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

Yu Huang* Upenn
Zixin Wen* CMU
Aarti Singh CMU
Yuejie Chi Yale
Yuxin Chen Upenn

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 that require longer chain-of-thoughts (CoT). In this work, we present a theoretical analysis of transformers trained via gradient descent on synthetic data for various state tracking tasks, revealing how length-generalizable reasoning can emerge. Specifically, we prove that: (i) for tasks with simple algebraic structure such as cyclic-group composition, transformers trained on short, constant-length chains learn a solution pattern that extrapolates to much longer chains; and (ii) for more complex tasks such as symmetric-group composition, a recursive self-training curriculum bootstraps longer reasoning and generalizes well beyond the training horizon, up to the natural limit of our setting. Our results demonstrate that transformers can learn sequential reasoning skills that scale with problem complexity. Moreover, we provide the first optimization-based guarantee demonstrating that constant-depth transformers can learn the state tracking problems in NC^1, which exceeds the prior barrier limited to TC^0, unless the famous conjecture $\mathsf{TC}^0 \neq \mathsf{NC}^1$ is false.

1 Introduction

Reasoning is central to artificial intelligence [1, 2, 3, 4, 5, 6]. Transformer-based [7] large language models (LLMs) achieve state-of-the-art results on complex reasoning tasks via chain-of-thought (CoT) reasoning [8, 9, 10, 11, 12, 13, 14, 15], where the model generates intermediate steps before producing a final answer. Recent frontier models such as OpenAI o1 [16] and DeepSeek R1 [17] typically produce longer CoT traces at inference time, often induced by reinforcement learning and by supervised fine-tuning (SFT) that distills from longer chains [18, 19]. These advances correlate with improved performance on harder problems [20, 21, 22, 23, 24], but the mechanisms and limits underlying these behaviors remain poorly understood, posing fundamental theoretical challenges.

Theoretical studies on transformers with CoT have advanced along two fronts: expressiveness [25, 26, 27, 28, 29] and statistical learnability [30, 31, 32, 33, 34, 35, 36]. Seminal works (e.g., [26, 27]) show that transformers with CoT can express far more powerful circuits: [26] proves that constant-depth transformers without CoT behave as highly parallel devices limited to AC^0, a proper subset of TC^0,where "parallel" means the computation proceeds in a fixed number of synchronized layers that aggregate many inputs per layer so no dependency chain can grow with input length; by contrast, with $\mathsf{poly}(L)$ CoT steps for input length L they can simulate polynomial-size circuits, i.e., problems whose solutions advance stage by stage so that the number of essential serial updates scales with L and cannot be compressed into $O(1)$ layers. These results establish that CoT allows transformers to transcend purely parallel computation and tackle problems requiring **inherently sequential** reasoning.

*The first two authors contributed equally.

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

In contrast, there is limited understanding of *how* transformers actually acquire such reasoning abilities via training. Prior optimization-based analyses [37, 38, 39, 40] have largely focused on simple tasks (e.g., parity) that are fully parallelizable and do not require sequential steps. However, inherently sequential problems beyond TC^0 remain largely unexplored, even though they include many common reasoning tasks in practice (e.g., playing chess). As implied by the expressive power argument, solving such problems with a constant-depth transformer requires CoT. This leaves a gap between what transformers can *express* in principle and what they can *learn* through training.

Another key question, prompted by the success of longer CoT traces, is whether language models can extrapolate their reasoning beyond the CoT lengths of the training data, known as **length generalization**. Longer CoT by itself is not a virtue; it helps on longer and harder instances only when the learned solution composes reliably rather than merely extending the chain [41]. Empirical evidence on the length generalization of transformers on reasoning tasks is mixed [42, 43, 44, 45, 46, 47, 48], with several studies reporting limited extrapolation despite strong in-distribution performance. Architectural choices, including positional encoding and training context size, can materially affect generalization [49, 50, 51, 52, 53, 54, 55]. On the theoretical front, prior work largely provides existence or statistical guarantees, showing that transformers can, in principle, represent length-generalizing algorithms or enjoy favorable sample complexity independent of the length of the CoT [56, 57, 58, 59, 60], but it remains unclear whether transformers can provably learn such reliable length generalization via optimization.

Given these gaps from the optimization perspective, we ask the following two questions:

Research Questions

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

To address these questions theoretically, we analyze a minimally viable transformer: a one-layer transformer block with softmax attention and a feed-forward network (FFN), trained end-to-end by GD under *no positional encoding* (NoPE). We study this model on synthetic *state-tracking* tasks, LEGO [42, 61], which distill core LLM skills such as entity tracking, game-state updates, and code evaluation [62]. This setup is tractable for analysis yet retains the mechanisms needed for step-by-step computation via CoT. We analyze the training dynamics with CoT on two LEGO task families: a simple class in TC^0 and a more challenging class in NC^1. By tracking attention patterns throughout training, we show how reasoning capabilities emerge and how length generalization is enabled by the structural properties of these tasks. Our main theoretical contributions are as follows:

1. **Strong length generalization on a simple LEGO task in TC^0.** We prove that for a class of simple LEGO state-tracking tasks in TC^0, a one-layer NoPE transformer with CoT, trained by GD on constant-length tasks, directly generalizes to tasks with significantly longer reasoning chains of length d^{c^*} (where d is the input dimension and constant $0 < c^* < 1$). By tracking attention patterns during training, we identify an **attention concentration** mechanism at convergence and show how it enables length generalization.
2. **Recursive self-training extends solvable length for NC^1 tasks.** For a class of challenging LEGO state-tracking tasks in NC^1, length generalization may saturate due to insufficient attention concentration compared with the TC^0 case. We introduce a *self-training* curriculum that recursively trains the model on its own CoT traces for slightly longer, constant-factor extensions, motivated by the empirical observations in [51]. We show that this scheme provably bootstraps the solvable length up to d over rounds of self-training, thereby providing a theoretical account of self-improvement.
3. **Constant-depth transformers learn inherently sequential problems beyond TC^0 via CoT.** Our self-training scheme further establishes that the model can learn to solve tasks in NC^1; in our case the task class is NC^1-complete, which lies outside TC^0 unless the widely held conjecture $\mathsf{TC}^0 \neq \mathsf{NC}^1$ fails. Therefore, we provide *the first optimization guarantee* that a one-layer transformer with CoT learns to solve reasoning tasks beyond TC^0, matching the expressivity result of [26] with a trained (rather than hand-crafted) model.

1.1 Overview of Main Results

We begin with an informal description of the reasoning task, commonly known as the *state-tracking problem*. Given a group $\mathcal{G}$ acting on a state space $\mathcal{Y}$, the goal is to find the final state obtained by applying a sequence of group actions to an initial state. For example, let $\mathcal{G}$ be the cyclic group C_n, acting on $\mathbb{Z}_n = \{0, 1, \ldots, n-1\}$ by rotations (i.e., addition modulo n). The problem is: given an initial state (e.g., 0) and a sequence of rotations (e.g., "rotate by 2", "rotate by 3"), predict the final state (e.g., $5 \bmod n$). This task naturally lends itself to CoT reasoning, in which intermediate steps explicitly track how the state evolves under successive actions. Our first main result shows that transformers can learn to solve such problems via CoT:

Theorem 1.1 (Learning CoT, informal)**.** *One-layer transformers trained via GD can provably learn to solve the state tracking problems involving cyclic and symmetric group actions using CoT reasoning.*

Simply transitive vs. symmetric. A simply transitive action is free and transitive: there is exactly one move sending any state to another one. The canonical example is the rotation action of C_n on $\mathbb{Z}_n$. In this regime, state tracking reduces to modular increments with a global commutative coordinate, enabling parallelprefix aggregation and placing the task within TC^0 for our instances. By contrast, in the symmetric case we consider the natural action of S_n on $\mathbb{Z}_n$ by permutations, where many group elements send a given i to j. Actions in S_n compose noncommutatively, depriving us of scanstyle shortcuts and making the problem inherently sequential; we situate this state tracking in NC^1 [61].

Moving on to length generalization guarantees, we show that transformers length-generalize to much longer CoT on simple tasks. When direct length generalization falls short, a recursive bootstrapping scheme enables further extension. We now present an informal version of the results.

Theorem 1.2 (Length generalization, informal)**.** *The algebraic structure dictates how far length generalization goes:*

- ***Simply transitive actions.*** *Training on constant length already yields generalization to sequences of length d^{c^*} (here d is the input dimension and $0 < c^* < 1$).*
- ***Symmetric actions.*** *Naive training only generalizes to a constant factor of the training length. However, solvable length up to d can be achieved via recursive self-training, where the model learns from its own CoT outputs.*

Optimization perspective on CoT. From an optimization view, recent theoretical progress shows that transformers can provably learn parity with CoT [37, 38, 40], but parity lies in TC^0. Moreover, [39] proves that a one-layer transformer learns to perform multi-step gradient descent for linear regression with CoT, which also remains in TC^0 because the objective size does not scale with the number of steps. Our result is significant from two aspects: (i) State tracking with symmetric actions lies in NC^1, so our optimization results provide the first provable guarantees for CoT beyond TC^0 and bridge expressiveness and optimization by matching the linear-length CoT expressiveness of [26] with end-to-end training guarantees. (ii) None of previous optimization results provide provable guarantees for length extrapolation even for simple tasks in TC^0, whereas we characterize when and how length generalization occurs.

Towards understanding length extrapolation of reasoning. Our work complements recent advances on length generalization in transformers. Prior studies [56, 57, 58, 59, 60] primarily establish statistical guarantees and *non–gradient-based* learnability. Notably, [59] identifies conditions under which specialized positional encodings and sparse contextual dependencies enable length extrapolation, provided the model fits the source length. [60] analyzes time-invariant autoregressive models (a fixed next-token generator) and shows that sample complexity can be independent of the CoT length: perfect one-step learning implies learnability for longer chains. We instantiate a concrete setting that mirrors these insights and, crucially, show that GD on a one-layer NoPE transformer *actually finds* such solutions. The learned model provably generalizes from fixed-length CoT to sequences that are significantly longer than the training horizon, thereby bridging statistical learnability with optimization dynamics. We further demonstrate that this length generalization can *bootstrap* longer CoT data for **self-training**, effectively extending the model's reasoning length beyond the annotated training corpus.

Significance of our techniques. Our proof techniques draw inspiration from recent advances in understanding the dynamics of feature learning in neural networks [63, 64, 65, 66, 67, 68, 69, 70],

which highlight how gradient-based training gives rise to useful internal patterns and representations. We build on these ideas to analyze how transformers gradually acquire length-generalizable reasoning ability through CoT training.

2 LEGO Language and State Tracking

We now introduce the background of the LEGO (*Learning Group and Equality Operations*) task [42], which was originally proposed as a synthetic task to study the reasoning behavior of transformers empirically. A typical LEGO instance in [42] has the form

$$\texttt{b = + a}, \quad \texttt{c = - b}, \quad \ldots, \quad \texttt{t = - s}, \quad \texttt{a = -1}, \quad \ldots$$

Here, `a, b, c, ...` are **variables**, each taking a **value** in $\{-1, +1\}$ in this example. Short expressions such as `b = + 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 `b = + a` and `a = -1`, it follows that `b = -1`. Formally, the LEGO language is defined as follows:

Definition 2.1 (LEGO language [42])**.** 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 $g : \mathcal{Y} \to \mathcal{Y}$. 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 $L' \leq L$:

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

which describes the chain of transitions:

$$\underbrace{x_0 \xrightarrow{g_1} x_1 \xrightarrow{g_2} x_2 \xrightarrow{g_3} \cdots \xrightarrow{g_{L'}} x_{L'}}_{\text{with answers } y_1, \ldots, y_{L'} \text{ up to } L'} \cdots \xrightarrow{g_L} x_L, \quad \text{starting with } x_0 = y_0.$$

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

Connection to the state tracking problem. LEGO instantiates a fundamental algorithmic reasoning task often termed *state tracking* [61, 71, 72]: given an initial state and a sequence of transformations, compute the resulting state. Equivalently, this can be viewed as evaluating a semiautomaton [43] or solving a word problem [61]. Beyond this formalization, state tracking is central to practical LLM abilities such as narrative entity tracking, chess move analysis, and code evaluation [62]. Consequently, it has become a standard synthetic testbed for probing the reasoning abilities of language models, both theoretically [42, 61, 71] and empirically [42, 72, 43]. In LEGO, predicate clauses encode transformations, while answer clauses encode observed states, which reduces state tracking to predicting the next answer consistent with the composed actions.

3 Theoretical Setup

As introduced in Section 2, we adapt the LEGO framework to probe the reasoning capabilities of transformer models. Define the **vocabulary** as $\mathcal{V} := \mathcal{X} \cup \mathcal{G} \cup \mathcal{Y} \cup \{\langle \mathsf{blank} \rangle\}$. The *blank* token $\langle \mathsf{blank} \rangle$ is a null symbol indicating the absence of other tokens. Let $d := |\mathcal{V}|$ denote the (finite) vocabulary size. For theoretical purposes, we analyze an asymptotic regime where $d \to \infty$; $|\mathcal{X}|$, $|\mathcal{G}|$, and $|\mathcal{Y}|$ may depend on d, and we will 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 2.1, we consider an asymptotic regime where both the vocabulary size d and the number of variables $|\mathcal{X}|$ tend to infinity. We assume $|\mathcal{G}| \leq \log^{C_0} d$ for some constant $C_0 \in [1, 100)$, ensuring that $\mathcal{Y}$ and $\mathcal{G}$ remain significantly smaller than $\mathcal{X}$.

3.1 LEGO Tokenization and Distribution

We begin by specifying how LEGO clauses are tokenized, and then define the LEGO distribution.

Definition 3.1 (LEGO encoding)**.** For any LEGO clause, we encode it as a fixed-length 5-token tuple $Z \in \mathcal{V}^5$. Specifically:

- For a predicate $x = g(x')$, set $Z_{\mathsf{pred}} \triangleq (x,\, g,\, x',\, \langle\mathsf{blank}\rangle,\, \langle\mathsf{blank}\rangle) \in \mathcal{V}^5$;
- For an answer $x = y$, set $Z_{\mathsf{ans}} \triangleq (\langle\mathsf{blank}\rangle,\, \langle\mathsf{blank}\rangle,\, \langle\mathsf{blank}\rangle,\, x,\, y) \in \mathcal{V}^5$.

By Definition 3.1, we can encode a LEGO sentence of the form (1) into a sequence $Z^{L,L'}$:

$$Z^{L,L'} = (Z_{\mathsf{pred},1}, \ldots, Z_{\mathsf{pred},L}, Z_{\mathsf{ans},0}, \ldots, Z_{\mathsf{ans},L'}), \tag{2}$$

where $Z_{\mathsf{pred},k}$ and $Z_{\mathsf{ans},k}$ are the corresponding predicate and answer clauses. We denote $\mathcal{I}^{L,L'} = \{(\mathsf{pred}, \ell)\}_{\ell\in[L]} \cup \{(\mathsf{ans}, \ell)\}_{\ell\in[L']}$ as the index set for the clauses in $Z^{L,L'}$. If $L = L'$, we simply write the sequence as Z^L and index set as $\mathcal{I}^L$.

To feed LEGO tokens into the 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 and fix notation used throughout the analysis.

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

With Definition 3.2, we can transform the LEGO sequence encoding into vector embeddings that can be used as inputs to any neural network models.

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

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

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

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

where each column $\mathbf{Z}_{\mathsf{pred},\ell}(resp.\mathbf{Z}_{\mathsf{ans},\ell}) \in \mathbb{R}^{d_c}$ is the embedding of clause $Z_{\mathsf{pred},\ell}(resp.Z_{\mathsf{ans},\ell})$.

With the token embedding defined, we now turn to the LEGO distribution.

Assumption 3.2 (LEGO distribution $\mathcal{D}^L, \mathcal{D}^{L,L'}$)**.** Given $\mathbf{LEGO}(\mathcal{X}, \mathcal{G}, \mathcal{Y})$ following Definition 2.1, letting L be the sequence length, we assume distribution $\mathcal{D}^L$ of length-L LEGO sentences satisfy the following properties.

1. All LEGO sentences $Z^L \sim \mathcal{D}^L$ are of the form (1) with $L' = L$ and are encoded by Definition 3.1 into the representation (2).
2. The variables $x_0, x_1, \ldots, x_L \in \mathcal{X}$ 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 $g_1, g_2, \ldots, g_L \in \mathcal{G}$ are sampled uniformly at random from $\mathcal{G}$ with replacement.
5. The intermediate values $y_1, y_2, \ldots, y_L$ 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 $Z^{L,L'}$ that contains all the predicates and the first $L'+1$ many answer clauses. $Z^{L,L'}$ is obtained by first sampling $Z^L \sim \mathcal{D}^L$ and then remove the answer clauses $Z_{\mathsf{ans},\ell}, \forall \ell > L'$.

One can easily see that the sequences sampled from $\mathcal{D}^L$ or $\mathcal{D}^{L,L'}$ represents valid LEGO sentences per Definition 2.1. 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 $Z^{L,L'}$ sampled from $\mathcal{D}^{L,L'}$.

3.2 Transformer Architecture

We first introduce a smoothed activation function that will be used in our network.

Definition 3.4 (Smooth ReLU)**.** We use a continuously differentiable variant of ReLU [73, 67]:

$$\mathbf{sReLU}(x) \;:=\; \begin{cases} \dfrac{\varrho}{q}, & x \le -\varrho, \\ \dfrac{x^q}{\varrho^{\,q-1} q}, & x \in (-\varrho, \varrho], \\ x - \varrho\Big(1 - \frac{1}{q}\Big), & x > \varrho, \end{cases}$$

where $q = O(1)$ is a large even integer and $\varrho = \Theta\big(1/\mathrm{polylog}(d)\big)$.

Transformer layers. We use an autoregressive transformer whose block [7] consists of a softmax attention layer followed by a position-wise feed-forward network (FFN). Given LEGO sentence embeddings $\mathbf{Z}^{L,L'}$ and indices $\boldsymbol{j}, \boldsymbol{k} \in \mathcal{I}^{L,L'}$, the attention from clause $\mathbf{Z}_{\boldsymbol{j}}$ to clause $\mathbf{Z}_{\boldsymbol{k}}$ is

$$\mathbf{Attn}_{\boldsymbol{j}\to\boldsymbol{k}}(\mathbf{Q}, \mathbf{Z}^{L,L'}) \;:=\; \frac{\exp(\mathbf{Z}_{\boldsymbol{j}}^\top \mathbf{Q}\, \mathbf{Z}_{\boldsymbol{k}})}{\sum_{\boldsymbol{r}\in\mathcal{I}^{L,L'}} \exp(\mathbf{Z}_{\boldsymbol{j}}^\top \mathbf{Q}\, \mathbf{Z}_{\boldsymbol{r}})}.$$

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

$$\mathrm{Attention}(\mathbf{Q}, \mathbf{Z}^{L,L'}) \;:=\; \sum\nolimits_{\boldsymbol{k}\in\mathcal{I}^{L,L'}} \mathbf{Attn}_{\mathsf{ans},L'\to\boldsymbol{k}}(\mathbf{Q}, \mathbf{Z}^{L,L'}) \cdot \mathbf{Z}_{\boldsymbol{k}}.$$

In the standard formulation, the score takes the form $\mathbf{Z}_{\boldsymbol{j}}^\top \mathbf{W}^{Q\top}\mathbf{W}^K \mathbf{Z}_{\boldsymbol{k}}$. Here, we fold $\mathbf{W}^{Q\top}\mathbf{W}^K$ into a single bilinear parameter $\mathbf{Q}$ so that the score is $\mathbf{Z}_{\boldsymbol{j}}^\top \mathbf{Q} \mathbf{Z}_{\boldsymbol{k}}$; this is an equivalent reparameterization that simplifies analysis without changing expressivity [69, 70, 74]. The FFN with parameter $\mathbf{W} \in \mathbb{R}^{5\times d\times m\times d_c}$ is defined by

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

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

Definition 3.5 (Transformer language model)**.** We assume that our learner neural network F is a one-layer decoder transformer block composed of an attention layer with NoPE (No Positional Encoding) and an FFN layer: $F = \mathrm{MLP} \circ \mathrm{Attention}$, with parameter $\mathbf{W} \in \mathbb{R}^{5\times d\times m\times d}$ and $\mathbf{Q} \in \mathbb{R}^{d_c\times d_c}$. Formally, for the i-th token position and the j-th vocabulary index,

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

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. Let $\mathcal{V}$ be the vocabulary with $|\mathcal{V}| = d$ and $\tau : \mathcal{V} \to [d]$ the index map. Given an encoded LEGO sequence $Z^\ell = (Z_1, \ldots, Z_\ell)$ with embedding $\mathbf{Z}^\ell$, the model's predictive distribution for the i-th token of the $(\ell+1)$-th clause is the following softmax:

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

Now we can sample the next clause $Z_{\ell+1}$ by sampling from the product distribution $Z_{\ell+1} = (Z_{\ell+1,1}, \ldots, Z_{\ell+1,5}) \sim \bigotimes_{i=1}^5 p_{F_i} := p_F$, in an autoregressive manner.

3.3 LEGO Task via CoT Reasoning and Training Objective

With the distribution and network in place, we now formalize the task within the LEGO framework via CoT reasoning. We view solving a length-L LEGO problem as generating a sequence of CoT steps that produce an intermediate answer for predicting $x_\ell = y_\ell$ at each step ℓ, before arriving at the final solution $x_L = y_L$. Specifically, we define the following reasoning task.

