled from the ground-truth LEGO distribution; the subscript will be explicitly included only when the expectation is taken over the bootstrapped distribution. Moreover, we abbreviate the wrong attention error $\epsilon_{\text{attn}}^{L_k, 2}$ as $\epsilon_{\text{attn}}^{L_k}$ and the attention gap $\Delta^{L_k, 2}$ as Δ^{L_k} when the context is clear.

F.1 Preliminaries

We first present preliminaries on the recursive learning attention layer, including its gradient computations and some useful probability lemmas.

F.1.1 Gradient Computations

Fact F.1 (Gradients of \mathbf{Q}). Given $F^{(T_{k-1})}$ from the previous stage $\mathcal{T}^{L_{k-1}}$ with $L_{k-1} = 2^{k-1}$ for $k \geq 2$, for $(p, q) \in \{(4, 3), (4, 4)\}$, we have

$$-\nabla_{\mathbf{Q}_{p, q}} \text{Loss}_{F^{(T_{k-1})}, 5}^{L_k, 2} = -\nabla_{\mathbf{Q}_{p, q}} \text{Loss}_5^{L_k, 2}.$$

Remark F.1. This fact is straightforward to verify. Since the attention error $\epsilon_{\text{attn}}^{L_{k-1},2}$ has already been controlled at a small constant level in the previous stage for the model $F^{(T_{k-1})}$, when transitioning from length 2^{k-1} to 2^k , the incorrect attention produced by $F^{(T_{k-1})}$ cannot increase significantly. In particular, we have

$$\epsilon_{\text{attn}}^{L_k} \leq 2\epsilon_{\text{attn}}^{L_{k-1}},$$

which remains small. Moreover, the attention gap $\Delta^{L_k} = |\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}|$ is non-increasing due to the doubling of irrelevant clauses. As a result, the probability assigned by $F^{(T_{k-1})}$ to the correct prediction of $Z_{\text{ans},2,5}$ given $Z^{L_{k,1}}$ remains significantly larger than that of any incorrect prediction. Consequently, under the greedy data annotator, the self-training loss $\text{Loss}_{F^{(T_{k-1})},5}^{L_{k,2}}$ coincides with the original loss $\text{Loss}_5^{L_{k,2}}$ on the ground-truth LEGO distribution.

Lemma F.1 (Gradients of $\mathbf{Q}_{4,3}$). Given $F^{(T_{k-1})}$ from the previous stage and $s \in \tau(\mathcal{X})$, for the diagonal entry $[\mathbf{Q}_{4,3}]_{s,s}$ of the block $\mathbf{Q}_{4,3}$, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_{F^{(T_{k-1})},5}^{L_{k,2}} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},2} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_1)} \right]. \end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_{F^{(T_{k-1})},5}^{L_{k,2}} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{pred},1} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_1), s'=\tau(x_0)} \right]. \end{aligned}$$

Lemma F.2 (Gradients of $\mathbf{Q}_{4,4}$). Given $F^{(T_{k-1})}$ from the previous stage and $s \in \tau(\mathcal{X})$, for the diagonal entry $[\mathbf{Q}_{4,4}]_{s,s}$ of the block $\mathbf{Q}_{4,4}$, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{F^{(T_{k-1})},5}^{L_{k,2}} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{ans},1} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_1)} \right]. \end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,4}]_{s,s'}$ with $s \neq s'$, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{F^{(T_{k-1})},5}^{L_{k,2}} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot \left(\sum_{j \in [d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r \in [m]} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \right. \\ &\quad \left. \left. \left(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\text{ans},0} \rangle - \Lambda_{5,j,r} + b_{5,j,r} \right) \right) \mathbb{1}_{s=\tau(x_1), s'=\tau(x_0)} \right]. \end{aligned}$$

Notations for activated neurons. Recall that

$$\mathfrak{A} = \bigcup_{j \in \tau(\mathcal{Y})} \mathfrak{A}_j, \quad \text{where } \mathfrak{A}_j = \{r_{j,y} \mid y \in \mathcal{Y}\}.$$

Given $\mathbf{Z}^{L,\ell-1}$, let

$$\widehat{\mathcal{G}}(\mathbf{Z}^{L,\ell-1}) = \bigcup_{\ell'=1}^L \{g_{\ell'}\}$$

be the collection of all group elements chosen in the predicate clauses, and similarly define

$$\widehat{\mathcal{Y}} = \bigcup_{\ell'=0}^{\ell-1} \{y_{\ell'}\}.$$

We then introduce

$$\widehat{\mathfrak{A}}_j(\mathbf{Z}^{L,\ell-1}) = \left\{ r_{g,y} \mid g \in \text{Fiber}_{j,y}, g \in \widehat{\mathcal{G}}(\mathbf{Z}^{L,\ell-1}) \vee y \in \widehat{\mathcal{Y}} \right\}.$$

For simplicity, we omit the dependence on $\mathbf{Z}^{L,\ell-1}$ in the notation of $\widehat{\mathfrak{A}}_j$ when it is clear from context.

The above recalls the relevant notations from Appendix E.2. With these preliminaries in place, we now state the following lemma, which also relies on the feature-magnitude bounds established therein.

Lemma F.3 (Characterizations of Lambda). *Given $\mathbf{Z}^{L_k,1}$ with $\{\mathbf{Attn}_{\text{ans},1 \rightarrow \mathbf{k}}\}_{\mathbf{k} \in \mathcal{I}^{L_k,1}}$, then*

(a) *for $j \in \tau(\mathcal{Y})$, for activated neuron $r \in \mathfrak{A}_j$, we have*

$$\Lambda_{5,j,r} = \sum_{\ell'=1}^2 \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell'} V_{j,r}(g_{\ell'}) + \sum_{\ell'=1}^{L_k} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},\ell'-1} V_{j,r}(y_{\ell'-1}) \pm \widetilde{O}(\sigma_0).$$

(b) *for $j \in \tau(\mathcal{Y})$, for any non-activated neuron $r \notin \mathfrak{A}_j$ we have*

$$|\Lambda_{5,j,r}| \leq O(\delta).$$

(c) *for $j \notin \tau(\mathcal{Y})$, for any $r \in [m]$, we have*

$$|\Lambda_{5,j,r}| \leq \widetilde{O}(\sigma_0).$$

Lemma F.4. *Given $j \in \tau(\mathcal{Y})$, for $r \in \mathfrak{A}_j \setminus \widehat{\mathfrak{A}}_j$, we have $\mathbf{sReLU}'(\Lambda_{5,j,r}) = 0$.*

We are now ready to derive the gradients of the attention layer, building on Lemmas F.1 and F.2 and the properties established above.

Lemma F.5 (Refined expression for the gradient of $\mathbf{Q}_{4,3}$). *Given $F^{(T_{k-1})}$ from the previous stage and $s \in \tau(\mathcal{X})$, for the diagonal entry $[\mathbf{Q}_{4,3}]_{s,s}$ of the block $\mathbf{Q}_{4,3}$, letting $j_2 = \tau(g_2(y_1))$, we have*

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_{F^{(T_{k-1})},5}^{L_k,2} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \right. \\ &\quad \left((1 - \text{logit}_{5,j_2}) \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot (V_{j_2,r}(g_2) - \Lambda_{5,j_2,r} \pm \widetilde{O}(\sigma_0)) \right) \pm \widetilde{O}(\delta^q) \right) \\ &\quad - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \text{logit}_{5,j} \cdot \left(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot (V_{j,r}(g_2) - \Lambda_{5,j,r} \pm \widetilde{O}(\sigma_0)) \pm \widetilde{O}(\delta^q) \right) \\ &\quad \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \text{logit}_{5,j} \widetilde{O}(\sigma_0^q) \right) \mathbb{1}_{\tau(x_1)=s} \right]. \end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_{F^{(T_{k-1})},5}^{L_k,2} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot \right. \\ &\quad \left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \hat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot (V_{j_2,r}(g_1) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \right. \\ &\quad - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \hat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot (V_{j,r}(g_1) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \\ &\quad \left. \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_1)=s, \tau(x_0)=s'} \right]. \end{aligned}$$

Lemma F.6 (Refined expression for the gradient of $\mathbf{Q}_{4,4}$). *Given $F^{(T_{k-1})}$ from the previous stage and $s \in \tau(\mathcal{X})$, for the diagonal entry $[\mathbf{Q}_{4,4}]_{s,s}$ of the block $\mathbf{Q}_{4,4}$, letting $j_2 = \tau(g_2(y_1))$, we have*

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{F^{(T_{k-1})},5}^{L_k,2} \right]_{s,s} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot \right. \\ &\quad \left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \hat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot (V_{j_2,r}(y_1) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \right. \\ &\quad - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \hat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot (V_{j,r}(y_1) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \\ &\quad \left. \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_1)=s} \right]. \end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\begin{aligned} \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{F^{(T_{k-1})},5}^{L_k,2} \right]_{s,s'} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot \right. \\ &\quad \left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \hat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot (V_{j_2,r}(y_0) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \right. \\ &\quad - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \hat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot (V_{j,r}(y_0) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \\ &\quad \left. \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \right) \mathbf{1}_{\tau(x_1)=s, \tau(x_0)=s'} \right]. \end{aligned}$$

Following the above calculations, we can further obtain the gradient summation of $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ as follows:

Lemma F.7 (Gradient sum of $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$). *Given $F^{(T_{k-1})}$ from the previous stage and $s \in \tau(\mathcal{X})$, letting $j_2 = \tau(g_2(y_1))$, we have*

$$\begin{aligned} &\left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_{F^{(T_{k-1})},5}^{L_k,2} \right]_{s,s} + \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{F^{(T_{k-1})},5}^{L_k,2} \right]_{s,s} \\ &= \mathbb{E} \left[\left((1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \hat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \right. \right. \right. \end{aligned}$$

$$\begin{aligned}
& \left(-\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot V_{j_2,r}(y_0) - \sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell} \cdot V_{j_2,r}(g_\ell) \right. \\
& \quad \left. + (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \\
& + \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \hat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \\
& \quad \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot V_{j,r}(y_0) + \sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell} \cdot V_{j_2,r}(g_\ell) \right. \\
& \quad \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \\
& \quad \left. \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \tilde{O}(\sigma_0^q) \right) \mathbb{1}_{\tau(x_1)=s} \Big].
\end{aligned}$$

Notations for gradient decompositions. Next we define some useful notations to further simplify the expressions of gradient.

Lemma F.8. *Given $F^{(T_k-1)}$ from the previous stage and $s \in \tau(\mathcal{X})$, we define the following notations for the gradient decompositions:*

1. for $[\mathbf{Q}_{4,3}]_{s,s}$ we have $[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_{F^{(T_k-1)},5}^{L_k,2}]_{s,s} = \mathbb{E}[\mathcal{N}_{s,3,L_k,i} + \mathcal{N}_{s,3,L_k,ii} + \mathcal{N}_{s,3,L_k,iii}]$, where

$$\begin{aligned}
\mathcal{N}_{s,3,L_k,i} &= \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot (1 - \mathbf{logit}_{5,j_2}) \cdot \\
& \quad \left(\sum_{r \in \hat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot (V_{j_2,r}(g_2) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s},
\end{aligned} \tag{120}$$

$$\begin{aligned}
\mathcal{N}_{s,3,L_k,ii} &= -\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \\
& \quad \left(\sum_{r \in \hat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot (V_{j,r}(g_2) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s},
\end{aligned} \tag{121}$$

$$\mathcal{N}_{s,3,L_k,iii} = \pm \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \mathbb{1}_{\tau(x_1)=s}. \tag{122}$$

2. for $[\mathbf{Q}_{4,4}]_{s,s}$, we have $[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{F^{(T_k-1)},5}^{L_k,2}]_{s,s} = \mathbb{E}[\mathcal{N}_{s,4,L_k,i} + \mathcal{N}_{s,4,L_k,ii} + \mathcal{N}_{s,4,L_k,iii}]$, where

$$\begin{aligned}
\mathcal{N}_{s,4,L_k,i} &= \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot (1 - \mathbf{logit}_{5,j_2}) \cdot \\
& \quad \left(\sum_{r \in \hat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot (V_{j_2,r}(y_1) - \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s},
\end{aligned} \tag{123}$$

$$\begin{aligned}
\mathcal{N}_{s,4,L_k,ii} &= -\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \\
& \quad \left(\sum_{r \in \hat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot (V_{j,r}(y_1) - \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0)) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s},
\end{aligned} \tag{124}$$

$$\mathcal{N}_{s,4,L_k,iii} = \pm \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \tilde{O}(\sigma_0^q) \mathbb{1}_{\tau(x_1)=s}. \tag{125}$$