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` $\mathbf{W}$

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

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

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` $\mathbf{Q}$

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

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

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

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:

$$\mathrm{Acc}_L(F) = \frac{1}{L} \sum_{0\leq L'<L} \mathbb{E}_{Z^L\sim\mathcal{D}^L} \Big[\mathbb{E}_{\widehat{Z}_{\mathsf{ans},L'+1}\sim p_F(\cdot|Z^{L,L'})} \big[\mathbb{1}\{\widehat{Z}_{\mathsf{ans},L'+1} = Z_{\mathsf{ans},L'+1}\} \big] \Big], \tag{5}$$

where p_F is induced by the model F from (4). Clearly, $\mathrm{Acc}_L(F) \in [0, 1]$. We say $\mathcal{T}^L$ is solved if $\mathrm{Acc}_L(F) \approx 1$. 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 $Z_{\mathsf{ans},L'+1}$. To enforce the model to generate the CoT trace step by step, 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

$$\mathsf{Loss}^L(F) \triangleq \sum_{1\leq L'\leq L} \mathsf{Loss}^{L,L'}(F), \tag{6a}$$

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

We also define the per token loss $\mathsf{Loss}_i^{L,L'}(F) \triangleq \mathbb{E}_{Z^{L,L'}\sim\mathcal{D}^{L,L'}}[-\log p_{F_i}(Z_{\mathsf{ans},L',i} \mid Z^{L,L'-1})]$.

This objective is a teacher forcing style CoT training: at each step, the model is given the ground truth answers so far and is guided to match the next answer [75, 37]. We further adopt the following initialization for training.

Assumption 3.3 (Initialization)**.** Let F be the transformer network in Definition 3.5 with parameters $\mathbf{W}, \mathbf{Q}$. The attention parameter is zero-initialized: $\mathbf{Q}^{(0)} = \mathbf{0}_{d_c\times d_c}$. The FFN weights are initialized 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 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.

4 Main Results

In this section, we present our main results on the length-generalizable learning process of CoT reasoning on different LEGO problems.

4.1 Learning CoT on Simply Transitive Actions

We start with the simple 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 $\mathbb{Z}_n$. This simple state tracking problem is in TC^0.

Assumption 4.1 (Simply transitive group action)**.** We assume that $\mathcal{G}$ acts simply transitively on $\mathcal{Y}$, i.e., 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 assume $\mathcal{Y} = \{0, 1, \ldots, n_y - 1\}$ where $n_y \in \big[\Omega(\log\log d), \log d\big]$.

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, A.2, A.1, and 4.1, for some constant $0 < c^* < 1$, the transformer model $F^{(T_1+T_2)}$ obtained by Algorithm 1 with learning rate $\eta = \frac{1}{\mathsf{poly}(d)}$, and stage 1 and 2 iteration $T_1 = \widetilde{O}\Big(\frac{1}{\eta(\sigma_0)^{q-2}}\Big)$, $T_2 = \widetilde{O}\Big(\frac{\mathsf{poly}(d)}{\eta\sigma_0}\Big)$ satisfies*

1. **Direct short-to-long length generalization:**

$$\mathrm{Acc}_L\big(F^{(T_1+T_2)}\big) \;\geq\; 1 - \frac{1}{\mathsf{poly}(d)}, \text{ for every } L \leq O(d^{c^*}), \tag{7}$$

   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:** *given $Z^{2,\ell}$ with $\ell \in \{0, 1\}$, we have*

$$\mathbf{Attn}^{(T_1+T_2)}_{\mathsf{ans},\ell\to\mathsf{pred},\ell+1} + \mathbf{Attn}^{(T_1+T_2)}_{\mathsf{ans},\ell\to\mathsf{ans},\ell} \geq 1 - O\Big(\frac{1}{d^{c^*}}\Big). \tag{8}$$

Mechanism for solving LEGO tasks. Given $Z^{L,\ell}$, predicting the next answer $y_{\ell+1} = g_{\ell+1}(y_\ell)$ requires two steps: (i) **retrieve** the correct group element $g_{\ell+1}$ from the context clause $Z_{\mathsf{pred},\ell+1}$ and the current value y_ℓ from the answer clause $Z_{\mathsf{ans},\ell}$; and (ii) **apply the group operation**. It is well established that attention can implement content-based retrieval [76], and that FFN can represent the group operation [77]. Building on these insights, Algorithm 1 decouples learning in the attention (retrieval) and FFN (operation) components, thereby simplifying the analysis while preserving essential behavior. For task $\mathcal{T}^1$, the transcript $Z^{1,0}$ contains only the two relevant clauses, $Z_{\mathsf{pred},1}$ and $Z_{\mathsf{ans},0}$, without useless contents. Fixed uniform attention ($\mathbf{Q}$ initlized 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 FFN has indeed *learned* to apply the operation correctly. For task $\mathcal{T}^2$, with the FFN already trained, the attention layer only need 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.

How does attention concentration induce strong length generalization? As we increase the chain length in $\mathcal{T}^\ell$ 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 $Z_{\mathsf{pred},\ell+1}$ and the fourth token of the answer clause $Z_{\mathsf{ans},\ell}$. This co-occurrence furnishes a strong, consistent signal that enables robust retrieval across longer chains.

Choice of NoPE. Empirically, standard positional embeddings often hinder length extrapolation, while NoPE has been favored for its stronger length-generalization performance [49, 78]. Prior works [37, 38, 40] 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 since relevant information is retrieved from long, unordered contexts by *content* (variables) instead of position. This provides concrete architectural guidance for practice and identifies positional embeddings as a plausible cause of observed failures to generalize to unseen lengths.

4.2 Learning CoT on Symmetry Groups

We now turn to the case where the action group $\mathcal{G}$ is isomorphic to the symmetric group, under Assumption 4.2. In this case, the problem is NC^1-complete for $n_y \geq 5$.

Assumption 4.2 (Symmetry group actions)**.** We assume $\mathcal{Y} = \{0, 1, \ldots, n_y - 1\}$ and $\mathcal{G} = \mathbf{Sym}(\mathcal{Y})$, i.e. the symmetry group of order $|\mathcal{G}| = n_y!$. We assume $n_y = \Theta(\frac{\log\log d}{\log\log\log d})$ and $|\mathcal{G}| = n_y! = \mathrm{polylog} d$.

Hardness of the symmetry task. When predicting the next answer $y_{\ell+1}$, ambiguity arises when clauses other than $Z_{\mathsf{pred},\ell+1}$ and $Z_{\mathsf{ans},\ell}$ provide an element g' and an input y' such that $g'(y') = y_{\ell+1}$;

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 $\mathsf{E}_1 > 0$ (constant) ; $\tau_1, \tau_2 = \widetilde{O}(\frac{\mathsf{poly}d}{\eta})$; Total Stage K.

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

for $t = 1$ **to** τ_1 **do** `// Update the FFN parameter` $\mathbf{W}$

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

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

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` $\mathbf{Q}$

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

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

$T_1 \leftarrow t$;

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

for $k = 2$ **to** K **do** `// Stage` k `to solve` $\mathcal{T}^{2^k}$

$L \leftarrow 2^k$, $\widetilde{F}^{(k)} \leftarrow F^{(T_{k-1})}$;

while $\mathsf{Loss}^{L,2}_{\widetilde{F}^{(k)},5}\big(F^{(t-1)}\big) > \frac{1}{d^{\mathsf{E}_1}}$ **do** `// Update the attention parameter` $\mathbf{Q}$

$t \leftarrow t + 1$;

$\mathbf{Q}^{(t)} \leftarrow \mathbf{Q}^{(t-1)} - \eta \nabla_{\mathbf{Q}} \mathsf{Loss}^{L,2}_{\widetilde{F}^{(k)},5}\big(F^{(t-1)}\big)$;

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

$T_k \leftarrow t$;

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

we call such clauses *distractors*. For example, there may be other predicate clauses whose group elements also send y_ℓ to $y_{\ell+1}$. In the symmetric 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 helps extend reasoning length. Recent empirical studies [79, 80, 51] show that *length generalization* can be bootstrapped via model **self-improvement**: e.g., transformers trained on n-digit arithmetic often handle $n+1$ digits and improve further by training on their own predictions. This 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 $\widehat{p}_F$ induced by the network F is defined by

$$\widehat{p}_F(Z_{\mathsf{ans},L'+1}|Z^{L,L'}) = \begin{cases} 1, & \text{if } Z_{\mathsf{ans},L'+1} = \mathrm{argmax}_Z\, p_F(Z|Z^{L,L'}), \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

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

Definition 4.1 (Bootstrapped LEGO distribution)**.** We define $\mathcal{D}_F^{L,L'}$ as the LEGO distribution in Assumption 3.2 except that the answers $Z_{\mathsf{ans},\ell}, 1 \le \ell \le L'$ is given recursively by sampling the prediction $Z_{\mathsf{ans},\ell} \sim \widehat{p}_F(\cdot|Z^{L,\ell-1}), 1 \le \ell \le L'$ from the greedy language model $\widehat{p}_F$.

Definition 4.2 (Self-training loss)**.** Given a (fixed) model $\widetilde{F}$ and length L, The self-training next-clause-prediction loss is defined by replacing $\mathcal{D}^{L,L'}$ with $\mathcal{D}_F^{L,L'}$ (Definition 4.1) in (6):

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

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

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 4.2. *Assume the distribution $\mathcal{D}^L$ induced from **LEGO**$(\mathcal{X}, \mathcal{G}, \mathcal{Y})$ satisfies Assumption 3.1, 3.2 and 4.2, and assume the transformer network satisfies Assumption 3.3, A.1 and A.2. Then for any $1 \le k < \log_2 |\mathcal{X}|$, the transformer $F^{(T_k)}$ trained via Algorithm 2 up to length $L_k = 2^k$ and $T_k = O(\frac{\mathsf{poly}(d)}{\eta})$ satisfies:*

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

$$\mathrm{Acc}_{L_{k+1}}\big(F^{(T_k)}\big) = 1 - \frac{1}{\mathsf{poly}(d)}. \tag{11}$$

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

$$\mathbf{Attn}^{(T_k)}_{\mathsf{ans},\ell\to\mathsf{pred},\ell+1} + \mathbf{Attn}^{(T_k)}_{\mathsf{ans},\ell\to\mathsf{ans},\ell} \ge 1 - \widetilde{c}, \tag{12}$$

where $\widetilde{c}$ 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 $1/d^{\mathsf{E}_1}$ in Algorithm 2), (12) confirms that attention concentration is still insufficient. Nevertheless, Theorem 4.2 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 4.1 (Self-improvement for $|\mathcal{X}|$-length reasoning)**.** *Under the same assumptions as Theorem 4.2, letting $K = \Theta(\log d)$, for any length $L \le |\mathcal{X}|$, the model $F^{(T_K)}$ trained via Algorithm 2 achieves*

$$\mathrm{Acc}_L\big(F^{(T_K)}\big) \ge 1 - \frac{1}{\mathsf{poly}(d)}.$$

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: $L+1 \le |\mathcal{X}| = \Theta(d)$. Thus our guarantee attains the best possible length in this setting. Corollary 4.1 also demonstrates that the transformer can be trained to solve a task beyond TC^0 with linear-step CoT, matching the expressivity result of [26].[2] While prior empirical work reports self-improvement in practice, theoretical guarantees, especially for transformers and length generalization, have been scarce [81, 82, 83]. Theorem 4.2 provides, to our knowledge, the first rigorous evidence that transformers can *bootstrap* their reasoning via self-training without additional supervision.

Discussion on context rot. Context rot, the drop in accuracy and reliability as the input context grows even when the task is unchanged, has become a widely noted practical issue for LLM [84]. A prevailing empirical view is that longer contexts introduce many *distractors* and other irrelevant tokens, forcing the relevant signal to compete with them and thereby weakening retrieval. Our attention concentration perspective captures this at a high level: as irrelevant and distractor clauses accumulate, attention mass is diluted away from the correct clause, reducing performance at extended reasoning lengths. This offers a simple theoretical lens on context rot. We expect our analysis to extend to richer tasks and to inform practical mitigation strategies, e.g., *context engineering* [85].

5 Conclusions

In this paper, we theoretically analyzed how reasoning ability emerges during gradient-descent training of transformers on synthetic CoT tasks. For tasks in TC^0, we proved transformers can directly generalize from short constant-length chains to substantially longer tasks. For inherently sequential tasks in NC^1, we established novel convergence guarantees showing transformers can bootstrap their CoT length through recursive self-training. Our results bridge the gap between transformers' known expressive power beyond TC^0 and existing optimization theory of CoT , while also formally validating the effectiveness of self-improvement training observed empirically.

[2] 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.

Acknowledgments

The work of Z. Wen is supported in part by NSF DMS-2134080. 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.

References

[1] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

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

[3] Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

[4] Aitor Lewkowycz, Anders Andreassen, David Dohan, Ethan Dyer, Henryk Michalewski, Vinay Ramasesh, Ambrose Slone, Cem Anil, Imanol Schlag, Theo Gutman-Solo, et al. Solving quantitative reasoning problems with language models. *Advances in Neural Information Processing Systems*, 35:3843–3857, 2022.

[5] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, et al. Textbooks are all you need. *arXiv preprint arXiv:2306.11644*, 2023.

[6] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[7] Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

[8] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[9] Lucie Charlotte Magister, Jonathan Mallinson, Jakub Adamek, Eric Malmi, and Aliaksei Severyn. Teaching small language models to reason. *arXiv preprint arXiv:2212.08410*, 2022.

[10] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

[11] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[12] Laria Reynolds and Kyle 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.

[13] Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213, 2022.

[14] Aman Madaan and Amir Yazdanbakhsh. Text and patterns: For effective chain of thought, it takes two to tango. *arXiv preprint arXiv:2209.07686*, 2022.

[15] Denny Zhou, Nathanael Schärli, Le Hou, Jason Wei, Nathan Scales, Xuezhi Wang, Dale Schuurmans, Claire Cui, Olivier Bousquet, Quoc Le, et al. Least-to-most prompting enables complex reasoning in large language models. *arXiv preprint arXiv:2205.10625*, 2022.

[16] OpenAI. Openai o1 system card. *ArXiv*, abs/2412.16720, 2024.

[17] DeepSeek-AI. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *ArXiv*, abs/2501.12948, 2025.

[18] Yiwei Qin, Xuefeng Li, Haoyang Zou, Yixiu Liu, Shijie Xia, Zhen Huang, Yixin Ye, Weizhe Yuan, Hector Liu, Yuanzhi Li, et al. O1 replication journey: A strategic progress report–part 1. *arXiv preprint arXiv:2410.18982*, 2024.

[19] Yingqian Min, Zhipeng Chen, Jinhao Jiang, Jie Chen, Jia Deng, Yiwen Hu, Yiru Tang, Jiapeng Wang, Xiaoxue Cheng, Huatong Song, et al. Imitate, explore, and self-improve: A reproduction report on slow-thinking reasoning systems. *arXiv preprint arXiv:2412.09413*, 2024.

[20] Aviral Kumar, Vincent Zhuang, Rishabh Agarwal, Yi Su, John D Co-Reyes, Avi Singh, Kate Baumli, Shariq Iqbal, Colton Bishop, Rebecca Roelofs, et al. Training language models to self-correct via reinforcement learning. *arXiv preprint arXiv:2409.12917*, 2024.

[21] Yuxi Xie, Anirudh Goyal, Wenyue Zheng, Min-Yen Kan, Timothy P Lillicrap, Kenji Kawaguchi, and Michael Shieh. Monte carlo tree search boosts reasoning via iterative preference learning. *arXiv preprint arXiv:2405.00451*, 2024.

[22] Tianyang Zhong, Zhengliang Liu, Yi Pan, Yutong Zhang, Yifan Zhou, Shizhe Liang, Zihao Wu, Yanjun Lyu, Peng Shu, Xiaowei Yu, et al. Evaluation of openai o1: Opportunities and challenges of agi. *arXiv preprint arXiv:2409.18486*, 2024.

[23] Tianchen Gao, Jiashun Jin, Zheng Tracy Ke, and Gabriel Moryoussef. A comparison of deepseek and other llms. *arXiv preprint arXiv:2502.03688*, 2025.

[24] Cameron R Jones and Benjamin K Bergen. Large language models pass the turing test. *arXiv preprint arXiv:2503.23674*, 2025.

[25] Guhao Feng, Bohang Zhang, Yuntian Gu, Haotian Ye, Di He, and Liwei Wang. Towards revealing the mystery behind chain of thought: a theoretical perspective. *Advances in Neural Information Processing Systems*, 36:70757–70798, 2023.

[26] Zhiyuan Li, Hong Liu, Denny Zhou, and Tengyu Ma. Chain of thought empowers transformers to solve inherently serial problems. *arXiv preprint arXiv:2402.12875*, 1, 2024.

[27] William Merrill and Ashish Sabharwal. The expressive power of transformers with chain of thought. *arXiv preprint arXiv:2310.07923*, 2023.

[28] Lijie Chen, Binghui Peng, and Hongxun Wu. Theoretical limitations of multi-layer transformer. *arXiv preprint arXiv:2412.02975*, 2024.

[29] William Merrill and Ashish Sabharwal. The parallelism tradeoff: Limitations of log-precision transformers. *Transactions of the Association for Computational Linguistics*, 11:531–545, 2023.

[30] Emmanuel Abbe, Samy Bengio, Aryo Lotfi, Colin Sandon, and Omid Saremi. How far can transformers reason? the globality barrier and inductive scratchpad. *Advances in Neural Information Processing Systems*, 37:27850–27895, 2024.

[31] Noam Wies, Yoav Levine, and Amnon Shashua. Sub-task decomposition enables learning in sequence to sequence tasks. *arXiv preprint arXiv:2204.02892*, 2022.

[32] Clayton Sanford, Bahare Fatemi, Ethan Hall, Anton Tsitsulin, Mehran Kazemi, Jonathan Halcrow, Bryan Perozzi, and Vahab Mirrokni. Understanding transformer reasoning capabilities via graph algorithms. *Advances in Neural Information Processing Systems*, 37:78320–78370, 2024.

[33] Juno Kim, Denny Wu, Jason Lee, and Taiji Suzuki. Metastable dynamics of chain-of-thought reasoning: Provable benefits of search, rl and distillation. *arXiv preprint arXiv:2502.01694*, 2025.

[34] Xinyang Hu, Fengzhuo Zhang, Siyu Chen, and Zhuoran Yang. Unveiling the statistical foundations of chain-of-thought prompting methods. *arXiv preprint arXiv:2408.14511*, 2024.

[35] Ben Prystawski, Michael Li, and Noah Goodman. Why think step by step? reasoning emerges from the locality of experience. *Advances in Neural Information Processing Systems*, 36:70926–70947, 2023.

[36] Yingcong Li, Kartik Sreenivasan, Angeliki Giannou, Dimitris Papailiopoulos, and Samet Oymak. Dissecting chain-of-thought: Compositionality through in-context filtering and learning. *Advances in Neural Information Processing Systems*, 36:22021–22046, 2023.

[37] Juno Kim and Taiji Suzuki. Transformers provably solve parity efficiently with chain of thought. *arXiv preprint arXiv:2410.08633*, 2024.

[38] Kaiyue Wen, Huaqing Zhang, Hongzhou Lin, and Jingzhao Zhang. From sparse dependence to sparse attention: unveiling how chain-of-thought enhances transformer sample efficiency. *arXiv preprint arXiv:2410.05459*, 2024.

[39] Jianhao Huang, Zixuan Wang, and Jason D Lee. Transformers learn to implement multi-step gradient descent with chain of thought. *arXiv preprint arXiv:2502.21212*, 2025.

[40] Ruiquan Huang, Yingbin Liang, and Jing Yang. How transformers learn regular language recognition: A theoretical study on training dynamics and implicit bias. *arXiv preprint arXiv:2505.00926*, 2025.

[41] Tian Xie, Zitian Gao, Qingnan Ren, Haoming Luo, Yuqian Hong, Bryan Dai, Joey Zhou, Kai Qiu, Zhirong Wu, and Chong Luo. Logic-rl: Unleashing llm reasoning with rule-based reinforcement learning. *arXiv preprint arXiv:2502.14768*, 2025.

[42] Yi Zhang, Arturs Backurs, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, and Tal Wagner. Unveiling transformers with lego: a synthetic reasoning task. *arXiv preprint arXiv:2206.04301*, 2022.

[43] Bingbin Liu, Jordan T Ash, Surbhi Goel, Akshay Krishnamurthy, and Cyril Zhang. Transformers learn shortcuts to automata. *arXiv preprint arXiv:2210.10749*, 2022.

[44] Samy Jelassi, Stéphane d'Ascoli, Carles Domingo-Enrich, Yuhuai Wu, Yuanzhi Li, and François Charton. Length generalization in arithmetic transformers. *arXiv preprint arXiv:2306.15400*, 2023.

[45] Hattie Zhou, Arwen Bradley, Etai Littwin, Noam Razin, Omid Saremi, Josh Susskind, Samy Bengio, and Preetum Nakkiran. What algorithms can transformers learn? a study in length generalization. *arXiv preprint arXiv:2310.16028*, 2023.