3. for the summation of $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$, we have $[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2}]_{s,s} + [-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2}]_{s,s} = \mathbb{E}[\mathcal{N}_{s,2,i} + \mathcal{N}_{s,2,ii} + \mathcal{N}_{s,2,iii}]$, where

$$\begin{aligned}
\mathcal{N}_{s,L_k,i} &= (1 - \mathbf{logit}_{5,j_2}) \cdot \left(\sum_{r \in \mathfrak{A}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \right. \\
&\quad \left(-\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot V_{j_2,r}(y_0) - \sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell} \cdot V_{j_2,r}(g_\ell) \right. \\
&\quad \left. \left. + (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \Lambda_{5,j_2,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s}, \\
\mathcal{N}_{s,L_k,ii} &= \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \left(\sum_{r \in \mathfrak{A}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \right. \\
&\quad \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot V_{j,r}(y_0) + \sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell} \cdot V_{j,r}(g_\ell) \right. \\
&\quad \left. \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \Lambda_{5,j,r} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right) \mathbb{1}_{\tau(x_1)=s}, \\
\mathcal{N}_{s,L_k,iii} &= \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \tilde{O}(\sigma_0^q) \mathbb{1}_{\tau(x_1)=s}.
\end{aligned}$$

F.1.2 Probabilistic Events

We introduce the following events:

$$\begin{aligned}
\mathcal{E}_{L_k,1} &= \left\{ \frac{1}{L_k} \sum_{\ell \in [L_k] \setminus \{2\}} \mathbb{1}\{g_\ell(y_1) = g_2(y_1)\} \leq U_k \right\}, \\
\mathcal{E}_{L_k,2} &= \{y_0 = y_1\}, \\
\mathcal{E}_{L_k,3} &= \left\{ \max_{y \in \mathcal{Y}} \sum_{\ell \in [L_k] \setminus \{2\}} \mathbb{1}\{g_\ell(y) = g_2(y)\} \leq W_k \right\}.
\end{aligned}$$

We set

$$U_k = \frac{1}{L_k} \left\lceil 1 + \frac{L_k}{8} \right\rceil = \Theta(1), \quad W_k = \begin{cases} \Theta\left(\frac{\log n_y}{\log\left(\frac{4n_y}{L_k} \log n_y\right)}\right), & L_k \leq n_y \log n_y, \\ \Theta\left(\frac{L_k}{n_y}\right), & L_k \geq n_y \log n_y. \end{cases}$$

By standard balls-into-bins tail bounds and maximum-load estimates (e.g., [RS98]), we obtain

$$\Pr(\mathbf{Z}^{L_k,1} \notin \mathcal{E}_{L_k,1}) = \begin{cases} O(n_y^{-U_k L_k + 1}), & L_k \ll n_y, \\ \exp(-\Theta(n_y \log n_y)), & L_k = \Theta(n_y), \\ \exp(-\Theta(L_k \log n_y)), & L_k \gg n_y, \end{cases}$$

and

$$\Pr(\mathbf{Z}^{L_k,1} \in \mathcal{E}_{L_k,3}) = 1 - n_y^{-\Omega(1)}.$$

F.2 Reducing the Wrong Attention

As discussed in Remark F.1, both $\epsilon_{\text{attn}}^{L_k}$ and the attention gap remain bounded by a small constant at the beginning of stage k . Hence, we can directly proceed to stage 1.2.2 of the convergence analysis as \mathcal{T}^2 . Moreover, by the symmetry between $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$, we may, without loss of generality, restrict attention to a particular $s \in \tau(\mathcal{X})$ and analyze the corresponding loss $\text{Loss}_{5,s}^{L_k,2}$ in what follows.

Induction F.1. Given $F^{(T_{k-1})}$ from the previous stage $\mathcal{T}^{L_{k-1}}$ with $L_{k-1} = 2^{k-1}$ for $k \geq 2$, let T_k denote the first time that $\text{Loss}_{5,s}^{L_k,2}$ decreases below $e^{(-\frac{1}{2}+3.01c_1) \cdot 2B}$. For all iterations $t \leq T_k$, we have the following holds

1. $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} + [\mathbf{Q}_{4,4}^{(t)}]_{s,s}$ monotonically increases;
2. for any sample $\mathbf{Z}^{2,1}$, we have $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \leq c_1$ for some sufficiently small constant $c_1 = \frac{1.005 \log d}{2B} > 0$;
3. for any $\mathbf{Z}^{2,1}$, we have $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \leq \min \left\{ \frac{L_k-1}{L_k} \epsilon_{\text{attn}}^{L_k,2} - c_2, 0 \right\}$, where $c_2 = \frac{\log d}{4B} > 0$ is some sufficiently small constant.
4. for $p \in \{3, 4\}$, for $s' \in \tau(\mathcal{X}) \neq s$, $|\mathbf{Q}_{4,p}^{(t)}]_{s,s'}| \leq O\left(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,s}}{d}\right)$; otherwise $[\mathbf{Q}_{4,p}^{(t)}]_{s,s'} = 0$.

F.2.1 Attention and Lambda Preliminaries

Lemma F.9. If Induction F.1 holds for all iterations $[T_{k-1}, t]$, then for any sample $\mathbf{Z}^{L_k,1}$, we have

1. $\epsilon_{\text{attn}}^{L_k} \in \left[\frac{4}{3}c_1, 2\epsilon_{\text{attn}}^{L_{k-1}} \right]$;
2. $\mathbf{Attn}_{\text{ans},1 \rightarrow \mathbf{k}}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \mathbf{k}'}^{(t)} \geq \Omega(1)$ for any $\mathbf{k}, \mathbf{k}' \in \mathcal{I}^{L_k,1} \setminus \{(\text{ans}, 1), (\text{pred}, 2)\}$;
3. $|\mathbf{Attn}_{\text{ans},1 \rightarrow \mathbf{k}}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \mathbf{k}'}^{(t)}| \leq \tilde{O}\left(\frac{1}{d}\right)$ for any $\mathbf{k}, \mathbf{k}' \in \mathcal{I}^{L_k,1} \setminus \{(\text{ans}, 1)\}$.