[46] Yongchao Zhou, Uri Alon, Xinyun Chen, Xuezhi Wang, Rishabh Agarwal, and Denny Zhou. Transformers can achieve length generalization but not robustly. *arXiv preprint arXiv:2402.09371*, 2024.

[47] Kaiying Hou, David Brandfonbrener, Sham Kakade, Samy Jelassi, and Eran Malach. Universal length generalization with turing programs. *arXiv preprint arXiv:2407.03310*, 2024.

[48] Changnan Xiao and Bing Liu. Generalizing reasoning problems to longer lengths. In *The Thirteenth International Conference on Learning Representations*, 2025.

[49] Amirhossein Kazemnejad, Inkit Padhi, Karthikeyan Natesan Ramamurthy, Payel Das, and Siva Reddy. The impact of positional encoding on length generalization in transformers. *Advances in Neural Information Processing Systems*, 36:24892–24928, 2023.

[50] Mahdi Sabbaghi, George Pappas, Hamed Hassani, and Surbhi Goel. Explicitly encoding structural symmetry is key to length generalization in arithmetic tasks. *arXiv preprint arXiv:2406.01895*, 2024.

[51] Nayoung Lee, Ziyang Cai, Avi Schwarzschild, Kangwook Lee, and Dimitris Papailiopoulos. Self-improving transformers overcome easy-to-hard and length generalization challenges. *arXiv preprint arXiv:2502.01612*, 2025.

[52] Cem Anil, Yuhuai Wu, Anders Andreassen, Aitor Lewkowycz, Vedant Misra, Vinay Ramasesh, Ambrose Slone, Guy Gur-Ari, Ethan Dyer, and Behnam Neyshabur. Exploring length generalization in large language models. *Advances in Neural Information Processing Systems*, 35:38546–38556, 2022.

[53] Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. Self-attention with relative position representations. *arXiv preprint arXiv:1803.02155*, 2018.

[54] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of machine learning research*, 21(140):1–67, 2020.

[55] Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36:70293–70332, 2023.

[56] Annie Marsden, Evan Dogariu, Naman Agarwal, Xinyi Chen, Daniel Suo, and Elad Hazan. Provable length generalization in sequence prediction via spectral filtering. *arXiv preprint arXiv:2411.01035*, 2024.

[57] Kartik Ahuja and Amin Mansouri. On provable length and compositional generalization. *arXiv preprint arXiv:2402.04875*, 2024.

[58] Xinting Huang, Andy Yang, Satwik Bhattamishra, Yash Sarrof, Andreas Krebs, Hattie Zhou, Preetum Nakkiran, and Michael Hahn. A formal framework for understanding length generalization in transformers. *arXiv preprint arXiv:2410.02140*, 2024.

[59] Noah Golowich, Samy Jelassi, David Brandfonbrener, Sham M Kakade, and Eran Malach. The role of sparsity for length generalization in transformers. *arXiv preprint arXiv:2502.16792*, 2025.

[60] Nirmit Joshi, Gal Vardi, Adam Block, Surbhi Goel, Zhiyuan Li, Theodor Misiakiewicz, and Nathan Srebro. A theory of learning with autoregressive chain of thought. *arXiv preprint arXiv:2503.07932*, 2025.

[61] William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. *arXiv preprint arXiv:2404.08819*, 2024.

[62] Najoung Kim and Sebastian Schuster. Entity tracking in language models. *arXiv preprint arXiv:2305.02363*, 2023.

[63] Zeyuan Allen-Zhu and Yuanzhi Li. Forward super-resolution: How can gans learn hierarchical generative models for real-world distributions. *arXiv preprint arXiv:2106.02619*, 2021.

[64] Zeyuan Allen-Zhu and Yuanzhi 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.

[65] Zixin Wen and Yuanzhi Li. Toward understanding the feature learning process of self-supervised contrastive learning. In *International Conference on Machine Learning*, pages 11112–11122. PMLR, 2021.

[66] Zixin Wen and Yuanzhi Li. The mechanism of prediction head in non-contrastive self-supervised learning. *Advances in Neural Information Processing Systems*, 35:24794–24809, 2022.

[67] Yu Huang, Junyang Lin, Chang Zhou, Hongxia Yang, and Longbo 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.

[68] Samy Jelassi, Michael Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. *Advances in Neural Information Processing Systems*, 35:37822–37836, 2022.

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

[70] Yu Huang, Zixin Wen, Yuejie Chi, and Yingbin Liang. A theoretical analysis of self-supervised learning for vision transformers. In *The Thirteenth International Conference on Learning Representations*, 2025.

[71] William Merrill and Ashish Sabharwal. A little depth goes a long way: The expressive power of log-depth transformers. *arXiv preprint arXiv:2503.03961*, 2025.

[72] Belinda Z. Li, Zifan Carl Guo, and Jacob Andreas. (how) do language models track state? *ArXiv*, abs/2503.02854, 2025.

[73] Zeyuan Allen-Zhu and Yuanzhi Li. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

[74] Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. Multi-head transformers provably learn symbolic multi-step reasoning via gradient descent. *arXiv preprint arXiv:2508.08222*, 2025.

[75] Namgyu Ho, Laura Schmid, and Se-Young Yun. Large language models are reasoning teachers. *arXiv preprint arXiv:2212.10071*, 2022.

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

[77] Zhou Lu, Hongming Pu, Feicheng Wang, Zhiqiang Hu, and Liwei Wang. The expressive power of neural networks: A view from the width. *Advances in neural information processing systems*, 30, 2017.

[78] Zeyuan 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`.

[79] Avi Singh, John D Co-Reyes, Rishabh Agarwal, Ankesh Anand, Piyush Patil, Xavier Garcia, Peter J Liu, James Harrison, Jaehoon Lee, Kelvin Xu, et al. Beyond human data: Scaling self-training for problem-solving with language models. *arXiv preprint arXiv:2312.06585*, 2023.

[80] Caglar Gulcehre, Tom Le Paine, Srivatsan Srinivasan, Ksenia Konyushkova, Lotte Weerts, Abhishek Sharma, Aditya Siddhant, Alex Ahern, Miaosen Wang, Chenjie Gu, et al. Reinforced self-training (rest) for language modeling. *arXiv preprint arXiv:2308.08998*, 2023.

[81] Audrey Huang, Adam Block, Dylan J Foster, Dhruv Rohatgi, Cyril Zhang, Max Simchowitz, Jordan T Ash, and Akshay Krishnamurthy. Self-improvement in language models: The sharpening mechanism. *arXiv preprint arXiv:2412.01951*, 2024.

[82] Yifan Sun, Yushan Liang, Zhen Zhang, and Jiaye Teng. Theoretical modeling of llm self-improvement training dynamics through solver-verifier gap. *arXiv preprint arXiv:2507.00075*, 2025.

[83] Yuda Song, Hanlin Zhang, Carson Eisenach, Sham Kakade, Dean Foster, and Udaya Ghai. Mind the gap: Examining the self-improvement capabilities of large language models. *arXiv preprint arXiv:2412.02674*, 2024.

[84] Kelly Hong, Anton Troynikov, and Jeff Huber. Context rot: How increasing input tokens impacts llm performance. Technical report, Chroma, July 2025.

[85] Lingrui Mei, Jiayu Yao, Yuyao Ge, Yiwei Wang, Baolong Bi, Yujun Cai, Jiazhi Liu, Mingyu Li, Zhong-Zhi Li, Duzhen Zhang, et al. A survey of context engineering for large language models. *arXiv preprint arXiv:2507.13334*, 2025.

[86] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

[87] Alessandro Rinaldo. 36-755: Advanced statistical theory, lecture 27. Lecture notes, December 5 2016. Scribed by Xiao Hui Tai.

[88] Martin Raab and Angelika 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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes] ,

   Justification: Main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper discusses the limitations by acknowledging the specific assumptions and settings in Sections 3 and 4 and appendix under which their methods and results hold.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.

- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: For each theoretical result, the paper provides the full set of assumptions and includes a complete and correct proof in the supplementary appendix.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: The main contribution of this work is primarily theoretical findings without any experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.

(c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: The main contribution of this work is primarily theoretical findings without any experimental results.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [NA]

   Justification: The main contribution of this work is primarily theoretical findings without any experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [NA]

Justification: The main contribution of this work is primarily theoretical findings without experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [NA]

   Justification: The main contribution of this work is primarily theoretical findings without experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper conforms with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [NA]

Justification: This paper contributes to advancing the field of Theoretical Machine Learning. We do not foresee any immediate societal implications arising from this work that warrant specific discussion in this context.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: This paper poses no such risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [NA]

Justification: This paper does not use existing assets.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing nor research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

Appendix: Complete Proofs

A Proof Overview

In this section we outline the proof ideas of the main theorem. Our training scheme alternates two phases: we first train the FFN parameters $\mathbf{W}$ to solve the one-step task $\mathcal{T}^1$, and then train the attention parameters $\mathbf{Q}$ to solve $\mathcal{T}^2$ and, recursively, $\mathcal{T}^{2^k}$. This mirrors the high-level division of labor in our setting: the FFN layer learns the *local update rule* (the group operation), while the attention layer learns to *route and compose* these updates by locating the relevant context over long sequences.

- **Task $\mathcal{T}^1$.** To predict the fourth token in the answer clause $\mathbf{Z}_{\mathsf{ans},1}$, namely x_1, the model must retrieve the correct variable from the first token of the predicate clause $\mathbf{Z}_{\mathsf{pred},1}$. This retrieval can be realized directly by the FFN. To predict the fifth token $y_1 = g_1(y_0)$ in $\mathbf{Z}_{\mathsf{ans},1}$, the model needs g_1 from $\mathbf{Z}_{\mathsf{pred},1}$ and y_0 from $\mathbf{Z}_{\mathsf{ans},0}$. Since the sequence $\mathbf{Z}^{1,0}$ contains no additional clauses, uniform attention suffices to expose these two tokens to the FFN, so dedicated attention mechanisms are unnecessary here.
- **Task $\mathcal{T}^\ell$ with $\ell > 1$.** To predict the fifth token $y_\ell = g_\ell(y_{\ell-1})$ in $\mathbf{Z}_{\mathsf{ans},\ell}$, the FFN has already learned the one-step update. Thus, if the attention layer can *route* the correct context to the FFN input, namely g_ℓ from $\mathbf{Z}_{\mathsf{pred},\ell}$ and $y_{\ell-1}$ from $\mathbf{Z}_{\mathsf{ans},\ell-1}$, the FFN outputs the correct y_ℓ. Unlike $\mathcal{T}^1$, attention must now consistently locate *two* dispersed sources across a long context, a task that cannot be handled by the FFN alone.

We adopt this alternating schedule to streamline the analysis and expose the core mechanisms; it is not essential to the algorithm itself. This separation clarifies the complementary roles of FFN and attention in our synthetic CoT setting. Guided by this picture, the proof proceeds in three parts: (1) learning the one-step mechanism for the LEGO task $\mathcal{T}^1$, including in-context variable retrieval (Section A.1) and group operations (Section A.2); (2) establishing direct short-to-polynomial CoT-length generalization on $\mathcal{T}^2$ under simply transitive actions (Section A.3.1); and (3) proving recursive length generalization via self-training on $\mathcal{T}^{2^k}$ under symmetric-group actions (Section A.3.2).

To control rare large deviations in the logits during training-time analysis, we additionally adopt a bounded-output assumption stated below.

Assumption A.1 (Logit clipping)**.** There exists $B = C_B \log d$ for a sufficiently large constant $C_B > 0$ such that each coordinate of the raw model output F_i is clipped from above:

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

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

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

Assumption A.2 (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, zeroing most inter-token attention, is standard in recent theoretical analyses of transformer training dynamics [69, 70, 86, 74]. 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.

Notations For each $i \in [5]$, $r \in [m]$, $j \in [d]$, define

$$\Lambda_{i,j,r}\big(\mathbf{Z}^{L,\ell-1}\big) \triangleq \sum_{\mathbf{k}\in\mathcal{I}^{L,\ell-1}} \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathbf{k}} \cdot \langle \mathbf{W}_{i,j,r}, \mathbf{Z}_{\mathbf{k}} \rangle + b_{i,j,r}. \tag{13}$$

According to (3), given $\mathbf{Z}^{L,\ell-1}$, the model output at token position i and vocabulary index j can be written as

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

A.1 Learning In-Context Retrieval of Variables

For task $\mathcal{T}^1$, to predict the fourth token in the answer clause $\mathbf{Z}_{\mathsf{ans},1}$, we specialize (14) at $i = 4$ and the stage-1 representation $\mathbf{Z}^{1,0}$:

$$\big[F_4\big(\mathbf{Z}^{1,0}\big)\big]_j \;=\; \sum_{r\in[m]} \mathbf{sReLU}\Big(\Lambda_{4,j,r}\big(\mathbf{Z}^{1,0}\big)\Big), \qquad \forall j \in [d].$$

In stage 1, attention from the answer position splits uniformly between the first predicate token and the prior answer token, so

$$\Lambda_{4,j,r}\big(\mathbf{Z}^{1,0}\big) = \tfrac{1}{2}\left\langle \mathbf{W}_{4,j,r},\, \mathbf{Z}_{\mathsf{pred},1}\right\rangle + \tfrac{1}{2}\left\langle \mathbf{W}_{4,j,r},\, \mathbf{Z}_{\mathsf{ans},0}\right\rangle + b_{4,j,r},$$

which implies that the prediction depends only on the parameters $\{\mathbf{W}_{4,j,r}\}_{j\in[d],\, r\in[m]}$. Write $\mathbf{W}_{4,j,r} = [\mathbf{W}_{4,j,r,1}; \ldots; \mathbf{W}_{4,j,r,5}]$, aligned with the five token-type subspaces, and recall that $e_{\tau(\cdot)}$ denote the token embedding vectors. Hence,

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

Our gradient analysis tracks the magnitudes of $\langle \mathbf{W}_{4,j,r,p}, e_{s'}\rangle$ for $j, s' \in [d]$, $p \in [5]$, and $r \in [m]$. We will show that the learning signal is concentrated on the diagonal coordinate that retrieves the correct variable from $\mathbf{Z}_{\mathsf{pred},1}$ and reuses it at $\mathbf{Z}_{\mathsf{ans},1}$. Specifically, let $s = \tau(x_1)$ be the correct index for the current example. Then the gradient component along the correct diagonal, $\langle \mathbf{W}_{4,s,r,1}, e_s\rangle$, enjoys a strictly larger update than any other coordinate because x_1 co-occurs as an input feature at $\mathbf{Z}_{\mathsf{pred},1}$ and as the target at $\mathbf{Z}_{\mathsf{ans},1}$. In contrast, all non-target coordinates (off-diagonals, wrong variables, value tokens, group-action tokens) receive only negligible gradients and remain $o(1)$ throughout. This asymptotic advantage makes the signal order-wise larger than competing coordinates and thus dominates the learning dynamics. Consequently, the *active diagonal mass* $\sum_r \langle \mathbf{W}_{4,s,r,1}, e_s\rangle$ grows and concentrates, reaching $\Omega(\log d)$ for each $s \in \tau(\mathcal{X})$. Consequently, the model learns a robust in-context retrieval pathway that routes the variable from $\mathbf{Z}_{\mathsf{pred},1}$ to the fourth token of $\mathbf{Z}_{\mathsf{ans},1}$, while spurious coordinates remain negligible.

A.2 Learning Group Operations

We sketch how the model learns to track simply transitive and symmetric group operations. **Cyclic (simply transitive) case.** We start with the cyclic (simply transitive) case, and introduce some useful notations.

Definition A.1 (Feature Combinations, Cyclic Group)**.** Suppose the group $\mathcal{G}$ satisfies Assumption C.1. For each $j \in \tau(\mathcal{Y})$, let

$$\mathfrak{F}_j := \{(g, y) \in \mathcal{G} \times \mathcal{Y} \mid \tau(g(y)) = j\}.$$

We call $\mathfrak{F} = \bigcup_{j\in\tau(\mathcal{Y})} \mathfrak{F}_j$ the set of *feature combinations*, and each $\mathfrak{F}_j$ the set of feature combinations predicting $y = \tau^{-1}(j)$.

Definition A.2 (Neuron Feature Indices, Cyclic Group)**.** Define the set of neuron feature indices as

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

Definition A.3 (ψ, Ψ notation)**.** For $j \in \tau(\mathcal{Y})$, $r \in [m]$, and any $\phi = (g, y)$, set

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

and define the *composite feature magnitude*

$$\Psi_{j,r}(\phi) \;:=\; \tfrac{1}{2}\big(\psi_{j,r}(g) + \psi_{j,r}(y)\big).$$

If $(j, r, \phi) = \mathbf{u} \in \mathcal{U}$, we also write $\Psi_{\mathbf{u}}$. The coefficient $\frac{1}{2}$ reflects the uniform attention on $Z_{\mathsf{pred},1}$ and $Z_{\mathsf{ans},0}$ in stage 1.

Combining with Appendix A.1, the score $\Psi_{j,r}(\phi)$ approximates the logit contribution $\Lambda_{5,j,r}(\mathbf{Z}^{1,0})$ and measures the signal for predicting index j when $g_1 = g$ and $y_0 = y$. For each $\phi \in \mathfrak{F}_j$ there is a unique neuron r whose coordinates $\psi_{j,r}(g)$ and $\psi_{j,r}(y)$ grow monotonically until the composite score $\Psi_{j,r}(\phi)$ reaches $\Omega(\log d)$. After this point, $\Psi_{j,r}(\phi)$ can continue to increase, while for any $g' \neq g$ or $y' \neq y$ the quantities $\psi_{j,r}(g')$ and $\psi_{j,r}(y')$ are driven negative so that the mixed pairs (g, y') and (g', y) are canceled, i.e., $\Psi_{j,r}(\phi') \approx 0$. We denote the activated neuron corresponding to $(g, y) \in \mathfrak{F}_j$ by $r_{g\cdot y}$ and set

$$\mathfrak{A}_j \;:=\; \{\, r :\; \exists\, (g,y) \in \mathfrak{F}_j,\; r = r_{g\cdot y} \,\}.$$

Thus the associated indices $\mathbf{u} = (j, r_{g\cdot y}, (g, y))$ for all $(g, y) \in \mathfrak{F}_j$ and $j \in \tau(\mathcal{Y})$ form a subset $\mathcal{U}^\star \subset \mathcal{U}$, and these $\mathbf{u}$ emerge under GD in an order analogous to a power iteration (via the smooth part of $\mathbf{sReLU}$) determined by their initialization magnitudes:

$$\mathbf{u} \prec \mathbf{u}' \quad\Longleftrightarrow\quad \Psi^{(0)}_{\mathbf{u}} \;\geq\; \Psi^{(0)}_{\mathbf{u}'}.$$

Moreover, by the symmetry of the simply transitive structure, the learned pair $(\psi_{j,r_{g\cdot y}}(g), \psi_{j,r_{g\cdot y}}(y))$ remains nearly balanced throughout training. At the end of stage 1, for any $(g, y) \in \mathfrak{F}_j$ with $j \in \tau(\mathcal{Y})$, we have

$$\tfrac{1}{2}\big(\psi_{j,r_{g\cdot y}}(g) + \psi_{j,r_{g\cdot y}}(y)\big) \;\geq\; B - o(1), \qquad \big|\psi_{j,r_{g\cdot y}}(g) - \psi_{j,r_{g\cdot y}}(y)\big| \;\leq\; o(1), \tag{15a}$$

$$\big|\psi_{j,r_{g\cdot y}}(g) + \psi_{j,r_{g\cdot y}}(y')\big| \;\leq\; o(1), \quad \psi_{j,r_{g\cdot y}}(y') \;<\; 0 \quad \text{for all } y' \neq y, \tag{15b}$$

$$\big|\psi_{j,r_{g\cdot y}}(g') + \psi_{j,r_{g\cdot y}}(y)\big| \;\leq\; o(1), \quad \psi_{j,r_{g\cdot y}}(g') \;<\; 0 \quad \text{for all } g' \neq g, \tag{15c}$$

$$\big|\psi_{j,r}(g)\big|,\; \big|\psi_{j,r}(y)\big| \;\leq\; o(1) \quad \text{for all } r \notin \mathfrak{A}_j. \tag{15d}$$

Symmetric group case. The strategy mirrors the cyclic case—*emergence*, *refinement*, and *convergence*—but symmetric actions create richer interactions because multiple group elements can map the same y to the same j. Accordingly, for each (j, y) we aggregate candidates over the *fiber*

$$\mathrm{Fiber}_{j,y} \;:=\; \{g \in \mathcal{G} :\; \tau(g(y)) = j\}, \qquad n_y \;:=\; |\mathrm{Fiber}_{j,y}|, \qquad \mathfrak{F}_j \;:=\; \{(\mathrm{Fiber}_{j,y}, y) : y \in \mathcal{Y}\}.$$

As in the cyclic case, a single neuron becomes responsible for each target pair (j, y); denote it by $r_{j,y}$ and set $\mathfrak{A}_j := \{\, r_{j,y} : y \in \mathcal{Y} \,\}$.

At convergence, the correct fiber dominates while incompatible compositions are canceled. Quantitatively, for any $j \in \tau(\mathcal{Y})$, $y \in \mathcal{Y}$, and $g \in \mathrm{Fiber}_{j,y}$,

$$\tfrac{1}{2}\big(\psi_{j,r_{g\cdot y}}(g) + \psi_{j,r_{g\cdot y}}(y)\big) \;\geq\; B - o(1), \qquad \big|\, n_y\, \psi_{j,r_{g\cdot y}}(y) - \psi_{j,r_{g\cdot y}}(g) \,\big| \;\leq\; o(1), \tag{16a}$$

$$\big|\psi_{j,r_{g\cdot y}}(g) + \psi_{j,r_{g\cdot y}}(y')\big| \;\leq\; o(1), \quad \psi_{j,r_{g\cdot y}}(y') < 0 \text{ for all } y' \neq y, \tag{16b}$$

$$\big|\psi_{j,r_{g\cdot y}}(g') + \psi_{j,r_{g\cdot y}}(y)\big| \;\leq\; o(1), \quad \psi_{j,r_{g\cdot y}}(g') < 0 \text{ for all } g' \notin \mathrm{Fiber}_{j,y}, \tag{16c}$$

$$\big|\psi_{j,r}(g)\big|,\; \big|\psi_{j,r}(y)\big| \;\leq\; o(1) \text{ for all } r \notin \mathfrak{A}_j. \tag{16d}$$

Different learned structures: symmetric vs. simply transitive Under (16), the model spreads mass across the n_y group elements in the fiber $\mathrm{Fiber}_{j,y} = \{g \in \mathcal{G} : \tau(g(y)) = j\}$, i.e., the preimage of j under $g \mapsto \tau(g(y))$ with y fixed. For the winning neuron $(j, r_{g\cdot y})$ we have

$$\psi_{j,r_{g\cdot y}}(g) \;\approx\; n_y\, \psi_{j,r_{g\cdot y}}(y) \quad \text{and} \quad \tfrac{1}{2}\big(\psi_{j,r_{g\cdot y}}(g) + \psi_{j,r_{g\cdot y}}(y)\big) \;\geq\; B - o(1).$$

Solving these relations yields, uniformly up to $o(1)$ terms,

$$\psi_{j,r_{g\cdot y}}(g) \;=\; 2B - \Theta(B/n_y), \qquad \psi_{j,r_{g\cdot y}}(y) \;=\; \Theta(B/n_y).$$

In contrast, in the simply transitive case (15) the mass does not split across multiple group elements; for the winning neuron, $\psi_{j,r_{g\cdot y}}(g) \;\approx\; \psi_{j,r_{g\cdot y}}(y) \;\approx\; B$.

A.3 Learning the Attention Layer

Successful training on $\mathcal{T}^1$ shows that the model has learned the one-step update $y_1 = g_1(y_0)$. We now turn to the more challenging task $\mathcal{T}^2$. Since the group operation is already implemented by the trained FFN, the main remaining difficulty is *routing*: directing attention to the appropriate locations.

For example, given $\mathbf{Z}^{2,1}$, the model must identify the predicate clause $\mathbf{Z}_{\mathsf{pred},2}$ (which contains g_2) and the answer clause $\mathbf{Z}_{\mathsf{ans},1}$ (which contains y_1) in order to compute $y_2 = g_2(y_1)$.

We show that training induces a routing pattern we call **attention concentration**: given an input $\mathbf{Z}^{L,\ell-1}$, the attention mass concentrates—approximately evenly—on $\mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{pred},\ell}$ and $\mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{ans},\ell-1}$. We quantify routing quality via the *attention concentration degree*

$$\epsilon_{\mathsf{attn}}^{L,\ell}\big(\mathbf{Z}^{L,\ell-1}\big) \;=\; 1 - \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{pred},\ell}\big(\mathbf{Z}^{L,\ell-1}\big) - \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{ans},\ell-1}\big(\mathbf{Z}^{L,\ell-1}\big)\,, \tag{17}$$

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

$$\Delta^{L,\ell}\big(\mathbf{Z}^{L,\ell-1}\big) \;=\; \Big|\mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{pred},\ell}\big(\mathbf{Z}^{L,\ell-1}\big) - \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{ans},\ell-1}\big(\mathbf{Z}^{L,\ell-1}\big)\Big|, \tag{18}$$

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

This behavior is tightly linked to the structure of the query matrix $\mathbf{Q}$ and the clause embeddings. Under Assumption A.2, for an input $\mathbf{Z}^{L,\ell-1}$ the (unnormalized) attention score from $\mathbf{Z}_{\mathsf{ans},\ell-1}$ to a clause $\mathbf{Z}_{\mathbf{k}}$ decomposes as

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

where $\mathbf{Z}_{\mathbf{k}} = [\mathbf{Z}_{\mathbf{k},1},\ldots,\mathbf{Z}_{\mathbf{k},5}]$ 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 $\langle\mathsf{blank}\rangle$. Consequently, $\mathbf{Q}_{4,3}$ governs attention to predicate clauses, while $\mathbf{Q}_{4,4}$ governs attention to answer clauses.

The key observation is that the desired allocation can be realized by growing the diagonal entries $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$ for $s\in\tau(\mathcal{X})$. This holds because the same variable x_ℓ appears as the third token in $\mathbf{Z}_{\mathsf{pred},\ell}$ and as the fourth token in $\mathbf{Z}_{\mathsf{ans},\ell-1}$, creating a strong co-occurrence signal. As a result, these diagonal coordinates receive asymptotically larger gradient magnitudes than all other entries, and the training dynamics are dominated by their growth. 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 $\epsilon_{\mathsf{attn}}^{L,\ell}$ 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.

A.3.1 Simply Transitive Group

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

– **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

$$\Lambda_{5,j,r}(\mathbf{Z}^{2,0}) \;\approx\; \tfrac{1}{3}\,\psi_{j,r}(g_1) + \tfrac{1}{3}\,\psi_{j,r}(g_2) + \tfrac{1}{3}\,\psi_{j,r}(y_0)\,,$$
$$\Lambda_{5,j,r}(\mathbf{Z}^{2,1}) \;\approx\; \tfrac{1}{4}\,\psi_{j,r}(g_1) + \tfrac{1}{4}\,\psi_{j,r}(g_2) + \tfrac{1}{4}\,\psi_{j,r}(y_0) + \tfrac{1}{4}\,\psi_{j,r}(y_1)\,.$$

By the cancellation in (15), for $\ell=2$ all Λ's 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, $-\nabla_{\mathbf{Q}}\mathsf{Loss}_5^{2,2}$ is negligible, while $-\nabla_{\mathbf{Q}}\mathsf{Loss}_5^{2,1}$ 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 $\mathsf{Loss}_5^{2,1}$). 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 $\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}$ 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 $\epsilon_{\mathsf{attn}}^{2,2}$ small. We show $\Delta^{2,2}$ cannot remain above $o(1)$ for long; otherwise an incorrect logit $\mathbf{logit}_{5,\tau(g_2(y_0))}$ would acquire a stronger gradient and force $\mathbf{Q}_{4,4}$ to outpace $\mathbf{Q}_{4,3}$, which contradicts stability. At convergence: (i) $\epsilon_{\mathsf{attn}}^{2,2} \le 1/\mathsf{poly}(d)$; (ii) both $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ reach $\Omega(\log d)$; and (iii) $\Delta^{2,2} \le o(1)$.

Direct short-to-long length generalization. The key is that attention concentrates cleanly: $\epsilon_{\mathsf{attn}}^{2,2} \le 1/d^{\Omega(1)}$ with $\mathbf{Q}_{4,3}, \mathbf{Q}_{4,4} = \Omega(\log d)$, so the model tolerates a polynomial number of irrelevant clauses. Hence for $\mathcal{T}^L$ with $L \le \mathsf{poly}(d)$,

$$\epsilon^{L,\ell} \;\le\; \frac{O(1)\cdot L}{O(1)\cdot L + \exp\big([\mathbf{Q}_{4,3}]_{\tau(x_\ell),\tau(x_\ell)}\big) + \exp\big([\mathbf{Q}_{4,4}]_{\tau(x_\ell),\tau(x_\ell)}\big)} \;=\; o(1),$$

and moreover $\Delta^{L,\ell} \le \Delta^{2,2} \le o(1)$. Together these imply

$$1 - \mathbf{logit}_{5,\tau(g_{\ell+1}(y_\ell))}\Big(F^{(T^\star)}, \mathbf{Z}^{(L,\ell)}\Big) \;\le\; \frac{O(1)\cdot d + e^{o(1)}}{O(1)\cdot d + e^{o(1)} + e^{\Omega(\log d)}} \;\le\; \frac{1}{\mathsf{poly}(d)},$$

so $\mathcal{T}^L$ is solved with accuracy $1 - 1/\mathsf{poly}(d)$.

A.3.2 Symmetry Group

We now turn to symmetry-group tasks $\mathcal{T}^L$ and analyze GD updates with respect to the per-token loss $\mathsf{Loss}_5^{L,2}$ (i.e., predicting the value token in $\mathbf{Z}_{\mathsf{ans},2}$ 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 FFN structure (16) spreads mass across the n_y preimages. Thus the roles of $\psi_{j,r_{g\cdot y}}(g)$ and $\psi_{j,r_{g\cdot y}}(y)$ are unbalanced: for the winning neuron $(j, r_{g\cdot y})$ with $g \in \mathrm{Fiber}_{j,y}$, we have $\psi_{j,r_{g\cdot y}}(g) \approx n_y\, \psi_{j,r_{g\cdot y}}(y)$. This makes 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) $\epsilon_{\mathsf{attn}}^{2,2} \le C_1$; (ii) $\Delta^{2,2} \le C_2$, for sufficiently small constants C_1, C_2; and since $B = \Theta(\log d)$, we still obtain $\mathsf{Loss}_5^{2,2} \le 1/\mathsf{poly}(d)$.

Recursive learning for $\mathcal{T}^{2^k}$, $k \ge 2$. Because $\epsilon_{\mathsf{attn}}^{2^{k-1},2}$ is already a small constant, initialization for $\mathcal{T}^{2^k}$ satisfies $\epsilon_{\mathsf{attn}}^{2^k,2} \le 2\,\epsilon_{\mathsf{attn}}^{2^{k-1},2}$ (still small), and $\Delta^{2^k,2} \le \Delta^{2^{k-1},2}$. 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 $\widehat{p}_{F^{(T_{k-1})}}$, 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 $\epsilon_{\mathsf{attn}}^{2^k,2}$ and $\Delta^{2^k,2}$ decrease to a small constant, yielding stable, inductive concentration across recursive reasoning depths.

B Learning In-Context Retrieval of Variables

B.1 Preliminaries

First we define some notations for the presentation of gradients.

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

$$\mathcal{E}_{i,j}(\mathbf{Z}^{L,\ell-1}) \triangleq \mathbb{1}_{\mathbf{Z}_{\mathsf{ans},\ell,i}=e_j} - \mathbf{logit}_{i,j}(F, \mathbf{Z}^{(L,\ell-1)}),$$
$$\Lambda_{i,j,r}(\mathbf{Z}^{L,\ell-1}) \triangleq \sum_{\mathbf{k}\in\mathcal{I}^{L,\ell-1}} \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathbf{k}} \cdot \big\langle \mathbf{W}_{i,j,r}, \mathbf{Z}_{\mathbf{k}} \big\rangle + b_{i,j,r}.$$

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

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

Fact B.1. For any $i \in [5]$, $j \in [d]$, $r \in [m]$

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

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 $\widehat{\mathcal{X}}^{(L)}$ to denote the appeared variables in the context clauses, i.e. $\widehat{\mathcal{X}}^{(L)} = \{x_0, x_1, \ldots, x_L\}$. We write $\widehat{\mathcal{X}}^{(L)}$ as $\widehat{\mathcal{X}}$ for simplicity. Throughout this section, we write $[F_i]_j$ as $F_{i,j}$ for simplicity.

B.2 Induction Hypothesis

In this stage, we consider the learning process for $\mathbf{W}_{4,\cdot}$.

Induction B.1. *For $t \le T = \frac{\mathsf{poly}d}{\eta}$, all of the following holds:*

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

(b). for $j \in \tau(\mathcal{X})$, $g \in \mathcal{G}$

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

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

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

(d). Else, $\big|\langle \mathbf{W}_{4,j,r,p}^{(t)}, e_s\rangle\big| \le \widetilde{O}(\sigma_0)$;

Claim B.1. If Induction B.1 holds at iteration t, then for a sequence $\mathbf{Z}$

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

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

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

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

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

Claim B.2. If Induction B.1 holds at iteration t, then for a sequence $\mathbf{Z}$,

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

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

$$0 \le F_{4,j}^{(t)}(\mathbf{Z}) \le \sum_{r\in[m]} [\Lambda_{4,j,r}^{(t)}]^+ \le \big(\Phi_{4,j}^{(t)} + O(\frac{\max\{\Phi_{4,j}^{(t)}, \Phi_{4,j^*}^{(t)}\}}{|\mathcal{G}|})\big) + \widetilde{O}(\sigma_0) + O(m\varrho \log d)$$
$$= \big(\Phi_{4,j}^{(t)} + O(\frac{\Phi_{4,j}^{(t)}}{|\mathcal{G}|})\big) + \widetilde{O}(\sigma_0) + O(\frac{1}{\mathrm{polylog} d})$$

for $j \in \tau(\mathcal{X}) \neq \tau(x_1)$, $F_{4,j}^{(t)}(\mathbf{Z}) \le \widetilde{O}(\sigma_0) + O(\frac{\max\{\Phi_{4,j}^{(t)}, \Phi_{4,j^*}^{(t)}\}}{|\mathcal{G}|})$; else $F_{4,j}^{(t)}(\mathbf{Z}) \le \widetilde{O}(\sigma_0)$. Combining together, we prove the result. □

B.3 Gradient Lemma

Starting with the gradient computation:

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

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

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

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

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

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

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

$$\langle -\nabla_{\mathbf{W}_{4,j,r,2}^{(t)}} \mathsf{Loss}^1, e_s\rangle = \frac{1}{2}\mathbb{E}\Big[(1 - \mathbf{logit}_{4,j}^{(t)})\mathbf{sReLU}'\big(\Lambda_{4,j,r}^{(t)}\big)\mathbb{1}_{\tau(x_1)=j,g_1=g}$$
$$- \mathbf{logit}_{4,j}^{(t)}\mathbf{sReLU}'\big(\Lambda_{4,j,r}^{(t)}\big)\mathbb{1}_{\tau(x_1)\neq j,g_1=g}\Big]$$

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

(1) $s = j$, $\langle -\nabla_{\mathbf{W}_{4,j,r,3}} \mathsf{Loss}^1, e_j\rangle = \frac{1}{2}\mathbb{E}\Big[-\mathbf{logit}_{4,j}^{(t)}\mathbf{sReLU}'\big(\Lambda_{4,j,r}^{(t)}\big)\mathbb{1}_{\tau(x_0)=j}\Big]$

(2) $s \neq j$

$$\langle -\nabla_{\mathbf{W}_{4,j,r,3}} \mathsf{Loss}^1, e_s\rangle = \frac{1}{2}\mathbb{E}\Big[(1 - \mathbf{logit}_{4,j}^{(t)})\mathbf{sReLU}'\big(\Lambda_{4,j,r}^{(t)}\big)\mathbb{1}_{\tau(x_0)=s,\tau(x_1)=j}$$
$$- \mathbf{logit}_{4,j}^{(t)}\mathbf{sReLU}'\big(\Lambda_{4,j,r}^{(t)}\big)\mathbb{1}_{\tau(x_0)=s,j\notin\tau(\widehat{X})}\Big]$$

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

$$\langle -\nabla_{\mathbf{W}_{4,j,r,5}^{(t)}} \mathsf{Loss}^1, e_s\rangle = \frac{1}{2}\mathbb{E}\Big[(1 - \mathbf{logit}_{4,j}^{(t)})\mathbf{sReLU}'\big(\Lambda_{4,j,r}^{(t)}\big)\mathbb{1}_{\tau(x_1)=j,y_0=y}$$
$$- \mathbf{logit}_{4,j}^{(t)}\mathbf{sReLU}'\big(\Lambda_{4,j,r}^{(t)}\big)\mathbb{1}_{\tau(x_1)\neq j,y_0=y}\Big]$$

Then for $j \notin \tau(\mathcal{X})$, we can obtain

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

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

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

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

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

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

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

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

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

B.4 Growth of Gamma

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

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

Proof. By Lemma B.1, we have

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

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

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

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

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

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

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

Therefore, for sufficiently small η, we have

$$\begin{aligned}
\Phi_{4,j}^{(t+1)} &= \Phi_{4,j}^{(t)} + \sum_{r \in \mathcal{A}_{4,j}^{(t)}} \frac{\eta}{2}\mathbb{E}\Big[(1 - \mathbf{logit}_{4,j}^{(t)}) \mathbf{sReLU}'\big(\Lambda_{4,j,r}^{(t)}\big) \mathbb{1}_{\tau(x_1) = j}\Big] + O(\varrho \log d) \cdot |\mathcal{A}_{4,j}^{(t+1)} \setminus \mathcal{A}_{4,j}^{(t)}| \\
&= \Phi_{4,j}^{(t)} + \sum_{r \in \mathcal{A}_{4,j}^{(t)}} \frac{\eta}{2}\mathbb{E}\Big[(1 - \mathbf{logit}_{4,j}^{(t)}) \mathbf{sReLU}'\big(\Lambda_{4,j,r}^{(t)}\big) \mathbb{1}_{\tau(x_1) = j}\Big] + \frac{1}{\mathrm{polylog} d}
\end{aligned}$$

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

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

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

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

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

B.5 Group and Value Correlations Are Not Large

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

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

Proof. By Lemma B.1, we have

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

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

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

When $t \leq T_1$, by Claim B.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

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

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

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

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

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

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

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

Due to the arbitrary of r', we have

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

□

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

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

Proof. The proof is similar as Lemma B.6. □

B.6 Off-diagonal Correlations Are Small

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

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

Proof. By Lemma B.1

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

Notice that by Claim B.1,

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

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

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

From Induction B.1, we have

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

□

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

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

Proof. When $s=j$, we have

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

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

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

When $s \neq j$,

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

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 $O(\frac{1}{d})\cdot|\langle -\nabla_{\mathbf{W}_{4,j,r,1}}\mathsf{Loss}^1, e_j\rangle|$. 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. □

B.7 Non-target Correlations Are Negligible

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

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

Proof. By Lemma B.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 $\langle \mathbf{W}_{4,j',r,p}^{(t)}, e_s \rangle \le \langle \mathbf{W}_{4,j',r,p}^{(0)}, e_s \rangle = \widetilde{O}(\sigma_0)$. Moreover, $\Lambda_{4,j',r}^{(t)} \le \widetilde{O}(\sigma_0)$ 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 $\Lambda_{4,j',r}^{(t)}\big|_{g_1=g} \le -3\mu + \frac{5}{2}\mu + \widetilde{O}(\sigma_0) \le 0$, and thus $\langle -\nabla_{\mathbf{W}_{4,j,r,2}^{(t)}} \mathsf{Loss}^1, e_s \rangle = \frac{1}{2}\mathbb{E}\Big[-\mathbf{logit}_{4,j}^{(t)} \mathbf{sReLU}'\big(\Lambda_{4,j,r}^{(t)}\big) \mathbb{1}_{g_1=g} \Big] = 0$, which implies $\langle \mathbf{W}_{4,j',r,2}^{(t)}, e_s \rangle \ge -3\mu$. Hence, $|\langle \mathbf{W}_{4,j',r,2}^{(t)}, e_s \rangle| \le \widetilde{O}(\sigma_0)$. Following the similar argument, we can prove the result for $\langle \mathbf{W}_{4,j',r,5}^{(t)}, e_s \rangle$ for $s \in \tau(\mathcal{Y})$. □

B.8 Convergence

Lemma B.11. *For $|\mathcal{G}| \ge |\mathcal{Y}| \ge \Omega(\frac{\log\log d}{\log\log\log d})$, $\mathrm{polylog} d \ge m \ge |\mathcal{Y}|$, $\varrho \ll \frac{1}{m \log d}$ and sufficiently small $\eta \le \frac{1}{\mathsf{poly} d}$, Induction B.1 holds for all iterations $t \le T = \frac{\mathsf{poly} d}{\eta}$.*

Proof. Putting the results in Lemmas B.5 to B.8 and B.10, we can directly establish the results in Induction B.1. □

Lemma B.12 (Convergence)**.** *For sufficiently large $T_1 \le t = \frac{\mathsf{poly} d}{\eta}$, we have*

(a) Objective convergence: $\mathsf{Loss}^1 \le \frac{1}{\mathsf{poly} d}$;

(b) Successful learning of diagonal feature: $\Phi_{4,j}^{(t)} \ge \Omega(\log d)$ for any $j \in \tau(\mathcal{X})$.

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

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

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

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

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

C Learning The Group Actions: Cyclic Group

In the LEGO language, one step of the state transition corresponds predicting the next answer clause from the current sequence $Z^{L,L'} \sim \mathcal{D}^{L,L'}$. As in Algorithm 1 and 2, we start with training the model F on length-1 sequences, which only requires the model to predict one answer $Z_{\mathsf{ans},1}$ given input clauses $Z_{\mathsf{pred},1}$ and $Z_{\mathsf{ans},0}$. In this appendix section, we show how the model learns to predict the 5-th token of $Z_{\mathsf{ans},1}$, that is, the value of x_1, in the LEGO sentence $Z^{(1)}$ in (2).