Proof. The highest loss occurs when the sample \mathbf{Z}^{L_k} makes the wrong answer highly confused with the correct answer. Thus, we consider the case that $g_\ell \cdot y_0 = g_2 \cdot y_0$ for all $\ell \in [L_k] \setminus \{2\}$ and $y_0 \neq y_1$.

$$\begin{aligned}
& \log p_F(\mathbf{Z}_{\text{ans},2,5} | \mathbf{Z}^{(L_k,1)}) \\
& \leq \Theta(1) \cdot \max \left\{ e^{\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \frac{L_k-1}{L_k} \epsilon_{\text{attn}}^{L_k,2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) 2B - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \frac{1}{L_k} \epsilon_{\text{attn}}^{L_k,2} \right) 2B}, \right. \\
& \quad \left. e^{\log d - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \frac{1}{L_k} \epsilon_{\text{attn}}^{L_k,2} \right) 2B} \right\} \\
& \leq \Theta \left(e^{\left(\frac{L_k-1}{L_k} \epsilon_{\text{attn}}^{L_k,2} + c_1 \right) 2B - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \frac{1}{L_k} \epsilon_{\text{attn}}^{L_k,2} \right) 2B} \right) \\
& \leq \Theta \left(e^{-\left(\frac{1}{2} - \frac{3}{2} \epsilon_{\text{attn}}^{L_k,2} - c_1 \right) 2B} \right).
\end{aligned}$$

Thus, if $\epsilon_{\text{attn}}^{L_k} \leq \frac{4}{3}c_1$, we must have

$$\text{Loss}_{5,s}^{2,2} \leq \Theta \left(e^{(-\frac{1}{2} + 3 \cdot \frac{2c_1}{3} + c_1) 2B} \right) = \Theta \left(e^{(-\frac{1}{2} + 3c_1) 2B} \right),$$

which contradicts the definition of $T_{1,2,2,s}$. Therefore, it must hold that $\epsilon_{\text{attn}}^{L_k} \geq \frac{4}{3}c_1$ for all $t \leq T_k$. \square

Lemma F.10. If Induction F.1 holds for all iterations $[T_{k-1}, t]$, then given $\mathbf{Z}^{L_k,1} \notin \mathcal{E}_{L_k,2}$, we have

1. for the prediction j_2 , we have

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = \left(\sum_{\ell \in [L_k]} \mathbb{1}_{g_\ell(y_1)=g_2(y_1)} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) \cdot 2B + O\left(\frac{B}{n_y}\right) + O(\delta).$$

2. for the prediction $j'_2 = \tau(g_2(y_0))$, we have

$$\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} = \left(\sum_{\ell \in [L_k]} \mathbb{1}_{g_\ell(y_0)=g_2(y_0)} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \cdot 2B + O\left(\frac{B}{n_y}\right) + O(\delta).$$

3. for other $j = g_2(y) \in \tau(\mathcal{Y})$, noticing that for this case $y \neq y_0, y_1$, then we have

$$\begin{aligned} \Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} &= \left(\sum_{\ell \in [L_k]} \mathbb{1}_{g_\ell(y)=g_2(y)} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) \cdot 2B \\ &\quad + O\left(\frac{B}{n_y}\right) + O(\delta). \end{aligned}$$

Lemma F.11. If Induction F.1 holds for all iterations $[T_{k-1}, t)$, then given $\mathbf{Z}^{L_k,1} \in \mathcal{E}_{L_k,2}$, we have

1. for the prediction j_2 , we have

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} = \sum_{\ell \in [L_k]} \mathbb{1}_{g_\ell(y_1)=g_2(y_1)} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \cdot 2B + O\left(\frac{B}{n_y}\right) + O(\delta).$$

2. for other $j = \tau(g_2(y)) \in \tau(\mathcal{Y})$ with $y \neq y_1$, then we have

$$\begin{aligned} \Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} &= \left(\sum_{\ell \in [L_k]} \mathbb{1}_{g_\ell(y)=g_2(y)} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \right) \cdot 2B \\ &\quad + O\left(\frac{B}{n_y}\right) + O(\delta). \end{aligned}$$

F.2.2 Gradient Lemma

Lemma F.12. If Induction F.1 holds for all iterations $t \in [T_{k-1}, T_k)$, given $s \in \tau(\mathcal{X})$, we have

$$\begin{aligned} &\left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} \\ &\geq \Omega(B/d) \mathbb{E} \left[\epsilon_{\text{attn}}^{L_k} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) | \tau(x_1) = s, \mathcal{E}_{L_k,1} \cap \mathcal{E}_{L_k,2}^c \right]. \end{aligned}$$

Proof. By Induction F.1, Lemma F.10, and Lemma F.11, we have

$$\begin{aligned} &\mathbb{E} \left[\mathcal{N}_{s,L_k,i}^{(t)} \right] \\ &= \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(- \sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(g_\ell) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(y_0) \right. \right. \\ &\quad \left. \left. + (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \pm \tilde{O}(\delta^q) \right] \\ &\quad \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \leq B}. \end{aligned} \tag{126}$$

Similarly,

$$\begin{aligned} &\mathbb{E} \left[\mathcal{N}_{s,L_k,ii}^{(t)} \right] \\ &= \mathbb{E} \left[\sum_{y \neq y_1} \mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) \right. \\ &\quad \left. \left(\sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \cdot V_{\tau(g_2(y)),r_{g_2 \cdot y}}(g_\ell) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{\tau(g_2(y)),r_{g_2 \cdot y}}(y_0) \right) \right] \end{aligned}$$

$$- \left(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}\right) \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \pm \tilde{O}(\delta^q) \Big] \mathbb{1}_{\tau(x_1)=s}.$$

Consider the event $\mathcal{E}_{L_k,1}$. For $\mathbf{Z}^{L_k,1} \in \mathcal{E}_{L_k,1} \cap \mathcal{E}_{L_k,2}^c$, we have $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \leq B$, since

$$\begin{aligned} \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} &= \sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(g_\ell) \mathbb{1}_{g_\ell(y_1)=j_2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(y_0) \\ &\leq 2 \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + (U_k - 1/L_k) \cdot \epsilon_{\text{attn}}^{L_k} \right) B \\ &\leq 2 \left(\frac{1 - \epsilon_{\text{attn}}^{L_k} + c_1}{2} + (U_k - 1/L_k) \cdot \epsilon_{\text{attn}}^{L_k} \right) B \\ &= 2 \left(\frac{1 + c_1}{2} - \left(\frac{1}{2} + 1/L_k - U_k \right) \cdot \epsilon_{\text{attn}}^{L_k} \right) B. \end{aligned}$$

Thus, provided $\epsilon_{\text{attn}}^{L_k} \geq 4c_1/3$ and choosing $U_k = \lfloor 1 + \frac{L_k}{8} \rfloor / L_k$, we indeed have $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \leq B$.

- For $j = \tau(g_2(y))$ in $\mathcal{N}_{s,L_k,ii}^{(t)}$:

$$- y = y_0 \neq y_1,$$

$$\begin{aligned} &\sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \cdot V_{\tau(g_2(y)),r_{g_2 \cdot y}}(g_\ell) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{\tau(g_2(y)),r_{g_2 \cdot y}}(y_0) \\ &\geq \sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \cdot V_{\tau(g_2(y)),r_{g_2 \cdot y}}(g_\ell) \mathbb{1}_{g_\ell(y) \neq g_2(y)} \\ &\geq -O\left(\frac{B}{n_y}\right) \cdot \left(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}\right). \end{aligned} \tag{127}$$

- $y \neq y_0, y_1$. If there exists at least one $\ell \neq 2$ such that $g_\ell(y) = g_2(y)$, then by cancellation,

$$\begin{aligned} &\sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \cdot V_{\tau(g_2(y)),r_{g_2 \cdot y}}(g_\ell) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{\tau(g_2(y)),r_{g_2 \cdot y}}(y_0) \\ &\geq -O\left(\frac{B}{n_y}\right) \cdot \left(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}\right). \end{aligned} \tag{128}$$

Else, pick $y' \neq y, y_1$ so that $\Lambda_{5,\tau(g_2(y')),r_{g_2 \cdot y'}}^{(t)} \geq \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}$. Then

$$\mathbf{logit}_{5,\tau(g_2(y))} \leq O\left(\frac{1}{n_y}\right) (1 - \mathbf{logit}_{5,j_2}),$$

which implies

$$\begin{aligned} &\mathbf{logit}_{5,j}^{(t)} \left(\sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \cdot V_{\tau(g_2(y)),r_{g_2 \cdot y}}(g_\ell) + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{\tau(g_2(y)),r_{g_2 \cdot y}}(y_0) \right) \\ &\geq -O\left(\frac{B}{L_k n_y}\right) \cdot \left(1 - \mathbf{logit}_{5,j_2}\right) \cdot \left(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}\right). \end{aligned} \tag{129}$$

- For $\mathcal{N}_{s,L_k,i}^{(t)}$:

- If $\mathbf{Z}^{L_k,1} \in \mathcal{E}_{L_k,1} \cap \mathcal{E}_{L_k,2}^c$, then

$$\begin{aligned} &-\sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(g_\ell) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(y_0) \\ &\quad + \left(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}\right) \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \\ &\geq -\sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \mathbb{1}_{g_\ell=j_2} \cdot V_{j_2,r_{g_2 \cdot y_1}}(g_\ell) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(y_0) \end{aligned}$$

$$\begin{aligned}
& + (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \\
& \geq \left(- (U_k - \frac{1}{L_k}) \cdot 2B + \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \cdot \epsilon_{\text{attn}}^{L_k} \geq \left(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} - \frac{B}{4} \right) \cdot \epsilon_{\text{attn}}^{L_k}.
\end{aligned} \tag{130}$$

– Else, for $\mathbf{Z}^{L_k,1} \in \mathcal{E}_{L_k,1}^c \cup \mathcal{E}_{L_k,2}$, note that having more identical predictions $g_\ell(y_1) = j_2$ or less distraction (i.e., $y_0 = y_1$) increases $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}$. Hence

$$\mathbb{E}[(1 - \mathbf{logit}_{5,j_2}) \mid \mathcal{E}_{L_k,1}^c \cup \mathcal{E}_{L_k,2}] \leq O(1) \cdot \mathbb{E}[(1 - \mathbf{logit}_{5,j_2}) \mid \mathcal{E}_{L_k,1} \cap \mathcal{E}_{L_k,2}^c],$$

and

$$\begin{aligned}
& \left| - \sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(g_\ell) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \cdot V_{j_2,r_{g_2 \cdot y_1}}(y_0) \right. \\
& \quad \left. + (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right| \leq O(1) \cdot \epsilon_{\text{attn}}^{L_k} \cdot B.
\end{aligned} \tag{131}$$

Putting the above together, we have

$$\begin{aligned}
& \left[- \nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[- \nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} \\
& = \mathbb{E} \left[\mathcal{N}_{s,L_k,1}^{(t)} \mathbb{1}_{\mathcal{E}_{L_k,1} \cap \mathcal{E}_{L_k,2}^c} \right] + \mathbb{E} \left[\mathcal{N}_{s,L_k,1}^{(t)} \mathbb{1}_{\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \leq B} \mathbb{1}_{\mathcal{E}_{L_k,1}^c \cup \mathcal{E}_{L_k,2}} \right] \\
& \quad + \mathbb{E} \left[\mathcal{N}_{s,L_k,2}^{(t)} \right] \pm \tilde{O}(\sigma_0^q) \\
& = \mathbb{E} \left[(\mathcal{N}_{s,L_k,1}^{(t)} + \mathcal{N}_{s,L_k,2}^{(t)}) \mathbb{1}_{\mathcal{E}_{L_k,1} \cap \mathcal{E}_{L_k,2}^c} \right] \\
& \quad + \mathbb{E} \left[\mathcal{N}_{s,L_k,1}^{(t)} \mathbb{1}_{\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \leq B} \mathbb{1}_{\mathcal{E}_{L_k,1}^c \cup \mathcal{E}_{L_k,2}} \right] + \mathbb{E} \left[\mathcal{N}_{s,L_k,2}^{(t)} \mathbb{1}_{\mathcal{E}_{L_k,1} \cup \mathcal{E}_{L_k,2}^c} \right] \pm \tilde{O}(\sigma_0^q).
\end{aligned}$$