Let's recall the structure of the uniform case setting. Note that the group action defined in Assumption 4.1 is equivalent to the following group action by group isomorphism:

Assumption C.1 (Assumption 4.1, restated)**.** Let $\mathbf{LEGO}(\mathcal{X}, \mathcal{G}, \mathcal{Y})$ be the LEGO language. We assume $\mathcal{Y} = \{0, 1, \ldots, n_y - 1\}$ and $\mathcal{G} = C_{|\mathcal{Y}|}$, i.e., the cyclic group of order $|\mathcal{Y}|$, and $n_y \in [\Omega(\log \log d), \log d]$.

We define some notations for this section here.

Notations. Let $\mathcal{D}^1$ be the LEGO distribution of length 1 under the language $\mathbf{LEGO}(\mathcal{X}, \mathcal{G}, \mathcal{Y})$. We define $\mathcal{D}^1_{\mathcal{X}}$, $\mathcal{D}^1_{\mathcal{G}}$ and $\mathcal{D}^1_{\mathcal{Y}}$ be the distribution of (x_0, x_1), g_0 and (y_0, y_1) in $\mathcal{D}^1$ respectively. That is, given a LEGO sentence

$$Z^{(1,0)} = (Z_{\mathsf{pred},1}, Z_{\mathsf{ans},0}, Z_{\mathsf{ans},1}) \sim \mathcal{D}^1,$$
$$Z_{\mathsf{pred},1} = (x_0, g_1, x_1, \langle \mathsf{blank} \rangle, \langle \mathsf{blank} \rangle), \quad Z_{\mathsf{ans},i} = (\langle \mathsf{blank} \rangle, \langle \mathsf{blank} \rangle, \langle \mathsf{blank} \rangle, x_i, y_i), i \in \{0, 1\}$$

The sampling distribution of (x_0, x_1) is $\mathcal{D}^1_{\mathcal{X}}$, and similarly for g_0 and (y_0, y_1).

C.1 Preliminaries and Induction Hypotheses

First we compute the expression of gradients for $\mathbf{W}$ here. With slight abuse of notation, we write $\mathsf{Loss}^{(t)} \equiv \mathsf{Loss}(F^{(t)}) \equiv \mathsf{Loss}^{1,0}(F^{(t)})$ in this stage. The gradients are given by:

$$\langle \nabla_{\mathbf{W}_{5,j,r,p}} \mathsf{Loss}^{(t)}, e_v \rangle$$
$$= \mathbb{E}_{\mathbf{Z}^1 \sim \mathcal{D}^1} \left[\mathcal{E}_{i,j}(\mathbf{Z}^1) \cdot \mathbf{sReLU}'(\Lambda^{(t)}_{5,j,r}(\mathbf{Z}^{1,0})) \sum_{\mathbf{k} \in \mathcal{I}^{1,0}} \mathbb{1}_{Z_{\mathbf{k},p}=v} \right], \quad j \in [d], p \in [5], r \in [m], v \in \mathcal{V}$$

where $\mathcal{E}_{i,j}(\mathbf{Z}^1) = \mathbb{1}_{\tau(Z_{\mathsf{ans},1,i})=j} - \mathbf{logit}_{i,j}(F, \mathbf{Z}^{1,0})$ is the

To analyze the learning of the group actions, we need to first define the set of features that the model will learn, corresponding to the group action.

Definition C.1 (feature combinations, cyclic group)**.** Asuming the group $\mathcal{G}$ follows Assumption C.1. For each $j \in \tau(\mathcal{Y})$, let

$$\mathfrak{F}_j := \{(g, y) \in \mathcal{G} \times \mathcal{Y} \mid \tau(g(y)) = j\}$$

we call the set $\mathfrak{F} = \bigcup_{j \in \tau(\mathcal{Y})} \mathfrak{F}_j$ the set of **feature combinations**, and the sets $\mathfrak{F}_j$ are called set of feature combinations for predicting $y = \tau^{-1}(j)$. Furthermore, for any $\phi = (g, y) \in \mathfrak{F}$, we write

$$\mathfrak{F}_{\mathsf{conf}}(\phi) := \{\phi' \in \mathfrak{F} \mid \phi' = (g', y) \text{ or } \phi' = (g, y'), \text{ where } g \neq g', y \neq y'\}$$

as the set of **confounding features** for ϕ, that is, the features that share exactly one component with ϕ, either g or y.

Each $\mathfrak{F}_j$ includes all possible combinations of g, y that transform $y \in \mathcal{Y}$ to a new state $y' = g \cdot y$ by predicting the corresponding token index $j = \tau(y')$. It is the set of features we want our network to learn in the neurons of output coordinate j, while sets $\mathfrak{F}_{j'}, j' \neq j$ are the sets of features we do not want to learn in the neurons of output coordinate j'.

The set of confounding features $\mathfrak{F}_{\mathsf{conf}}(\phi)$ for a given feature $\phi = (g, y) \in \mathfrak{F}$ contains the features that share exactly one component with ϕ, either g or y. Confounding features are the combination of features that are similar to ϕ but are incorrect for predicting j-th output.

Features in sets $\mathfrak{F}_j$ exist in the neurons. We define a short notation for the set of all indices of feature combinations at coordinate $j \in \tau(\mathcal{Y})$ and neuron $r \in [m]$.

Definition C.2 (neuron feature indices)**.** We define

$$\mathcal{U} = \{\mathbf{u} = (j, r, \phi) \mid j \in \tau(\mathcal{Y}), r \in [m], \phi \in \mathfrak{F}\}$$

be the set of all indices of compositional features.

Now we present the some notations that could help us determine which features are learned first.

Definition C.3 (ψ, Ψ-notations)**.** Let $j \in \tau(\mathcal{Y})$ and $r \in [m]$, for $\phi = (g, y) \in \mathfrak{F}_j$, Let's define the following notation:

$$\psi_{j,r}(g) := \langle \mathbf{W}_{5,j,r,2}, e_g \rangle, \quad \psi_{j,r}(y) := \langle \mathbf{W}_{5,j,r,5}, e_y \rangle \tag{19}$$

When cG follows Assumption C.1, we define an index $\mathbf{u} = (j, r, \phi)$ and a feature magnitude $\Psi_{\mathbf{u}} \equiv \Psi_{j,r}(g, y)$ as follows:

$$\Psi_{\mathbf{u}} \equiv \Psi_{j,r}(\phi) := \frac{1}{2}(\psi_{j,r}(g) + \psi_{j,r}(y))$$

which is the combination of features we need to predict the correct answer $y' = g \cdot y$ with fixed attention weights $\frac{1}{2}$ for both $Z_{\mathsf{pred},1}$ and $Z_{\mathsf{ans},0}$ during training phase I.

A key technical ingredient of our proof is the characterization of the learning order of the features. By leveraging the smoothness of the $\mathbf{sReLU}$ activationfunction, we can show that the features are learned in a specific order that relates to the feature magnitude $\Psi_{\mathbf{u}}^{(0)}$ at initialization. We define the learning order, encoded by the order $\prec$ on $\mathcal{U}^\star$ as follows:

Definition C.4 (learning order)**.** The *learning order* is the ordered set $\mathcal{U}^\star$ that we obtain from the following process: Define a total order on $\mathcal{U}$ as follows:

$$\mathbf{u} \prec \mathbf{u}' \iff \Psi_{\mathbf{u}}^{(0)} \geq \Psi_{\mathbf{u}'}^{(0)} \quad \forall \mathbf{u}, \mathbf{u}' \in \mathcal{U} \tag{20}$$

We construct the sets $\mathcal{U}^\star$ by the following procedure: initialize an empty neuron set $\mathcal{W}_{tmp}^{(0)} = \varnothing$, and an empty feature set $\mathcal{R}_{tmp}^{(0)} = \varnothing$, and the initial index set $\mathcal{U}^{(0)} = \varnothing$. Starting from $k = 1$, we do the following:

(1) Find the index $\mathbf{u} = (j, r, \phi) \in \mathcal{U}$, where $(j, r, \phi) = \arg\max_{j',r',\phi'} \Psi_{j',r'}^{(0)}(\phi')$ such that the feature $\phi \in \mathfrak{F} \setminus \mathcal{R}_{tmp}^{(k-1)}$ and $(j, r) \in \tau(\mathcal{Y}) \times [m] \setminus \mathcal{W}_{tmp}^{(k-1)}$.

(2) Update $\mathcal{R}_{tmp}^{(k)} \leftarrow \mathcal{R}_{tmp}^{(k-1)} \cup \{\phi\}$, $\mathcal{W}_{tmp}^{(k)} \leftarrow \mathcal{W}_{tmp}^{(k-1)} \cup \{(j, r)\}$, and $\mathcal{U}^{(k)} \leftarrow \mathcal{U}^{(k-1)} \cup \{\mathbf{u}\}$.

(3) Iterate the (1) and (2) steps until $k = n_y^2$, then yield $\mathcal{U}^\star \equiv \mathcal{U}^{(n_y^2)}$.

This process yields the ordered set $\mathcal{U}^\star$, equipped with the total order $\prec$ defined in (20).

Note that $\mathcal{U}^\star$ is a smaller subset of $\mathcal{U}$. In fact, $\mathcal{U}^\star$ encodes the order that the neural network F learn the features ϕ in the neurons $(j, r) \in [d] \times [m]$, and leave out the indices $\mathbf{u} \in \mathcal{U} \setminus \mathcal{U}^\star$ that are not learned. The order $\prec$ is induced by the feature magnitude $\Psi_{j,r}^{(0)}(\phi)$. In the proof we shall show with high probability, the order that each ϕ is learned is according to its position in $\mathcal{U}^\star$.

In order to characterize the feature updates before they start to become significant, we define the following notion of *pseudo weights*.

Definition C.5 (pseudo weights)**.** Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}$. We define the pseudo weight $\widetilde{\mathbf{W}}^{(t)}(\mathbf{u})$ where $\widetilde{\mathbf{W}}_{i,j',r'}^{(t)}(\mathbf{u}) \equiv \mathbf{W}_{i,j',r'}^{(t)}$ for all $(i, j', r') \in [5] \times [d] \times [m]$ except for when $i = 5$. We initialize $\widetilde{\mathbf{W}}_{5,j,r}^{(0)}(\mathbf{u}) \equiv \mathbf{W}_{5,j,r}^{(0)}$, and let $\widetilde{\Lambda}_{5,j,r}^{(t)}$ be the corresponding activation with weights $\widetilde{\mathbf{W}}_{5,j,r}^{(t)}$. We define the update rule of $\widetilde{\mathbf{W}}_{5,j,r}^{(t)}(\mathbf{u})$ in the following manner:

- if $(p, v) \notin \{(2, g), (5, y)\}$, we define $\langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t+1)}(\mathbf{u}), e_v \rangle \equiv \langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v \rangle$;
- if $(p, v) \in \{(2, g), (5, y)\}$, then we let the update rule to be:

$$\langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t+1)}(\mathbf{u}), e_v \rangle = \langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t)}(\mathbf{u}), e_v \rangle + \eta \mathbb{E}[\mathbf{sReLU}'(\widetilde{\Lambda}_{5,j,r}^{(t)}) \mathbb{1}_{\mathcal{B}_\phi} \mathbb{1}_{F_{5,j} \leq B}]$$

This means that we are updating the pseudo weights for the 2nd and 5th components of the weight vectors differently, while keeping all other components' updates unchanged.

In order to introduce the induction hypothesis, we first define the probability events of feature appearance.

Definition C.6 (probability events of feature appearance)**.** Let $j \in \tau(\mathcal{Y})$, $r \in [m]$ and $\phi = (g, y) \in \mathfrak{F}_j$ be a pair that predicts j-th output. Denote events $\mathcal{B}_\phi, \mathcal{B}(g,y), \mathcal{B}_j(g), \mathcal{B}_j(y)$ and $\widetilde{\mathcal{B}}_\phi, \widetilde{\mathcal{B}}_j(g), \widetilde{\mathcal{B}}_j(y)$ as follows:

1. $\mathcal{B}_\phi \equiv \mathcal{B}(g,y) \equiv \mathcal{B}_j(g) \equiv \mathcal{B}_j(y) := \{g_1 = g, y_0 = y\}$;
2. $\widetilde{\mathcal{B}}_j(g) := \{g_1 = g, y_0 \neq y\}$, the event that g is the incorrect feature for predicting j-th output;
3. $\widetilde{\mathcal{B}}_j(y) := \{g_1 \neq g, y_0 = y\}$, the event that y is the incorrect feature for predicting j-th output;
4. $\widetilde{\mathcal{B}}_\phi := \widetilde{\mathcal{B}}_j(g) \cup \widetilde{\mathcal{B}}_j(y)$, the event that $\phi = (g, y)$ is the incorrect feature for predicting j-th output.

We then define the notion of the gradient conditions, which will be used in the proof of the induction.

Definition C.7 (gradient criterion)**.** Let $\mathbf{u} = (j, r, \phi) \in \mathcal{O}^\star$ and $t \leq T_1$, we define the following two conditions:

- Given $\delta > 0$, the positive gradient criterion $\mathcal{K}_{\mathsf{pos}}(\mathbf{u}, \delta)$:

$$\mathcal{K}_{\mathsf{pos}}(\mathbf{u}, \delta) \text{ is true } \iff \left|\mathbb{E}\left[(1 - \mathbf{logit}_{5,j}) \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\mathcal{B}_\phi} \mathbb{1}_{F_{5,j} \leq B}\right]\right| \leq \delta \tag{21}$$

- Given $\delta > 0$, the negative gradient criterion $\mathcal{K}_{\mathsf{neg}}(\mathbf{u}, \delta)$:

$$\mathcal{K}_{\mathsf{neg}}(\mathbf{u}, \delta) \text{ is true } \iff \left|\mathbb{E}\left[\mathbf{logit}_{5,j} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\widetilde{\mathcal{B}}_\phi} \mathbb{1}_{F_{5,j} \leq B}\right]\right| \leq \delta \tag{22}$$

These conditions control the magnitude of the gradient of the feature $\mathbf{u}$ at iteration t. Now we define the following intermediate time-steps:

Definition C.8 (phase decomposition)**.** Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$, we define the following intermediate time-steps:

$$\begin{aligned}
T_{\mathbf{u},1a} &:= \min\{t \geq 0 \mid \Psi_{\mathbf{u}}^{(t)} \geq d^{0.01}\sigma_0\} \\
T_{\mathbf{u},1} &:= \min\{t \geq 0 \mid \mathcal{K}_{\mathsf{pos}}(\mathbf{u}, \delta_1), \text{ where } \delta_1 := 1 - d^{-0.1}\} \\
T_{\mathbf{u},2a} &:= \min\left\{t \geq 0 \mid \mathcal{K}_{\mathsf{pos}}(\mathbf{u}, \delta_2) \wedge \mathcal{K}_{\mathsf{neg}}(\mathbf{u}, \delta_2), \text{ where } \delta_2 := \boldsymbol{\lambda}/d^{1.1}\right\} \\
T_{\mathbf{u},2} &:= \min\left\{t \geq 0 \mid \mathcal{K}_{\mathsf{pos}}(\mathbf{u}, \delta_3) \wedge (\Psi_{\mathbf{u}}^{(t)} \geq B - O(d^{0.001}\sigma_0)), \text{ where } \delta_3 := (d^{0.001}\sigma_0)^{q-2}\right\}
\end{aligned}$$

Below we shall introduce some induction hypotheses that will be used in the proof. We first introduce a induction hypothesis for the pseudo weights.

Induction C.1 (induction on pseudo weight bounds)**.** *Let $j \in \tau(\mathcal{Y})$ and $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$, $\phi = (g, y) \in \mathfrak{F}_j$. Let $\mathbf{u}' \prec \mathbf{u} \in \mathcal{U}^\star$ be the immediate predecessor of $\mathbf{u}$ in $\mathcal{U}^\star$. Then at $t = T_{\mathbf{u}',2}$, it holds that the pseudo weights $\widetilde{\mathbf{W}}_{5,j,r}^{(t)}(\mathbf{u})$ defined in Definition C.5 satisfies*

$$\left|\langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t)}(\mathbf{u}), e_v\rangle - \langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v\rangle\right| \leq \widetilde{O}(\sigma_0/d), \quad \forall (p, v) \in \{(2, g), (5, y)\}$$

We maintain the following induction hypotheses for the case of Assumption 4.1.

Induction C.2 (induction on weight bounds)**.** *Assuming Assumption 4.1, for $t \leq T_1$, the following properties hold:*

(A) $|\langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v\rangle - \langle \mathbf{W}_{5,j,r,p}^{(0)}, e_v\rangle| \leq \widetilde{O}(1/d)$ *for all $v \in \mathcal{X}$ and $p \in \{1, 3, 4\}$;*

(B) $|\langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v\rangle - \langle \mathbf{W}_{5,j,r,p}^{(0)}, e_v\rangle| \leq \widetilde{O}(\sigma_0/d)$ *for all $v \in \mathcal{V}$ and $p \in [5], r \in [m]$ if $j \notin \tau(\mathcal{Y})$;*

The last induction hypothesis is about the checkpoints defined in Definition C.8.

Induction C.3 (induction on cyclic group actions)**.** *Under Assumption C.1, for $t \leq T_1$, in addition to the induction hypotheses in Induction C.1 and C.2, the following properties hold:*

(A) For any $\mathbf{u}' \prec \mathbf{u} \in \mathcal{U}^\star$, it holds that $T_{\mathbf{u},1a} \geq T_{\mathbf{u}',2} + \widetilde{\Omega}(1/\eta\sigma_0^{q-2})$;

(B) Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$, for any $t \in [T_{\mathbf{u},2}, T_1]$, it holds that $\Psi_{j,r}^{(t)}(\phi) \geq B - O(d^{0.01}\sigma_0)$ and

$$\max_{\phi' \in \mathfrak{F}_{\mathsf{conf}}(\phi)} \Psi_{j,r}^{(t)}(\phi') \leq ((d^{0.01}\sigma_0)^{q-2} d/\boldsymbol{\lambda})^{1/(q-1)}$$

C.2 Technical Lemmas

Here is a simple fact for the logits of 5-th token predictors. The proofs are trivial.

Fact C.1 (non-updating weights)**.** For $i = 5$, $j \in [d]$, $r \in [m]$ and $p \in [5]$, the following components of $\mathbf{W}_5$, i.e., $\langle \mathbf{W}_{i,j,r,p}, e_v \rangle$ would not be updated:

- $j \in [d], p \in \{1, 3, 4\}, v \notin \mathcal{X}$;
- $j \in [d], p = 2, v \notin \mathcal{G}$;
- $j \in [d], p \in [5], v \notin \mathcal{Y}$.

We also have some unconditional bounds on logits.

Fact C.2 (unconditional logit bounds)**.** Due to Assumption A.1, for the logits of 5-th token predictors, we have the following simple facts:

(a) Let $j \in [d]$ and $i \in [5]$, and suppose $F_{i,j}(\cdot) \leq B$, then

$$1 - \mathbf{logit}_{i,j}(F, \mathbf{Z}) \geq \frac{d-1}{d-1+e^B} =: \boldsymbol{\lambda}, \quad \forall \mathbf{Z} \in \mathrm{supp}(\mathcal{D}^1(\mathcal{Z}))$$

(b) Let $j \in \tau(\mathcal{Y})$ and input $\mathbf{Z} \in \mathrm{supp}(\mathcal{D}^1(\mathcal{Z}))$. Suppose there are no more than n coordinates in $[d]$ such that $F_{5,j}(\mathbf{Z}) \geq \frac{1}{\log d}$, then as long as $e^B \gg d$, it holds that

$$\mathbf{logit}_{5,j}(F, \mathbf{Z}) \geq \frac{1}{O(d) + ne^B} \geq \Omega(\boldsymbol{\lambda}/dn)$$

This is the logit lower bound for the prediction of the j-th head of the language model.

Lemma C.1 (gradient bounds)**.** *Let $j \in \tau(\mathcal{Y})$, and $\phi = (g, y) \in \mathfrak{F}$. Suppose Induction C.3 is satisfied at $t \leq T_1$, then for any $\delta \in [0, 0.4]$, if $\sum_{r\in[m]} \Psi_{j,r}^{(t)}(\phi) \leq (0.5 + \delta)\log d$, it holds that*

(a) $\mathbb{E}[(1 - \mathbf{logit}_{5,j}^{(t)}) \mid \mathcal{B}_\phi] \geq 1 - d^{-0.49+\delta}$

(b) $\mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mid \widetilde{\mathcal{B}}_\phi] \leq d^{-0.49+\delta}$

(c) for any $r \in [m]$, it holds that for $(p, v) \in \{(2, g), (5, y)\}$:

$$\langle \nabla_{\mathbf{W}_{5,j,r,p}} \mathsf{Loss}^{(t)}, e_v \rangle \geq (1 - \widetilde{O}(\frac{1}{d^{0.49-\delta}}))\mathbb{E}[\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbb{1}_{\mathcal{B}_\phi} \mathbb{1}_{F_{5,j} \leq B}]$$

Proof. We prove the statements separately.