First, by (127), (128), (130),

$$\begin{aligned}
& \mathbb{E} \left[(\mathcal{N}_{s,L_k,1}^{(t)} + \mathcal{N}_{s,L_k,2}^{(t)}) \mathbb{1}_{\mathcal{E}_{L_k,1} \cap \mathcal{E}_{L_k,2}^c} \right] \\
& \geq \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \left((\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} - \frac{B}{4}) \epsilon_{\text{attn}}^{L_k} \right) - (1 - \mathbf{logit}_{5,j_2}^{(t)}) \left(O(\frac{B}{n_y}) \epsilon_{\text{attn}}^{L_k} + \max_{y \neq y_1} \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \right) \right. \\
& \quad \left. \cdot \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_{L_k,1} \cap \mathcal{E}_{L_k,2}^c} \right] \\
& \geq \Omega(B/d) \cdot \mathbb{E} \left[\epsilon_{\text{attn}}^{L_k} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_{L_k,1} \cap \mathcal{E}_{L_k,2}^c \right].
\end{aligned}$$

Moreover, by (131),

$$\begin{aligned}
& \mathbb{E} \left[\mathcal{N}_{s,L_k,1}^{(t)} \mathbb{1}_{\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \leq B/2} \mathbb{1}_{\mathcal{E}_{L_k,1}^c \cup \mathcal{E}_{L_k,2}} \right] \\
& \geq -\Pr(\mathbf{Z}^{L_k,1} \notin \mathcal{E}_{L_k,1}^c) \cdot O(B/d) \cdot \mathbb{E} \left[\epsilon_{\text{attn}}^{L_k} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_{L_k,1}^c \cup \mathcal{E}_{L_k,2} \right] \\
& \geq - \left(O(1/n^{\Omega(1)}) + 1/n_y \right) \cdot O(B/d) \cdot \mathbb{E} \left[\epsilon_{\text{attn}}^{L_k} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_{L_k,1} \cap \mathcal{E}_{L_k,2}^c \right],
\end{aligned}$$

and, by (127), (128), (129),

$$\mathbb{E} \left[\mathcal{N}_{s,L_k,2}^{(t)} \mathbb{1}_{\mathcal{E}_{L_k,1}^c \cup \mathcal{E}_{L_k,2}} \right] \geq \frac{1}{n_y} \cdot O(B/d) \cdot \mathbb{E} \left[\epsilon_{\text{attn}}^{L_k} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_{L_k,1}^c \cup \mathcal{E}_{L_k,2} \right].$$

Therefore,

$$\left[- \nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[- \nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s}$$

$$\geq \Omega(B/d) \cdot \mathbb{E} \left[\epsilon_{\text{attn}}^{L_k} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \mid \tau(x_1) = s, \mathcal{E}_{L_k,1} \cap \mathcal{E}_{L_k,2}^c \right].$$

□

Lemma F.13. *If Induction F.1 holds for all iterations $t \in [T_{k-1}, T_k]$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,p}]_{s,s'}$, $p \in \{3, 4\}$, $s' \neq s \in \tau(\mathcal{X})$, we have*

$$\left| \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s'} \right| \leq O\left(\frac{1}{d}\right) \left| \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} \right|.$$

F.2.3 Attention gap is small

Lemma F.14. *If Induction F.1 holds for all iterations $t \in [T_{k-1}, T_k]$, then for any sample $\mathbf{Z}^{L_k,1}$, we have*

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \leq c_1,$$

where $c_1 > 0$ is a small constant.

Proof. The proof follows the same high-level idea as Lemma E.22. The key observation is that for each $\mathbf{Z}^{L_k,1}$, there always exists at least one $y \neq y_1$ such that, for some index j appearing in the gradient term $\mathcal{N}_{s,L_k,2}^{(t)}$, we have

$$\Lambda_{5,j,r_{g_2 \cdot y}}^{(t)} \geq 2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)})B.$$

Let $\hat{\mathcal{J}}$ denote the set of indices j achieving $\arg \max_{y \notin \{y_1\}} \Lambda_{5,j,r_{g_2 \cdot y}}^{(t)}$, and pick one such index $j' = g_2(y')$. Define \tilde{T} to be the first time when $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \geq c_1$. At this time, we have

$$\sum_{j \in \hat{\mathcal{J}}} \mathbf{logit}_{5,j}^{(\tilde{T})} = \left(1 - O(d^{-\Omega(1)})\right) (1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}).$$

Plugging this into the gradient expressions yields

$$\begin{aligned} & \mathcal{N}_{s,3,L_k,i}^{(\tilde{T})} + \mathcal{N}_{s,3,L_k,ii}^{(\tilde{T})} \\ &= \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \cdot \left(\left(1 - O(d^{-\Omega(1)})\right) (1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) \cdot \left(\Lambda_{5,j',r_{g_2 \cdot y'}}^{(\tilde{T})} - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})} \right) \right. \\ & \quad \left. + O(d^{-\Omega(1)}) (1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) \cdot \left(V_{5,j_2,r_{g_2 \cdot y_1},2}(g_2) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})} \right) \right) \mathbb{1}_{\tau(x_1)=s}, \end{aligned}$$

and similarly

$$\begin{aligned} & \mathcal{N}_{s,4,L_k,i}^{(\tilde{T})} + \mathcal{N}_{s,4,L_k,ii}^{(\tilde{T})} \\ &= \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \cdot \left(\left(1 - O(d^{-\Omega(1)})\right) (1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) \cdot \left(-V_{j_2',r_{g_2 \cdot y_0}}(y_1) + \Lambda_{5,j',r_{g_2 \cdot y'}}^{(\tilde{T})} - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})} \right) \right. \\ & \quad \left. + O(d^{-\Omega(1)}) (1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) \cdot \left(V_{5,j_2,r_{g_2 \cdot y_1},2}(y_1) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})} \right) \right) \mathbb{1}_{\tau(x_1)=s}. \end{aligned}$$

Since $\Lambda_{5,j',r_{g_2 \cdot y'}}^{(\tilde{T})} - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})} \leq 0$ and

$$-V_{j_2',r_{g_2 \cdot y_0}}(y_1) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})} + \Lambda_{5,j_2,r_{g_2 \cdot y'}}^{(\tilde{T})} \geq \Omega(B),$$

we obtain

$$\left[-\nabla_{\mathbf{Q}_{4,4}^{(\tilde{T})}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,2} \right]_{s,s} \geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E}[1 - \text{logit}_{5,j_2}^{(\tilde{T})} \mid \tau(x_1) = s].$$

Therefore, the quantity $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}$ cannot further increase, and we conclude that

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \leq c_1.$$

□

Lemma F.15. *If Induction F.1 holds for all iterations $t \in [T_{k-1}, T_k)$, then for any sample $\mathbf{Z}^{L_k,1}$, we have*

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \leq \frac{L_k - 1}{L_k} \epsilon_{\text{attn}}^{L_k} - c_2,$$

where $c_2 = \frac{\log d}{2B} > 0$ is a small constant.