- **Part (a)**: By Induction C.2, we have that All the irrelevant features $|\langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v \rangle| \leq \widetilde{O}(\sigma_0/d)$, so conditioned on $\mathcal{B}_{g,y}$ we have that

$$\begin{aligned} F_j(\mathbf{Z}) &\leq \sum_{r\in[m]} \mathbf{sReLU}(\Lambda_{5,j,r}^{(t)}(\mathbf{Z})) \leq \sum_{r\in[m]} \Psi_{j,r}^{(t)}(g, y) + \widetilde{O}(\sigma_0 m/d) \\ &\leq (0.5 + \delta)\log d + \widetilde{O}(\sigma_0) \\ &\leq 0.501 \log d \end{aligned}$$

  Then since all $F_j \geq 0$, we have that conditioned on $\mathcal{B}_{g,y}$, we have that

$$\mathbf{logit}_{5,j}^{(t)}(\mathbf{Z}) = \frac{e^{F_j(\mathbf{Z})}}{\sum_{j'\in[d]} e^{F_{j'}(\mathbf{Z})}} \leq \frac{e^{(0.51+\delta)\log d}}{d} \leq \frac{1}{d^{0.49-\delta}}$$

  which concludes the proof of (a). (b) can also be similarly proved.

- **Part (b)**: Firstly when Induction C.2 holds, for any $\phi = (g, y) \in \mathfrak{F}_j$, we have for all $\phi' = (g', y')$ such that exactly one of $g' = g$ or $y' = y$ is satisfied, it holds

$$\Psi_{j,r}(\phi') \le \Psi_{j,r}(\phi) + \widetilde{O}(\sigma_0)$$

Therefore by taking a sum over $r \in [m]$ it holds that

$$\sum_{r\in[m]} \Psi_{j,r}(\phi') \le \sum_{r\in[m]} \Psi_{j,r}(\phi) + \widetilde{O}(\sigma_0) \le (0.5+\delta)\log d + \widetilde{O}(\sigma_0) \le (0.5001+\delta)\log d$$

So by (a), we have the logit upper bound for all $\phi' = (g', y')$ such that it shares a component with ϕ as $\mathbb{E}[\mathbf{logit}_{5,j} \mid \mathcal{B}_{\phi'}] \le d^{-0.49+\delta}$. Note that since $\widetilde{\mathcal{B}}_\phi = (\bigcup_{(g',y),g'\neq g} \mathcal{B}_{g',y}) \cup (\bigcup_{(g,y'),y'\neq y} \mathcal{B}_{g,y'})$, we can obtain the desired result.

- **Part (c)**: By combining (a) and Induction C.3, we can compute

$$\begin{aligned}
&\left|\langle \nabla_{\mathbf{W}_{5,j,r,p}} \mathsf{Loss}^{(t)}, e_v\rangle - \mathbb{E}[(1-\mathbf{logit}_{5,j}^{(t)})\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{\mathcal{B}_\phi}\mathbb{1}_{F_{5,j}\le B}]\right| \\
&\le \mathbb{E}[\mathbf{logit}_{5,j}\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{\widetilde{\mathcal{B}}_\phi}\mathbb{1}_{F_{5,j}\le B}] \\
&\le \widetilde{O}(d^{-0.49+\delta})\mathbb{E}[(1-\mathbf{logit}_{5,j}^{(t)})\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{\mathcal{B}_\phi}\mathbb{1}_{F_{5,j}\le B}]
\end{aligned}$$

where the last inequality is because $\Lambda_{5,j,r}^{(t)}$ conditioned on $\widetilde{\mathcal{B}}_\phi$ is smaller than $\widetilde{O}(1)\mathbb{E}[\Lambda_{5,j,r}^{(t)} \mid \mathcal{B}_\phi]$ from Induction C.3, and the application of (a). Note that we ignore the event $\{F_{5,j} \le B\}$ because it always happens when $\sum_r \Psi_{j,r}(\phi) \le \log d$ (coupled with our induction about irrelavant features).

Now we have finished all proofs. □

Lemma C.2 (initialization gap between features)**.** *Assuming Assumption 4.1. Let $j \in \tau(\mathcal{Y})$, for all $r \in [m]$ and any two $\mathbf{u}, \mathbf{u}' \in \mathcal{U}$, we have with prob $\ge 1 - o(1)$ over the randomness at initialization that*

$$|\Psi_{\mathbf{u}}^{(0)} - \Psi_{\mathbf{u}'}^{(0)}| \gtrsim \frac{\sigma_0}{n_y^4 m^2 \log d}$$

Proof. We give a straightforward proof that every pair of $\Psi_{\mathbf{u}}$ has a gap of $\frac{\sigma_0}{n_y^4 m^2 \log d}$. First note that $\Psi_{\mathbf{u}}$ of different $\mathbf{u} \in \mathcal{U}$ 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 $\Psi_{\mathbf{u}}^{(0)} - \Psi_{\mathbf{u}'}^{(0)}$ 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. □

Lemma C.3 (consistency of gradients)**.** *Let $\phi = (g, y) \in \mathfrak{F}$, for any $j \in [d]$ and $r \in [m]$, we have with probability $1 - \frac{1}{d^{\Omega(\log d)}}$ over $\mathbf{W}_{5,j,r}^{(0)}$ that:*

$$\begin{aligned}
&\left|\mathbb{E}_{x_0,x_1}[\mathbf{sReLU}'(\Lambda_{5,j,r}^{(0)})\mathbb{1}_{\mathcal{B}_{g,y}} \mid \mathbf{W}^{(0)}] - \mathbb{E}_{x_0,x_1,\mathbf{W}^{(0)}}[\mathbf{sReLU}'(\Lambda_{5,j,r}^{(0)})\mathbb{1}_{\mathcal{B}_{g,y}}]\right| \\
&\le \begin{cases} \widetilde{O}(\frac{(\Psi_{j,r}^{(t)}(\phi))^{q-1}}{d}) & \text{if } \Psi_{j,r}^{(t)}(\phi) \le d^{-0.01} \\ \widetilde{O}(\frac{\Psi_{j,r}^{(t)}(\phi)}{d}) & \text{if } \Psi_{j,r}^{(t)}(\phi) > d^{-0.01} \end{cases}
\end{aligned}$$

Proof. We can view the expectation

$$\mathbb{E}_{(x_0,x_1)\sim\mathcal{X}^1}[\mathbf{sReLU}'(\Lambda_{5,j,r}^{(0)}) \mid g_0 = g, y_0 = y]$$

at iteration $t = 0$ as

$$\mathbb{E}_{(x_0,x_1)\sim\mathcal{X}^1}[\mathbf{sReLU}'(\Lambda_{5,j,r}^{(0)}) \mid g_0 = g, y_0 = y] = \frac{1}{\binom{|\mathcal{X}|}{2}} \sum_{x,x'\in\mathcal{X}} h(v_1, v_2)$$

where

$$h(v_1, v_2) = \mathbb{E}[\mathbf{sReLU}'(\Lambda_{5,j,r}^{(0)})\mathbb{1}_{\mathcal{B}_{g,y}} \mid x_0 = v_1, x_1 = v_2], \quad \text{for some } v_1, v_2 \in \mathcal{X}$$

Now we can apply Lemma E.1 by viewing the RHS as a U-statistics where the randomness of each $h(v_1, v_2)$ comes from the initialization of $\langle \mathbf{W}_{5,j,r,1}, e_{v_1} \rangle + \langle \mathbf{W}_{5,j,r,1}, e_{v_1} \rangle$ and $\langle \mathbf{W}_{5,j,r,1}, e_{v_2} \rangle$ which are identically distributed and jointly independent with any $h(v_1', v_2')$ if the sets $\{v_1, v_2\}$ is disjoint with the set $\{v_1', v_2'\}$. So by choosing $n = |\mathcal{X}| = \Theta(d)$ and $m = 2$, $M = (\widetilde{O}(\Psi_{j,r}^{(0)}(\phi))^{q-1} \log d)^{q-1}$ when $\Psi_{j,r}^{(0)}(\phi) \le d^{-0.01}$ and $M = \widetilde{O}(\Psi_{j,r}^{(0)}(\phi) \log d)^{q-1}$ when $\Psi_{j,r}^{(0)}(\phi) > d^{-0.01}$ and corresponding $t = nM \log d$ in Lemma E.1 we have the desired result. □

Fact C.3 (conditional expectation of logit)**.** Let $\phi = (g, y) \in \mathfrak{F}_j$, then we have the decompositions: For the negative gradient $\mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbb{1}_{\widetilde{\mathcal{B}}_\phi}]$, we have

$$\begin{aligned}\mathbb{E}\big[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mid \widetilde{\mathcal{B}}_\phi\big] &= \mathbf{Pr}(\widetilde{\mathcal{B}}_j(g) \mid \widetilde{\mathcal{B}}_\phi) \cdot \mathbb{E}\big[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mid \widetilde{\mathcal{B}}_j(g)\big] \\ &\quad + \mathbf{Pr}(\widetilde{\mathcal{B}}_j(y) \mid \widetilde{\mathcal{B}}_\phi) \cdot \mathbb{E}\big[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mid \widetilde{\mathcal{B}}_j(g)\big]\end{aligned}$$

Lemma C.4 (irrelavant features)**.** *Suppose Induction C.3 holds for all $t < T_1$, then Induction C.2a holds at iteration $t+1$. Moreover, let $\mathcal{A}_x = \{x \in \{x_0, x_1\}, (x_0, x_1) \in \mathcal{D}_{\mathcal{X}}^1\}$, then at each step the following holds:*

$$|\langle \mathbf{W}_{5,j,r,p}^{(t+1)}, e_x \rangle - \langle \mathbf{W}_{5,j,r,p}^{(t)}, e_x \rangle| \le \sum_{g \in \mathcal{G}} \eta \mathbf{Pr}(\mathcal{A}_x) \cdot |\langle \nabla_{\mathbf{W}_{5,j,r,2}} \mathsf{Loss}^{(t)}, e_g \rangle|$$

Proof. We shall be proving that the total feature growth of any $x \in \mathcal{X}$ at the end of training should be negligible, that is, $\langle \mathbf{W}_{5,j,r,p}^{(t)}, e_x \rangle - \langle \mathbf{W}_{5,j,r,p}^{(0)}, e_x \rangle \le \widetilde{O}(\frac{1}{d}) \ll \sigma_0$. In fact, suppose that something weaker happened before t, for example $\langle \mathbf{W}_{5,j,r,p}^{(t)}, e_x \rangle - \langle \mathbf{W}_{5,j,r,p}^{(0)}, e_x \rangle \le \sigma_0 / d^{0.1}$, then by the gradient formula for $\mathbf{W}_5$, for any $x \in \mathcal{X}$ and $p \in \{1, 3, 4\}$, we have the gradient of feature $\langle \mathbf{W}_{5,j,r,p}, e_x \rangle$ is:

$$\begin{aligned}&\langle \nabla_{\mathbf{W}_{5,j,r,p}} \mathsf{Loss}^{(t)}, e_x \rangle \\ &= \mathbb{E}_{x \in \{x_0, x_1\}}[\mathcal{E}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})] \\ &= \mathbb{E}_{x \in \{x_0, x_1\}}[(1 - \mathbf{logit}_{5,j}^{(t)}) \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbb{1}_{\mathcal{B}_j} - \mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbb{1}_{\widetilde{\mathcal{B}}_j}]\end{aligned}$$

By Induction C.3, we know that there is at most one feature $(g, y) \in \mathfrak{F}_j$ such that $\Psi_{j,r}^{(t)}(g, y) > d^{0.1} \sigma_0$, so we can decompose the gradient of e_x at any iteration $s \le t$ to:

$$\langle \nabla_{\mathbf{W}_{5,j,r,p}} \mathsf{Loss}^{(s)}, e_x \rangle \tag{23}$$

$$= \mathbb{E}[\mathbb{1}_{\mathcal{A}_x} (1 - \mathbf{logit}_{5,j}^{(s)}) \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j} \mathbb{1}_{F_{5,j} \le B} - \mathbf{logit}_{5,j}^{(s)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\widetilde{\mathcal{B}}_j} \mathbb{1}_{F_{5,j} \le B}] \tag{24}$$

$$\begin{aligned}&= \mathbb{E}[(1 - \mathbf{logit}_{5,j}^{(s)}) \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(g)} - \mathbf{logit}_{5,j}^{(s)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\widetilde{\mathcal{B}}_j(g)} \mathbb{1}_{F_{5,j} \le B} \mathbb{1}_{\mathcal{A}_x}] \\ &\quad + \mathbf{Pr}(\mathcal{A}_x) \sum\nolimits_{g' \in \mathcal{G}, g' \ne g} \mathbb{E}\left[(1 - \mathbf{logit}_{5,j}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(g')} \mathbb{1}_{F_{5,j} \le B} \mid \mathcal{A}_x\right] \\ &\quad + \mathbf{Pr}(\mathcal{A}_x) \sum\nolimits_{g' \in \mathcal{G}, g' \ne g} \mathbb{E}\left[\mathbf{logit}_{5,j}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbb{1}_{\widetilde{\mathcal{B}}_{g'}} \mathbb{1}_{F_{5,j} \le B} \mid \mathcal{A}_x\right]\end{aligned}$$

Now since $\langle \mathbf{W}_{5,j,r,p}^{(s)}, e_{x'} \rangle \le \sigma_0 \log d, \forall x' \in \mathcal{X}$, it holds that the sum of all updates of $\langle \mathbf{W}_{5,j,r,p}, e_x \rangle$ before t is bounded by

$$\begin{aligned}&\sum_{s \le t} O(\frac{\eta}{d}) \mathbb{E}[(1 - \mathbf{logit}_{5,j}^{(s)}) \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(g,y)} \mathbb{1}_{F_{5,j} \le B} - \mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\widetilde{\mathcal{B}}_j(g)} \mathbb{1}_{F_{5,j} \le B} \mid \mathcal{A}_x] \\ &\quad + \sum_{s \le t} \sum\nolimits_{g' \ne g, g' \in \mathcal{G}} O(\frac{\eta}{d}) \mathbb{E}\left[(1 - \mathbf{logit}_{5,j}^{(s)}(F)) \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(g')} \mathbb{1}_{F_{5,j} \le B} \mid \mathcal{A}_x\right] \\ &\quad + \sum_{s \le t} \sum\nolimits_{g' \ne g, g' \in \mathcal{G}} O(\frac{\eta}{d}) \mathbb{E}\left[\mathbf{logit}_{5,j}^{(s)} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\widetilde{\mathcal{B}}_j(g')} \mathbb{1}_{F_{5,j} \le B} \mid \mathcal{A}_x\right]\end{aligned}$$

$=: I_1 + I_2 + I_3$

Since by Induction C.3, we can bound $I_2 + I_3$ by

$$|I_2 + I_3| \le \sum\nolimits_{g' \neq g, g' \in \mathcal{G}} O(\frac{1}{d}) \left| \sum\nolimits_{s \le t} \eta \langle \nabla_{\mathbf{W}_{5,j,r,2}} \mathsf{Loss}^{(s)}, e_{g'} \rangle \right| \\ \lesssim \frac{n}{d} \times \widetilde{O}(\sigma_0) \ll \sigma_0 / d^{0.99}$$

and for I_1, we also have

$$|I_1| \le \sum_{s \le t} O(\frac{1}{d}) \eta \mathbf{Pr}(\mathcal{A}_x) \cdot \langle \nabla_{\mathbf{W}_{5,j,r,2}} \mathsf{Loss}^{(s)}, e_g \rangle \ll \widetilde{O}(\frac{1}{d})$$

Thus the total update to $\langle \mathbf{W}_{5,j,r,p}, e_x \rangle$ for all iterations before t is bounded by $\widetilde{O}(\frac{1}{d})$ which proves the induction hypothesis and also the final desired result. The second statement can be obtained by looking at (24). □

C.3 Phase I: Initial Growth

First suppose $\mathbf{u} = (j, r, (g, y))$, where j denotes the token index and r is the neuron index. In phase I.1 which spans the time period $t \in [0, T_{\mathbf{u},0}]$, we show that the feature $\mathbf{u}$ remains close to initial value while competing with other features. In phase I.2, we show that the feature $\mathbf{u}$ grows faster than other features and reach a certain magnitude.

C.3.1 Phase I.a: Emergence of the Feature

Before the feature $\mathbf{u}$ starts to grow, we know that the gradient for it could change, due to the confounding feature which affects the logits of the gradient. When the feature $\Psi_{\widetilde{\mathbf{u}}}$ grows, the erroneous prediction probability $\mathbf{logit}_{5,\widetilde{j}}$ could rise and therefore decrease the gradient of $\mathbf{u}$. We shall show below that this two effects will not affect the growth of $\mathbf{u}$ much, and thus we can safely ignore them.

In fact, we show the following bound on the "optimistic growth" by pseudo weights in Definition C.5 almost match the actual growth by the true weights, in phase I.a.

$$\left| \langle \widetilde{\mathbf{W}}^{(t)}_{5,j,r,p}, e_v \rangle - \langle \mathbf{W}^{(t)}_{5,j,r,p}, e_v \rangle \right| \le \widetilde{O}(\sigma_0 / d^{0.1}) \tag{25}$$

To show this, we need to bound the trajectory of $|\widetilde{\mathbf{W}}^{(t)}_{5,j,r,p} - \mathbf{W}^{(t)}_{5,j,r,p}|$ during the time peirod $t \in [0, T_{\mathbf{u}',2}]$, which is the point where the immediate predecessor has been learned almost optimally. We need to argue the two following conditions are satisfied:

- $\mathbf{C}_1$: The total amount of iterations where $\mathbb{E}[\mathbf{logit}^{(t)}_{5,j} \mid \mathcal{B}_{g,y}] \ge \frac{1}{d^{0.1}}$ is smaller than $O(\frac{d^{0.2}}{\eta})$ (here the $d^{0.2}$ can be losen to almost $1/\sigma_0^{q-2}$).
- $\mathbf{C}_2$: Throughout $t \in [0, T_{\mathbf{u}',2}]$, we have $\langle \widetilde{\mathbf{W}}^{(t)}_{5,j,r,p}, e_v \rangle \le \widetilde{O}(\sigma_0)$.

So we have the following proposition for $\Psi_{\mathbf{u}}$'s feature growth during the feature learning process of predecessor features.

Proposition C.1 (Phase I.a)**.** Let $\mathbf{u}'$ be the immediate predecessor of $\mathbf{u}$ in $\mathcal{U}^\star$, then Induction C.1 is satisfied for all iterations $t \le T_{\mathbf{u}',2}$. Moreover, (25) holds at $t = T_{\mathbf{u}',2}$.

To prove this proposition, we first prove the following lemma.

Lemma C.5 (gradient approximation)**.** *Let $j \in [d]$ and $r \in [m]$. Suppose at iteration $t \le \mathcal{T}_1$ there is a $\beta \in [0, O(\sigma_0)]$ such that $|\langle \mathbf{W}^{(t)}_{5,j,r,q} - \mathbf{W}^{(0)}_{5,j,r,q}, e_x \rangle| \le \beta$ for all $q \in \{1, 3, 4\}$ and all $x \in \mathcal{X}$, then if $\Lambda^{(t)}_{5,j,r} \ge 0$, we shall have*

$$\left| \mathbf{sReLU}'(\Lambda^{(t)}_{5,j,r}(\mathbf{Z})) - \mathbf{sReLU}'(\widetilde{\Lambda}^{(t)}_{5,j,r}(\mathbf{Z})) \right| \le \begin{cases} \beta, & \text{if } \Lambda^{(t)}_{5,j,r}(\mathbf{Z}) \ge \frac{1}{2}\varrho, \\ O(\beta(\Lambda^{(t)}_{5,j,r}(\mathbf{Z}) + \beta)^{q-2}), & \text{if } \Lambda^{(t)}_{5,j,r}(\mathbf{Z}) \le \frac{1}{2}\varrho; \end{cases}$$

Proof. Note that by Induction C.2(b) we have that

$$|\langle \mathbf{W}_{5,j,r,p}^{(t)}, e_x\rangle - \langle \mathbf{W}_{5,j,r,p}^{(0)}, e_x\rangle| \le \widetilde{O}(1/d), \qquad \forall p \in \{1,3,4\} \text{ and } x \in \mathcal{X}$$

which allows us to bound the difference between $\Lambda_{5,j,r}^{(t)}$ and $\widetilde{\Lambda}_{5,j,r}^{(t)}$ as any iteration t:

$$|\Lambda_{5,j,r}^{(t)}(\mathbf{Z}) - \widetilde{\Lambda}_{5,j,r}^{(t)}(\mathbf{Z})| \le \frac{1}{3}\sum_{p\in\{1,3,4\}} \big|\langle \mathbf{W}_{5,j,r,p}^{(t)} - \mathbf{W}_{5,j,r,p}^{(0)}, e_x\rangle\big| \le \widetilde{O}(1/d), \quad \forall \mathbf{Z} \in \mathrm{supp}(\mathcal{D}_{\mathcal{X}}^1)$$

Now we are able to bound the difference between $\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})$ and $\mathbf{sReLU}'(\widetilde{\Lambda}_{5,j,r}^{(t)})$ as follows:

- When $\Lambda_{5,j,r}^{(t)}(\mathbf{Z}) \ge \frac{1}{2}\varrho$, because of the monotonically increasing slope of $\mathbf{sReLU}'(x)$ for $x \in [0, \varrho]$ and flat slope when $x \ge \varrho$, we have that $|\mathbf{sReLU}'(x) - \mathbf{sReLU}'(x+\epsilon)| \le \epsilon$ for $\epsilon > 0$. Therefore
$$\Big|\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}(\mathbf{Z})) - \mathbf{sReLU}'(\widetilde{\Lambda}_{5,j,r}^{(t)}(\mathbf{Z}))\Big| \le O(|\Lambda_{5,j,r}^{(t)}(\mathbf{Z}) - \widetilde{\Lambda}_{5,j,r}^{(t)}(\mathbf{Z})|) \le \beta$$
- When $\Lambda_{5,j,r}^{(t)}(\mathbf{Z}) \le \frac{1}{2}\varrho$, we have that$|\mathbf{sReLU}'(x) - \mathbf{sReLU}'(x+\epsilon)| \le O(q^2\epsilon\cdot(|x|+\epsilon)^{q-2})$ for $x \le \frac{1}{2}\varrho$ and $\epsilon \ll \varrho$, therefore
$$\begin{aligned}\Big|\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}(\mathbf{Z})) - \mathbf{sReLU}'(\widetilde{\Lambda}_{5,j,r}^{(t)}(\mathbf{Z}))\Big| &\le O(|(\Lambda_{5,j,r}^{(t)}(\mathbf{Z}))^{q-1} - (\widetilde{\Lambda}_{5,j,r}^{(t)}(\mathbf{Z}))^{q-1}|)\\ &\le O\Big(\beta\cdot(\Lambda_{5,j,r}^{(t)}(\mathbf{Z}) + \beta)^{q-2}\Big)\end{aligned}$$

This concludes the proof. □

proof of Proposition C.1. Actually, for all $\widetilde{\mathbf{u}} \in \Phi$ such that $\widetilde{\mathbf{u}} \prec \mathbf{u}$ and $\widetilde{\mathbf{u}}_3$ share at least one feature component g or y with $\mathbf{u}_3$, we know that there time duration of $t \in [\mathcal{T}_{\widetilde{\mathbf{u}},1}, \mathcal{T}_{\widetilde{\mathbf{u}},2}]$ is at most $\mathcal{T}_{\widetilde{\mathbf{u}},2} - \mathcal{T}_{\widetilde{\mathbf{u}},1} \le \widetilde{O}(\frac{d^{0.1}}{\eta})$. Therefore the total number of iterations where $\mathbb{E}[\mathbf{logit}_{5,j} \mid \mathcal{B}_{g,y}] \ge \frac{1}{d^{0.1}}$ before $t = T_{\mathbf{u}',2}$ is at most $nO(\frac{d^{0.1}}{\eta}\log^2 d) \le O(\frac{d^{0.1}}{\eta}\log^4 d)$. Then for each step t, we have

- When $\mathbb{E}[\mathbf{logit}_{5,j} \mid \mathcal{B}_{g,y}] \ge \frac{1}{d^{0.1}}$, we have that
$$\begin{aligned}&\big|\langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t+1)}, e_v\rangle - \langle \mathbf{W}_{5,j,r,p}^{(t+1)}, e_v\rangle\big|\\ \le\ &\big|\langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t)}, e_v\rangle - \langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v\rangle\big| + \Big|\eta\Big(\mathbb{E}[\mathbf{sReLU}'(\widetilde{\Lambda}_{5,j,r}^{(t)})] - \mathbb{E}[\mathcal{E}_{5,j}^{(t)}\mathbf{sReLU}'(\widetilde{\Lambda}_{5,j,r}^{(t)})]\Big)\Big|\\ \le\ &\big|\langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t)}, e_v\rangle - \langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v\rangle\big|\\ &\quad + \eta\Big|\mathbb{E}\Big[\mathbf{logit}_{5,j}^{(t)}\mathbb{1}_{\mathcal{B}_{g,y}}\Big(\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) + \mathbf{sReLU}'(\widetilde{\Lambda}_{5,j,r}^{(t)})\Big)\Big]\Big|\\ &\quad + \eta\Big|\mathbb{E}\Big[\mathbf{logit}_{5,j}^{(t)}\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{\widetilde{\mathcal{B}}_{g,y}}\Big]\Big|\\ \le\ &\big|\langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t)}, e_v\rangle - \langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v\rangle\big| + \widetilde{O}(\eta\sigma_0) + \Big|\eta\mathbb{E}\Big[\mathbf{logit}_{5,j}^{(t)}\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{\widetilde{\mathcal{B}}_{g,y}}\Big]\Big|\\ \le\ &\big|\langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t)}, e_v\rangle - \langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v\rangle\big| + \widetilde{O}(\eta\sigma_0^{q-1})\end{aligned}$$
where the last inequality is because both $\mathbf{logit}_{5,j}^{(t)} \le O(1)$ and $\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) = \widetilde{O}(\sigma_0^{q-1})$ is much smaller conditioned on $\widetilde{\mathcal{B}}_{g,y}$. In fact, by Induction C.3 at each single iteration $t \le T_{\mathbf{u}',2}$ there can only be one $\widetilde{\mathbf{u}} = (\widetilde{j}, \widetilde{r}, \widetilde{\phi}) \in \mathcal{U}^\star$ such that $\mathbf{logit}_{5,j}^{(t)} \ge \frac{1}{d^{0.1}}$ conditioned on $(g_1, y_0) = \widetilde{\phi}$. Therefore we achieve the same bound.
- When $\mathbb{E}[\mathbf{logit}_{5,j} \mid \mathcal{B}_{g,y}] \le \frac{1}{d^{0.1}}$, by similar calculations, we have
$$\begin{aligned}&\big|\langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t+1)}, e_v\rangle - \langle \mathbf{W}_{5,j,r,p}^{(t+1)}, e_v\rangle\big|\\ \le\ &\Big|\langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t)}, e_v\rangle - \langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v\rangle\Big|\end{aligned}$$

$$
\begin{aligned}
&+\Big|\eta\mathbb{E}\Big[\mathbf{logit}_{5,j}^{(t)}\mathbb{1}_{\mathcal{B}_{g,y}}\Big(\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})+\mathbf{sReLU}'(\widetilde{\Lambda}_{5,j,r}^{(t)})\Big)\Big]\Big|\\
&+\Big|\eta\mathbb{E}\Big[\mathbf{logit}_{5,j}^{(t)}\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{\widetilde{\mathcal{B}}_{g,y}}\Big]\Big|\\
\leq&\big|\langle\widetilde{\mathbf{W}}_{5,j,r,p}^{(t)},e_v\rangle-\langle\mathbf{W}_{5,j,r,p}^{(t)},e_v\rangle\big|+\widetilde{O}(\frac{1}{d^{0.1}}\eta\sigma_0^{q-1})
\end{aligned}
$$

Since we know that $T_{\mathbf{u}',2}\leq\widetilde{O}(\frac{1}{\eta\sigma_0^{q-2}})$, we have that the total growth of difference between $\langle\widetilde{\mathbf{W}}_{5,j,r,p}^{(t)},e_v\rangle$ and $\langle\mathbf{W}_{5,j,r,p}^{(t)},e_v\rangle$ is bounded by

$$
\begin{aligned}
\big|\langle\widetilde{\mathbf{W}}_{5,j,r,p}^{(T_{\mathbf{u}',2})},e_v\rangle-\langle\mathbf{W}_{5,j,r,p}^{(T_{\mathbf{u}',2})},e_v\rangle\big|&\leq\widetilde{O}(\frac{d^{0.1}}{\eta})\cdot\widetilde{O}(\eta\sigma_0^{q-1})+\widetilde{O}(\frac{1}{\eta\sigma_0^{q-2}})\cdot\widetilde{O}(\frac{1}{d^{0.1}}\eta\sigma_0^{q-1})\\
&\leq\widetilde{O}(\frac{\sigma_0}{d^{0.1}})
\end{aligned}
$$

Therefore, we have shown that the (25) holds for all $t\in[0,T_{\mathbf{u}',2}]$. This provides a good initialization for future leanring. □

Now we are set to prove the proposition that concludes Phase I.a.

Proposition C.2 (Phase I.a, feature competition)**.** Let $\mathbf{u}'\prec\mathbf{u}\in\mathcal{U}^\star$, then Induction C.3 holds for all iterations $t\in[0,T_{\mathbf{u},1a}]$. Moreover, we have the following properties:

$$
\Psi_{\mathbf{u}}^{(T_{\mathbf{u},1a})}\geq d^{0.01}\sigma_0,\quad\text{while }\Psi_{\widetilde{\mathbf{u}}}^{(T_{\mathbf{u},1a})}=\widetilde{O}(\sigma_0),\quad\forall\widetilde{\mathbf{u}}\succ\mathbf{u}\in\mathcal{U}
$$

Proof. By Proposition C.1, we can deduce

$$
\begin{aligned}
\Psi_{\mathbf{u}}^{(T_{\mathbf{u}',2})}&\geq\Psi_{\mathbf{u}}^{(0)}-\widetilde{O}(\sigma_0/d^{0.1})\\
&\geq\Psi_{\widetilde{\mathbf{u}}}^{(0)}+\frac{1}{\log^{O(1)}d}-\widetilde{O}(\sigma_0/d^{0.1})\\
&\geq\Psi_{\widetilde{\mathbf{u}}}^{(T_{\mathbf{u}',2})}+\frac{1}{\log^{O(1)}d}-\widetilde{O}(\sigma_0/d^{0.1})
\end{aligned}
$$

Now we proceed to prove the result for $t\in[T_{\mathbf{u}',2},T_{\mathbf{u},1a}]$. Remember $\mathbf{u}'$ is the immediate predecessor of $\mathbf{u}=(j,r,(g,y))$ in $\mathcal{U}^\star$. At this point, by Lemma C.1, we have a basic gradient lower bound for both g and y in neuron r:

$$
\langle\nabla_{\mathbf{W}_{5,j,r,p}}\mathsf{Loss}^{(t)},e_v\rangle\geq\frac{1}{n^2}(1-O(n/d))\mathbb{E}[\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mid\mathcal{B}_{g,y}]\quad\text{for }(p,v)\in\{(2,g),(5,y)\}\tag{26}
$$

Using these lower bound, we can prove that the feature $\Psi_{\mathbf{u}}^{(t)}$ will outgrow all other feature $\Psi_{\widehat{\mathbf{u}}}^{(t)}$ where $\mathbf{u}\prec\widehat{\mathbf{u}}\in\mathcal{U}^\star$ with $\widehat{\mathbf{u}}_1=j$. In fact, let $\widehat{\mathbf{u}}=(j,\widehat{r},(\widehat{g},\widehat{y}))$ be an index from $\mathcal{U}$ where $(\widehat{g},\widehat{y})\in\mathfrak{F}_j$, Now assume the following induction hypotheses ① and ② during $t\in[T_{\mathbf{u}',2},T_{\mathbf{u},1a}]$:

① $\Psi_{\mathbf{u}}^{(t)}\geq\Psi_{\widehat{\mathbf{u}}}^{(t)}+\Omega(\frac{\sigma_0}{\log^{\Omega(1)}d})\leq\Psi_{\mathbf{u}}^{(t)}$ for all $\widehat{\mathbf{u}}\in\Sigma$ where $\widehat{\mathbf{u}}_1=j$.

② The condition (26) is satisfied for all $\widehat{\mathbf{u}}\in\Sigma$ where $\widehat{\mathbf{u}}_1=j$.

We shall prove that if both ① and ② are satisfied at $t=T_{\mathbf{u}',2}$, then they are also satisfied at each $t\in(T_{\mathbf{u}',2},T_{\mathbf{u},1a}]$. Let us assume they are satisfied at some iteration $t\in(T_{\mathbf{u}',2},T_{\mathbf{u},1a}]$, we show below it remains the case at $t+1$.

- **Proof of ② at $t+1$.** Since by the definition of $T_{\mathbf{u},1a}$, we have Lemma C.1 holds for feature $\mathbf{u}$ for all iteration $t\leq T_{\mathbf{u},1a}$. Moreover, by our induction hypoethsis, ① holds for all iteration s where $T_{\mathbf{u}',2}\leq s\leq t$, i.e. $\Psi_{\widehat{\mathbf{u}}}^{(s)}\leq\Psi_{\mathbf{u}}^{(s)}$ for all $s\in[T_{\mathbf{u}',2},t]$. Therefore we can obtain that for all $s\in[T_{\mathbf{u}',2},t]$, the gradient of $\Psi_{\widehat{\mathbf{u}}}$ satisfies

  $$
  \langle\nabla_{\mathbf{W}_{5,j,r',p}}\mathsf{Loss}^{(s)},e_v\rangle\leq\frac{1}{n^2}\mathbb{E}[\mathbf{sReLU}'(\Lambda_{5,j,\widehat{r}}^{(s)})\mid\mathcal{B}_{\widehat{g},\widehat{y}}]\text{ for }(p,v)\in\{(2,\widehat{g}),(5,\widehat{y})\}\tag{27}
  $$

  By applying both (26) and (27), couple with Induction C.3, we can show that the gradient of $\Psi_{\mathbf{u}}$ is always larger than that of $\Psi_{\widehat{\mathbf{u}}}$ as long as which proves ① for $t+1$.

- **Proof of ① at $t+1$** by ② we have $\Psi_{\widehat{\mathbf{u}}}^{(t+1)} \le \Psi_{\mathbf{u}}^{(t+1)}$, which, combined with Induction C.3, implies that

$$\sum_{r\in[m]} \Psi_{j,r}^{(t)}(\widehat{g},\widehat{y}) \le \widetilde{O}(\sigma_0 m) \le O(d^{0.1}\sigma_0)$$

then by Lemma C.1, we have the negative gradient for $\Psi_{\mathbf{u}}^{(t+1)}$ satisfies (26), which concludes the induction step.

Moreover, from Lemma E.2, we obtain that at $t = cT_{\mathbf{u},1a}$, we have the following result: $\Psi_{\mathbf{u}}^{(t)} \ge d^{0.01}\sigma_0$ while all $\Psi_{\widehat{\mathbf{u}}}^{(t)} \le \widetilde{O}(\sigma_0)$ for $\widehat{\mathbf{u}} \succ \mathbf{u}$ in Σ; So the proof of Phase Ia is complete. □

C.3.2 Phase I.b: Feature Growth

In this stage we prove the following result:

Proposition C.3 (Phase I, cyclic group)**.** Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$, then Induction C.3 holds for all $t \le T_{\mathbf{u},1}$, and we have the following results at $t = T_{\mathbf{u},1}$:

(A) $\Psi_{\mathbf{u}}^{(T_{\mathbf{u},1})} \ge \Omega(\log d)$

(B) For any $\widetilde{\mathbf{u}} \in \mathcal{U}$ such that $\widetilde{\mathbf{u}} \succ \mathbf{u}$, we have $\Psi_{\widetilde{\mathbf{u}}}^{(T_{\mathbf{u},1})} \le \widetilde{O}(\sigma_0)$.

Proof. The same result for phase I.a is proven in Proposition C.2. We only need to prove for the iterations $t \in [T_{\mathbf{u},1a}, T_{\mathbf{u},1}]$. Indeed, The total number of iterations in this stage is at most $\widetilde{O}(\frac{1}{\eta d^{0.01}\sigma_0^{q-2}})$. By Lemma C.1 and Lemma C.1, we know that as long as $\Psi_{\mathbf{u}}^{(t)} \le 0.5\log d$, it holds that

$$\nabla\psi_{j,r}^{(t)}(v) \ge \frac{1}{n_y^2}(1 - O(\frac{n_y}{d^{0.49}}))\mathbb{E}[\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mid \mathcal{B}_{g,y}], \quad v \in \{g, y\} \tag{28}$$

This ensured that we can use Corollary E.1 for $\widetilde{O}(\frac{1}{\eta d^{0.01}\sigma_0^{q-2}})$ many iterations starting from $T_{\mathbf{u},1a}$ to reach $T_{\mathbf{u},1}$, where $\Psi_{\mathbf{u}}^{(t)} \ge \varrho/2$. Moreover, since $\mathbf{sReLU}'(x)$ is a smooth polynomial for $x \in [0, \varrho]$ and a constant for $x \ge \varrho$. The same lower bound (28) can be applied and used to calculate the iterations needed for $\Psi_{\mathbf{u}}^{(t)}$ to reach $\Omega(\log d)$ after reaching $\varrho/2$, which is $O(\log d/\eta)$, as long as $\Psi_{\mathbf{u}}^{(t)} \le 0.5\log d$. □

C.4 Phase II: Cancellation and Convergence

In this stage, we shall show that the incorrect feature combination will move close to zero when the FFN weights reach a certain level of convergence.

Lemma C.6 (activeness of a neuron)**.** *Let $\mathbf{u} = (j, r, (g, y)) \in \mathcal{U}^\star$. If Induction C.3 holds, then for all $t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$, it holds that $\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) = 1$ conditioned on $\mathcal{B}_{g,y}$.*

Proof. It is satisfied at $t = T_{\mathbf{u},1}$ by Induction C.3. For any $t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$, we have that whenever $\Psi_{\mathbf{u}}^{(t)}$ falls slightly below $\frac{1}{2}\log d$ (which have to happen before it reaches even lower), by Lemma C.1 and Lemma C.1 we have that

$$\nabla\psi_{j,r}(v) \ge \frac{1}{n^2}(1 - O(\frac{n}{d^{0.49}}))\mathbb{E}[\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mid \mathcal{B}_{g,y}] \ge \frac{1}{n^2}(1 - O(\frac{n}{d^{0.49}})) \quad \text{(for } v = y \text{ or } g\text{)}$$

Therefore $\Psi_{\mathbf{u}}^{(t)}$ is increasing once it suppass a certain threshold, and that $\Psi_{\mathbf{u}}^{(t)} \ge 0.49\log d$ for all $t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$ and therefore the neuron r is active. □

Lemma C.7 (bounds on same feature in different neurons)**.** *Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}$, where $\phi = (g, y)$, and $\mathbf{u}' = (j, r', \phi) \in \mathcal{U}$ for some $r' \ne r$ such that $\mathbf{u}' \succ \mathbf{u}$, we have that $\Psi_{\mathbf{u}'}^{(t)} \le \widetilde{O}(\sigma_0)$ for all $t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$.*

Proof. We shall prove this by induction. It is true at $t = T_{\mathbf{u},1}$ by Proposition C.3. Since $T_{\mathbf{u},2} - T_{\mathbf{u},1} = O(\frac{1}{\eta d^{0.001}\sigma_0^{q-2}})$, we can simply bound the total growth of $\Psi_{\mathbf{u}'}^{(t)}$ as follows: let $v = y$ or g, we have that

$$\begin{aligned}
\psi_{j,r'}^{(t)}(v) &\le \widetilde{O}(\sigma_0) + \sum_{s\in[T_{\mathbf{u},1},t]} \nabla_{\psi_{j,r'}^{(t)}(v)} \mathsf{Loss}^{(s)} \\
&\le \psi_{j,r'}^{(T_{\mathbf{u},1})} + O(\frac{1}{d^{0.001}\eta\sigma_0^{q-2}}) \cdot \max_{s\in[T_{\mathbf{u},1},t]} \eta\mathbb{E}[\mathbf{sReLU}'(\Lambda_{5,j,r'}^{(s)})] \quad \text{(using the upper bound)} \\
&\le \psi_{j,r'}^{(T_{\mathbf{u},1})} + O(\frac{1}{d^{0.001}\eta\sigma_0^{q-2}}) \cdot \widetilde{O}(\sigma_0^{q-1}) \\
&\le \psi_{j,r'}^{(T_{\mathbf{u},1})} + O(\sigma_0/d^{0.001})
\end{aligned}$$

where the second last inequality is due to that the activation for $\Lambda_{5,j,r'}^{(t)}$ is at most σ_0^{q-1} by Lemma C.6. Therefore we have that $\psi_{j,r'}^{(t)} \le \widetilde{O}(\sigma_0)$ for all $t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$. This concludes the induction. □

Lemma C.8 (logits in phase II)**.** *Consider a $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$, let $\mathfrak{F}^{(t)}$ contains all the feature combinations from $\mathfrak{F}$ that are learned at iteration t, that is, for each $\phi' \in \mathfrak{F}^{(t)}$, there exists a $\mathbf{u}' \in \mathcal{U}^\star$ such that $\mathbf{u}' \preceq \mathbf{u}$, which also means $t \ge T_{\mathbf{u}',2}$. Then for any $\phi' \in \mathfrak{F}^{(t)} \cap \mathfrak{F}_{\mathsf{conf}}(\phi)$, we have that for all $t \ge T_{\mathbf{u},2}$,*

$$\mathbf{logit}_{5,j}^{(t)} \mathbb{1}_{\mathcal{B}_{\phi'}} = \Theta(\exp(\Psi_{j,r}^{(t)}(\phi') - B)) = \Theta(\exp(\Psi_{j,r}^{(t)}(\phi')\boldsymbol{\lambda}/d))$$

otherwise, we have that $\mathbf{logit}_{5,j}^{(t)} \mathbb{1}_{\mathcal{B}_{\phi'}} = \Theta(\exp(\Psi_{j,r}^{(t)}(\phi')/d))$.