Proof. The argument parallels Lemma E.34, but in the recursive setting we first note that for every $\mathbf{Z}^{L_k,1} \in \mathcal{E}_{L_k,2}^c$,

$$\Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)} \leq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} + \frac{L_k - 1}{L_k} \epsilon_{\text{attn}}^{L_k}.$$

Let \tilde{T} be the first time such that

$$\mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} + \frac{L_k - 1}{L_k} \epsilon_{\text{attn}}^{L_k} \mid \tau(x_1) = s \right] \leq \frac{\log d}{4.02 B}.$$

Then, for any y ,

$$\begin{aligned} & \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(\tilde{T})} \\ &= 2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} + \sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(\tilde{T})} \mathbf{1}_{g_\ell(y)=g_2(y)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(\tilde{T})}) B \leq \frac{\log d}{2.01}. \end{aligned}$$

Hence the corresponding wrong-class logit mass is small:

$$\text{logit}_{5,j_2}^{(1)} \leq O(d^{-1.01/2.01}) (1 - \text{logit}_{5,j_2}^{(t)}).$$

Plugging this into the gradient decomposition, we obtain

$$\begin{aligned} & \mathcal{N}_{s,3,L_k,i}^{(\tilde{T})} + \mathcal{N}_{s,3,L_k,ii}^{(\tilde{T})} \\ &= \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \left((1 - \text{logit}_{5,j_2}^{(\tilde{T})}) (V_{5,j_2,r_{g_2 \cdot y_1},2}(g_2) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})}) \right. \\ & \quad \left. - O(d^{-1.01/2.01}) (1 - \text{logit}_{5,j_2}^{(\tilde{T})}) \sum_{y \neq y_0} (V_{5,j_2,r_{g_2 \cdot y},2}(g_2) - \Lambda_{5,j_2,r_{g_2 \cdot y}}^{(\tilde{T})}) \right) \mathbf{1}_{\tau(x_1)=s} \\ &\geq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \cdot \Omega(1) \cdot (1 - \text{logit}_{5,j_2}^{(\tilde{T})}) \cdot (V_{5,j_2,r_{g_2 \cdot y_1},2}(g_2) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})}) \mathbf{1}_{\tau(x_1)=s}, \end{aligned}$$

and

$$\begin{aligned} & \mathcal{N}_{s,4,L_k,i}^{(\tilde{T})} + \mathcal{N}_{s,4,L_k,ii}^{(\tilde{T})} \\ &= \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \left((1 - \text{logit}_{5,j_2}^{(\tilde{T})}) (V_{5,j_2,r_{g_2 \cdot y_1},2}(y_1) - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(\tilde{T})}) \right. \\ & \quad \left. - O(d^{-1.01/2.01}) (1 - \text{logit}_{5,j_2}^{(\tilde{T})}) \sum_{y \neq y_0} (V_{5,j_2,r_{g_2 \cdot y},2}(y_1) - \Lambda_{5,j_2,r_{g_2 \cdot y}}^{(\tilde{T})}) \right) \mathbf{1}_{\tau(x_1)=s} \end{aligned}$$

$$\leq -\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} \cdot \Omega(1) \cdot (1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) \cdot \Lambda_{5,j_2,r_{g_2},y_1}^{(\tilde{T})} \mathbb{1}_{\tau(x_1)=s}.$$

Therefore,

$$[-\nabla_{\mathbf{Q}_{4,3}^{(\tilde{T})}} \text{Loss}_5^{2,2}]_{s,s} \geq \Omega\left(\frac{B}{d}\right) \cdot \mathbb{E}[1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \mid \tau(x_1) = s, \mathcal{E}_{L_k,2}^c], \quad (132)$$

$$[-\nabla_{\mathbf{Q}_{4,4}^{(\tilde{T})}} \text{Loss}_5^{2,2}]_{s,s} \leq -\Omega\left(\frac{B}{d}\right) \cdot \mathbb{E}[1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})} \mid \tau(x_1) = s, \mathcal{E}_{L_k,2}^c]. \quad (133)$$

Finally, recall that

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} = \frac{e^{[\mathbf{Q}_{4,3}^{(t)}]_{s,s}} - e^{[\mathbf{Q}_{4,4}^{(t)}]_{s,s}} - e^{\tilde{O}(1/d)}}{e^{[\mathbf{Q}_{4,3}^{(t)}]_{s,s}} + e^{[\mathbf{Q}_{4,4}^{(t)}]_{s,s}} + e^{\tilde{O}(1/d)}}.$$

Combining the two gradient inequalities shows that the quantity

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \frac{L_k - 1}{L_k} \epsilon_{\text{attn}}^{L_k} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}$$

must strictly decrease at time $\tilde{T} + 1$. \square

F.2.4 At the end of stage

Lemma F.16 (At the end of stage). *Induction F.1 holds for all iterations $T_{k-1} < t \leq T_k = O(\frac{\text{poly}(d)}{\eta B})$. At the end of stage k , we have*

- (a) *Attention concentration: $\epsilon_{\text{attn}}^{L_k} \leq 5.04c_1$ for some small constant $c_1 = \frac{1.005 \log d}{B} > 0$;*
- (b) *Loss convergence: $\text{Loss}_{F(T_{k-1}),5}^{L_k,2} \leq e^{(-\frac{1}{2} + 3.01c_1)2B} = \frac{1}{\text{poly}d}$.*

Proof. By Lemmas F.9 and F.12, the diagonal entries $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$ and $[\mathbf{Q}_{4,4}^{(t)}]_{s,s}$ keep increasing until the total wrong-attention mass satisfies $\epsilon_{\text{attn}} \geq \frac{4c_1}{3}$. Combining Lemmas F.12 to F.15, there exists a stopping time

$$T_k = O\left(\frac{\text{poly}(d)}{\eta B}\right),$$

such that Induction F.1 holds for all $t \in [T_1, T_k]$.

To obtain an upper bound on $\epsilon_{\text{attn}}^{L_k}$ at the end of stage k , it suffices to work on the main event $\mathcal{E}_{L_k,3}$.

Case 1: $L_k = O(1)$. For each $y \in \mathcal{Y}$, at most one index $\ell \neq 2$ can satisfy $g_\ell(y) = g_2(y)$, and thus the number of collisions per y is uniformly bounded by W_k . Hence,

$$\begin{aligned} & \log p_F(Z_{\text{ans},2,5} \mid Z^{(2,1)}) \\ & \geq \Theta(1) \cdot \exp\left(\log d - (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \mathbb{1}_{g_\ell(y_1)=g_2(y_1)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)}) 2B\right) \\ & \geq \Theta(1) \cdot \exp\left(\log d - (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + (1/L_k - 1/L_k) \epsilon_{\text{attn}}^{L_k}) 2B\right) \\ & \geq \Theta\left(\exp\left(-\left(\frac{1}{2} + \frac{c_1}{2} - \frac{\epsilon_{\text{attn}}^{L_k}}{2} - \frac{\log d}{B}\right) 2B\right)\right), \end{aligned}$$

where the last line uses $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \leq \frac{1}{2}(1 - \epsilon_{\text{attn}}^{L_k} + c_1) = \frac{1+c_1}{2} - \frac{\epsilon_{\text{attn}}^{L_k}}{2}$.

Case 2: $L_k = \omega(1)$. If $L_k \leq n_y \log n_y$ then $W_k/L_k = O(1/L_k)$; otherwise $W_k/L_k = O(1/n_y)$. In either regime $W_k/L_k = o(1)$, and we similarly obtain

$$\begin{aligned} \log p_F(Z_{\text{ans},2,5} \mid Z^{(2,1)}) &\geq \Theta(1) \cdot \exp\left(\log d - (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + (o(1) - 1/L_k) \epsilon_{\text{attn}}^{L_k}) 2B\right) \\ &\geq \Theta\left(\exp\left(-\left(\frac{1}{2} + \frac{c_1}{2} - \frac{\epsilon_{\text{attn}}^{L_k}}{2} - \frac{\log d}{B}\right) 2B\right)\right). \end{aligned}$$

At convergence we therefore have

$$\frac{1}{2} + \frac{c_1}{2} - \frac{\epsilon_{\text{attn}}^{L_k}}{2} - \frac{\log d}{B} \geq \frac{1}{2} - 3.01 c_1,$$

which rearranges to

$$\epsilon_{\text{attn}}^{L_k} \leq 2(3.02c_1 + 0.5c_1 - c_1) = 5.04c_1.$$

In particular, since $B/\log d$ can be taken as a sufficiently large constant, we obtain the clean stage-end bound

$$\epsilon_{\text{attn}}^{L_k} = O(c_1).$$

□

F.3 Proof of Main Theorem

Theorem F.1 (Recursive self-training (Restatement)). *Assume the distribution \mathcal{D}^L induced from **LEGO**($\mathcal{X}, \mathcal{G}, \mathcal{Y}$) satisfies Assumption 3.1, 3.2 and 5.1, and assume the transformer network satisfies Assumption 3.3, 3.4, 3.5, and $n_y < m \ll \log^2 d$. Then for any $1 \leq k < \log_2 |\mathcal{X}|$, the transformer $F^{(T_k)}$ trained via Algorithm 2 up to length $L_k = 2^k$ and $T_k = O(\frac{\text{poly}(d)}{\eta})$ satisfies: $F^{(T_k)}$ is able to solve $\mathcal{T}^{L_{k+1}}$ with $L_{k+1} = 2^{k+1}$, i.e.,*

$$\text{Acc}_{L_{k+1}}(F^{(T_k)}) = 1 - O(1/\text{poly}(d)).$$

Proof. By Lemma F.16, at the time T_k , we have $\epsilon_{\text{attn}}^{L_k} \leq 5.04c_1$, combining with Induction F.1, non-diagonal entry of $\mathbf{Q}_{p,q}$ remains close to 0, thus moving to the next stage, we have $\epsilon_{\text{attn}}^{L_{k+1},\ell} \leq 4\epsilon_{\text{attn}}^{L_k} \leq 20.16c_1$ (notice that $20.16c_1$ could be sufficiently small e.g., 0.01 since we can set a reasonably large B) for all $\ell < L_{k+1}$. Moreover, $\Delta^{L_{k+1},\ell} \leq \Delta^{L_k,1}$. Hence, we have

$$\begin{aligned} &\mathbb{E}_{Z^{L_{k+1}} \sim \mathcal{D}^{L_{k+1}}} \left[\mathbb{E}_{\widehat{Z}_{\text{ans},\ell+1} \sim p_F^{(T_k)}(\cdot \mid Z^{L_{k+1},\ell})} [\mathbb{1}_{\{\widehat{Z}_{\text{ans},\ell+1} \neq Z_{\text{ans},\ell+1}\}}] \right] \\ &\leq O(1) \cdot \mathbb{E}_{Z^{L_{k+1}} \sim \mathcal{D}^{L_{k+1}}} \left[1 - \text{logit}_{5,\tau}(\mathbf{z}_{\text{ans},\ell+1,5}) \right] \\ &\leq O(1) \cdot e^{\left(-\left(\frac{1}{2} - \epsilon_{\text{attn}}^{L_{k+1}} - (\epsilon_{\text{attn}}^{L_k} - c_2)/2 - \epsilon_{\text{attn}}^{L_{k+1}}\right) 2B\right)} \\ &\leq O(1) \cdot e^{\left(-\frac{1}{2} - \frac{\log d}{4B} + \frac{5}{2} \cdot 0.01\right) 2B} = O\left(\frac{1}{\text{poly}d}\right). \end{aligned}$$

Thus $\text{Acc}_{L_{k+1}}(F^{T_k}) \geq 1 - O\left(\frac{1}{\text{poly}d}\right)$.

□