Proof. It is direct to verify the above claim following Induction C.3(b) since there is only one neuron that learned ϕ' and it would be the only active neuron besides the current learning neuron. □

The above lemma also classified the different $\mathbf{logit}_{5,j}^{(t)}$ into two categories: one is the suppressed logits which are small and proportional to $\boldsymbol{\lambda}/d$, and the other is the unsuppressed logits which are proportional to $\frac{1}{d}$. This leads to the following lemma:

Lemma C.9 (different cancellation conditions)**.** *Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$, let $\mathfrak{F}^{(t)}$ contains all the feature combinations from $\mathfrak{F}$ that are learned at iteration t as in Lemma C.8.*

- *If $\phi' \in \mathfrak{F}_j \cap \mathfrak{F}^{(t)} \cap \mathfrak{F}_{\mathsf{conf}}(\phi)$, we have that before $\Psi_{\mathbf{u}}^{(t)}$ exceeds $B/2$, $\Psi_{j,r}^{(t)}(\phi') \ge \Omega(\log d)$.*
- *Otherwise, we have that $\Psi_{j,r}^{(t)}(\phi') \le \frac{1}{d^{\Omega(1)}}$ once $\Psi_{\mathbf{u}}^{(t)} \ge 2.01 \log d$.*

The above lemma showed how different logits interact with the feature cancellation condition. Now we are ready to prove the following lemma:

Lemma C.10 (convergence of positive gradient)**.** *Let $j \in \tau(\mathcal{Y})$ and $\phi = (g, y) \in \mathfrak{F}_j$. For any level $\delta > d\sigma_0^{q-2}$, the total number of iterations of which the condition $\mathcal{K}_{\mathsf{pos}}(\phi, \delta)$ does not hold must be smaller than $O(\frac{n_y^3 \log^2 d}{\eta\delta})$.*

Proof. This is due to the upper bound assumption we put on $F_{5,j}$. In fact, suppose conversely that the pair $(g, y) \in \mathcal{F}_j$ satisfies

$$\mathbb{E}[(1 - \mathbf{logit}_{5,j}^{(t)})\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{F_{5,j}\le B} \mid \mathcal{B}_{g,y}] \ge \delta$$

for more than $O(\frac{n_y^3 \log^2 d}{\delta})$ many iterations. We compute the update by

$$\begin{aligned}
\psi_{j,r}^{(t+1)}(y) &= \psi_{j,r}^{(t)}(y) + \eta\langle \nabla_{\mathbf{W}_{5,j,r,5}} \mathsf{Loss}^{(t)}, e_y \rangle \\
&= \psi_{j,r}^{(t)}(y) + \eta\mathbb{E}\left[(1 - \mathbf{logit}_{5,j}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{F_{5,j}\le B}\mathbb{1}_{\mathcal{B}_j(y)}\right]
\end{aligned}$$

$$- \eta\mathbb{E}\left[\mathbf{logit}_{5,j}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{\widetilde{\mathcal{B}}_j(y)}\mathbb{1}_{F_{5,j}\le B}\right]$$

Let $g' \in \mathcal{G} \setminus \{g\}$ be such that $(g', y) \notin \mathfrak{F}_j$, we have that

$$\begin{aligned}&\psi_{j,r}^{(t+1)}(y) - \sum_{g'\in\mathcal{G}\setminus\{g\}} \psi_{j,r}^{(t+1)}(g')\\ &= \psi_{j,r}^{(t)}(y) - \sum_{g'\in\mathcal{G}\setminus\{g\}} \psi_{j,r}^{(t)}(g') + \eta\mathbb{E}\left[(1-\mathbf{logit}_{5,j}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{F_{5,j}\le B}\mathbb{1}_{\mathcal{B}_j(y)}\right]\end{aligned}$$

By a telescoping sum, we have that

$$\psi_{j,r}^{(t)}(y) - \sum_{g'\in\mathcal{G}\setminus\{g\}} \psi_{j,r}^{(t)}(g') = \sum_{s\le t}\eta\mathbb{E}\left[(1-\mathbf{logit}_{5,j}^{(s)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)})\mathbb{1}_{F_{5,j}\le B}\mathbb{1}_{\mathcal{B}_j(y)}\right]$$

Due to the contradiction assumption, we know that for $T = \omega(\frac{n_y^3 \log d}{\delta})$ many iterations, we have that

$$\psi_{j,r}^{(t)}(y) - \sum_{g'\in\mathcal{G}\setminus\{g\}} \psi_{j,r}^{(t)}(g') \ge B + \Omega(n_y \log^2 d)$$

which is impossible because $\psi_{j,r}^{(t)}(y)$ and $\psi_{j,r}^{(t)}(g)$ is both absolutely bounded by $B + O(1)$. Therefore there is a contradiction, $\mathcal{K}_{\mathsf{pos}}(\phi, \delta)$ does not hold for at most $O(\frac{n_y^3 \log^2 d}{\delta})$ many iterations. This concludes the proof. □

The above proof also produced a corollary.

Corollary C.1 (monotonicity of cancellations)**.** *Let $j \in \tau(\mathcal{Y})$ and $\phi = (g, y) \in \mathfrak{F}_j$. The following quantity is non-decreasing:*

$$\psi_{j,r}^{(t)}(y) - \sum_{g'\in\mathcal{G}\setminus\{g\}} \psi_{j,r}^{(t)}(g'), \quad \textit{and} \quad \psi_{j,r'}^{(t)}(g) - \sum_{y'\in\mathcal{Y}\setminus\{y\}} \psi_{j,r'}^{(t)}(y')$$

for all $t \ge T_{\mathbf{u},1}$.

Using the above corollary, we can now prove the following lemma:

Lemma C.11 (convergence of negative gradient)**.** *Let $j \in \tau(\mathcal{Y})$ and $\phi = (g, y) \in \mathfrak{F}_j$. We have that*

(a) For all iterations $t \ge T_{\mathbf{u},1}$, it holds that $\Lambda_{5,j,r}^{(t)}(\mathbf{Z})\mathbb{1}_{\mathcal{B}_{\widetilde{\phi}}} \in (-\frac{\varrho}{2}, \frac{\varrho}{2})$ for any $\widetilde{\phi} \in \mathfrak{F}_{\mathsf{conf}}(\phi)$.

(b) For any level $\delta \in (\frac{\boldsymbol{\lambda}}{d^{1.1}}, (d^{0.01}\sigma_0)^{q-2})$, there exists an iteration $t \ge T_{\mathbf{u},1} + O(\frac{n_y^3 \log^2 d}{\delta})$ such that $\mathcal{K}_{\mathsf{neg}}(\phi, \delta)$ holds.

Proof. We prove this by contradition. By Corollary C.1, we know that the difference $\psi_{j,r}^{(t)}(g) - \sum_{y'\in\mathcal{Y}\setminus\{y\}} \psi_{j,r}^{(t)}(y')$ is non-decreasing with a growth speed

$$\eta\mathbb{E}\left[(1-\mathbf{logit}_{5,j}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{F_{5,j}\le B}\mathbb{1}_{\mathcal{B}_j(y)}\right]$$

Since by Lemma C.10, we know that the iterations of which the condition $\mathcal{K}_{\mathsf{pos}}(\phi, \delta)$ does not hold must be smaller than $O(\frac{n_y^3 \log^2 d}{\eta\delta})$, we have that for some $\delta < \frac{\boldsymbol{\lambda}}{d^{1.1}}$ and $t \ge T_{\mathbf{u},1} + \Omega(\frac{n_y^3 \log^2 d}{\eta\delta})$, it must hold that $\Psi_{\mathbf{u}}^{(t)} \ge B + o(1)$. However, for any $\delta < \frac{\boldsymbol{\lambda}}{d}$, if $\psi_{j,r}^{(t)}(g) + \psi_{j,r}^{(t)}(y') \le -\Omega(\frac{\varrho}{n_y \log d})$ for some $y' \ne y$, we have that

$$\left|\eta\mathbb{E}\left[\mathbf{logit}_{5,j}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{F_{5,j}\le B}\mathbb{1}_{\mathcal{B}_j(y')}\right]\right| \ge \Omega(\frac{\boldsymbol{\lambda}}{\mathrm{polylog} d}) \gg \frac{\boldsymbol{\lambda}}{d^{0.1}}$$

such that the gradient of $\psi_{j,r}^{(t)}(y')$ would greatly exceed the average growth speed of $\psi_{j,r}^{(t)}(g) - \sum_{y'\in\mathcal{Y}\setminus\{y\}} \psi_{j,r}^{(t)}(y')$. This would result in a fast increase of $\psi_{j,r}^{(t)}(y')$ to $-\psi_{j,r}^{(t)}(g)$. Due to this effect, we know that $\psi_{j,r}^{(t)}(y') + \psi_{j,r}^{(t)}(g) \ge \varrho/2$ for any $y' \ne y$ and $t \ge T_{\mathbf{u},1} + \Omega(\frac{n_y^3 \log^2 d}{\eta\delta})$.

Now we prove that there exist a iteration $t \geq T_{\mathbf{u},1} + \Omega(\frac{n_y^3 \log^3 d}{\eta\delta})$ such that $\mathcal{K}_{\mathsf{neg}}(\phi, \delta)$ does not hold. Suppose this is not the case, by the continuity of the gradient term

$$\eta\mathbb{E}\left[\mathbf{logit}_{5,j}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{F_{5,j}\leq B}\mathbb{1}_{\widetilde{\mathcal{B}}_\phi}\right]$$

we know that it must either be larger than δ or smaller than $-\delta$. Suppose the former is the case, then we have a always decreasing gradient for $\psi_{j,r}^{(t)}(g) + \psi_{j,r}^{(t)}(y')$ for more than $\Omega(\frac{n_y^3 \log^3 d}{\eta\delta})$ many iterations, which is impossible as that would make $\Psi_{j,r}^{(t)}(\phi) \leq O(1)$ after sufficient iterations. A similar argument can be applied to the case where the gradient is always negative. Therefore there must exist a iteration $t \geq T_{\mathbf{u},1} + \Omega(\frac{n_y^3 \log^3 d}{\eta\delta})$ such that $\mathcal{K}_{\mathsf{neg}}(\phi, \delta)$ holds. This concludes the proof. □

Corollary C.2 (incorrect feature cancellation)**.** *Let $j \in \tau(\mathcal{Y})$ and $\phi = (g, y) \in \mathfrak{F}_j$ and $\delta = (d^{0.01}\sigma_0)^{q-2}$, then there exists a iteration $T \geq T_{\mathbf{u},1} + \Omega(\frac{n_y^3 \log^2 d}{\delta})$ such that for all $t \geq T$, we have*

$$|\Psi_{j,r}^{(t)}(\widetilde{\phi})| \leq (d^{1+0.01\times(q-2)}\sigma_0^{q-2}/\boldsymbol{\lambda})^{1/(q-1)} \quad \textit{for any } \widetilde{\phi} \in \mathfrak{F}_{\mathsf{conf}}(\phi)$$

Proof. By choosing $\delta = d^{0.01}\sigma_0^{q-2}$ in Lemma C.11, we have that there exists an iteration $t \geq T_{\mathbf{u},1} + \Omega(\frac{n_y^3 \log^2 d}{\delta})$ such that

$$\left|\eta\mathbb{E}\left[\mathbf{logit}_{5,j}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{F_{5,j}\leq B}\mathbb{1}_{\widetilde{\mathcal{B}}_\phi}\right]\right| \leq (d^{0.01}\sigma_0)^{q-2}$$

accounting for the fact that there is an absolute logit lower bound $\mathbf{logit}_{5,j}^{(t)} \geq \Omega(\frac{\boldsymbol{\lambda}}{n_y d})$, we have that

$$\Lambda_{5,j,r}^{(t)}(\mathbf{Z})\mathbb{1}_{\widetilde{\mathcal{B}}_\phi} \in (-(d^{1+0.01\times(q-2)}\sigma_0^{q-2}/\boldsymbol{\lambda})^{1/(q-1)}, (d^{1+0.01\times(q-2)}\sigma_0^{q-2}/\boldsymbol{\lambda})^{1/(q-1)})$$

By looking into the features in $\Lambda_{5,j,r}^{(t)}\mathbb{1}_{\widetilde{\mathcal{B}}_\phi}$ we can conclude that

$$|\Psi_{j,r}^{(t)}(\widetilde{\phi})| \leq (d^{1+0.01\times(q-2)}\sigma_0^{q-2}/\boldsymbol{\lambda})^{1/(q-1)} \quad \text{for any } \widetilde{\phi} \in \mathfrak{F}_{\mathsf{conf}}(\phi)$$

for any $g' \neq g, y' \neq y$. One can also verify that at this point, $\nabla\psi_{j,r}^{(t)}(g')$ and $\nabla\psi_{j,r}^{(t)}(y')$ for $g' \neq g, y' \neq y$ contains only the terms with indicator function $\mathbb{1}_{\widetilde{\mathcal{B}}_\phi}$. Moreover, at this point $\Psi_{\mathbf{u}}^{(t)} \geq B - \widetilde{O}(\sigma_0)$ with gradient $\Omega(\boldsymbol{\lambda}/n_y^2)$ as long as it's all activated, thus the gradient of $\psi_{j,r}^{(t)}(g')$ and $\psi_{j,r}^{(t)}(y')$ cannot move beyond $(d^{0.01}\sigma_0)^{q-2}$ in the following iterations. This concludes the proof. □

Combining Lemma C.10 and Corollary C.2, we can now prove the following proposition:

Proposition C.4. Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$, then the following holds:

- At $t \geq T_{\mathbf{u},2} = T_{\mathbf{u},1} + O(\frac{1}{\eta d^{0.01}\sigma_0^{q-2}})$, we have $\Psi_{\mathbf{u}}^{(t)} \geq B - O(d^{0.01}\sigma_0)$.
- For any $\phi' \in \mathfrak{F}_{\mathsf{conf}}(\phi)$, we have $\Psi_{j,r}^{(t)}(\phi') \leq O((d^{0.01}\sigma_0)^{q-2}d/\boldsymbol{\lambda})^{1/(q-1)}$ if $t \geq T_{\mathbf{u},2}$.

This proved Induction C.3 for $t \in [T_{\mathbf{u},1}, T_1]$.

C.4.1 Feature Shape at Convergence

Next we shall characterize the magnitude of $\psi_{j,r}^{(t)}(g)$ and $\psi_{j,r}^{(t)}(y)$ for any $g \in \mathcal{G}$ and $y \in \mathcal{Y}$ at the end of training.

Lemma C.12 (symmetry of ϕ in the end)**.** *Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$ where $\phi = (g, y) \in \mathfrak{F}_j$. Suppose $\Psi_{\mathbf{u}}^{(t)} \geq B - O(d^{0.01}\sigma_0)$ and $\max_{\phi' \in \mathfrak{F}_{\mathsf{conf}}(\phi)} |\Psi_{j,r}^{(t)}(\phi')| \leq O(\delta)$ for all $t \geq T_{\mathbf{u},2}$, then we have*

$$\mathcal{J}_\phi^{(t)} := |\psi_{j,r}^{(t)}(g) - \psi_{j,r}^{(t)}(y)| \leq O(\delta) \tag{29}$$

Moreover, plugging in $\delta = O((d^{0.01}\sigma_0)^{q-2}d/\boldsymbol{\lambda})^{1/(q-1)}$ as in Proposition C.4, we have that $\mathcal{J}_\phi^{(t)} \leq O((d^{0.01}\sigma_0)^{q-2}d/\boldsymbol{\lambda})^{1/(q-1)}$.

Proof. The proof uses Proposition C.4. Let's look at the growth of $\psi_{j,r}^{(t)}(g) = \langle \mathbf{W}_{5,j,r,2}^{(t)}, e_g \rangle$ and $\psi_{j,r}^{(t)}(y) = \langle \mathbf{W}_{5,j,r,5}^{(t)}, e_y \rangle$ for every $g \in \mathcal{G}$ and $y \in \mathcal{Y}$ over the during the period $s \in [0, t]$. First for any $\phi = (g, y) \in \mathfrak{F}_j$, we have

$$
\begin{aligned}
&\psi_{j,r}^{(t)}(g) - \psi_{j,r}^{(0)}(g) \\
&= \sum_{s=0}^{t-1} \eta \langle \nabla_{\mathbf{W}_{5,j,r,2}} \mathsf{Loss}^{(s)}, e_g \rangle \\
&= \sum_{s=0}^{t-1} \eta \bigg(\mathbb{E}\big[(1 - \mathbf{logit}_{5,j}^{(s)}) \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(g)} \mathbb{1}_{F_{5,j}^{(s)} \le B}\big] - \mathbb{E}\big[\mathbf{logit}_{5,j}^{(s)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\widetilde{\mathcal{B}}_j(g)} \mathbb{1}_{F_{5,j}^{(s)} \le B}\big] \bigg) \\
&= \sum_{s=0}^{t-1} \eta \bigg(\mathbb{E}\big[(1 - \mathbf{logit}_{5,j}^{(s)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(g)} \mathbb{1}_{F_{5,j}^{(s)} \le B}\big] \\
&\qquad - \sum_{s=0}^{t-1} \sum_{y' \neq y} \eta \mathbb{E}\big[\mathbf{logit}_{5,j}^{(s)} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}(g,y')} \mathbb{1}_{F_{5,j}^{(s)} \le B}\big] \bigg) \\
&= U_{g,y} - \sum_{y' \neq y} R_{g,y'}
\end{aligned} \tag{30}
$$

where the terms $U_{g,y}$ and $R_{g,y'}$ are defined as follows:

$$
\begin{aligned}
U_{g,y} &:= \sum_{s=0}^{t-1} \eta \mathbb{E}\big[(1 - \mathbf{logit}_{5,j}^{(s)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(g)} \mathbb{1}_{F_{5,j}^{(s)} \le B}\big] \\
R_{g,y'} &:= \sum_{s=0}^{t-1} \eta \mathbb{E}\big[\mathbf{logit}_{5,j}^{(s)} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}(g,y')} \mathbb{1}_{F_{5,j}^{(s)} \le B}\big]
\end{aligned}
$$

Similarly, the total growth of $\psi_{j,r}^{(t)}(y) = \langle \mathbf{W}_{5,j,r,5}^{(t)}, e_y \rangle$ is given by

$$
\begin{aligned}
&\psi_{j,r}^{(t)}(y) - \psi_{j,r}^{(0)}(y) \\
&= \sum_{s=0}^{t-1} \eta \bigg(\mathbb{E}\big[(1 - \mathbf{logit}_{5,j}^{(s)}) \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(y)} \mathbb{1}_{F_{5,j}^{(s)} \le B}\big] - \mathbb{E}\big[\mathbf{logit}_{5,j}^{(s)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\widetilde{\mathcal{B}}_j(y)} \mathbb{1}_{F_{5,j}^{(s)} \le B}\big] \bigg) \\
&= \sum_{s=0}^{t-1} \eta \bigg(\mathbb{E}\big[(1 - \mathbf{logit}_{5,j}^{(s)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(y)} \mathbb{1}_{F_{5,j}^{(s)} \le B}\big] \\
&\qquad - \sum_{s=0}^{t-1} \sum_{g' \neq g} \eta \mathbb{E}\big[\mathbf{logit}_{5,j}^{(s)} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}(g',y)} \mathbb{1}_{F_{5,j}^{(s)} \le B}\big] \bigg) \\
&= U_{g,y} - \sum_{g' \neq g} R_{g',y}
\end{aligned} \tag{31}
$$

When $\widetilde{\phi} = (\widetilde{g}, \widetilde{y}) \in \mathfrak{F}_j$ but $\widetilde{\phi} \neq \phi$, suppose $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$, then we have $\Lambda_{5,j,r}^{(s)} \mathbb{1}_{\mathcal{B}_{\widetilde{g},\widetilde{y}}} \le \widetilde{O}(\sigma_0)$ by Induction C.3 for all $s \in [0, \mathcal{T}_1]$. Therefore, we can further compute

$$
\begin{aligned}
U_{\widetilde{g},\widetilde{y}} &= \sum_{s=0}^{t-1} \eta \mathbb{E}\big[(1 - \mathbf{logit}_{5,j}^{(s)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_{\widetilde{g},\widetilde{y}}} \mathbb{1}_{F_{5,j}^{(s)} \le B}\big] \\
&\le \sum_{s=0}^{t-1} O(\eta d^{0.001} \sigma_0^{q-1}) && \text{(by the smoothness of } \mathbf{sReLU}') \\
&\le O(\eta d^{0.001} \sigma_0^{q-1}) \cdot \widetilde{O}\Big(\frac{1}{\eta \sigma_0^{q-2}}\Big) && (t \le \mathcal{T}_1 \le \widetilde{O}(\tfrac{1}{\eta \sigma_0^{q-2}}) \text{ by Induction C.3}) \\
&\le O(d^{0.002} \sigma_0)
\end{aligned}
$$

Similarly, we have $R_{\widetilde{g},\widetilde{y}} = O(d^{0.002}\sigma_0)$ for all $t \le \mathcal{T}_1$. Thus for any $\phi' = (g', y') \neq \phi \in \mathfrak{F}_j$, we have

$$\psi_{j,r}^{(t)}(g') = \psi_{j,r}^{(0)}(g') + U_{g',y'} - \sum_{y'' \neq y'} R_{g',y''} = O(d^{0.003}\sigma_0) - R_{g',y}$$
$$\text{and } \psi_{j,r}^{(t)}(y') = \psi_{j,r}^{(0)}(y') + U_{g',y'} - \sum_{g'' \neq g} R_{g'',y'} = O(d^{0.003}\sigma_0) - R_{g,y'}$$

Now combined with Equation (30) and Equation (31), we have

$$\psi_{j,r}^{(t)}(g) = \psi_{j,r}^{(0)}(g) + U_{g,y} - \sum_{y' \neq y} R_{g,y'} = U_{g,y} + \sum_{y' \neq y} \psi_{j,r}^{(t)}(y') \pm O(d^{0.01}\sigma_0)$$
$$\text{and } \psi_{j,r}^{(t)}(y) = \psi_{j,r}^{(0)}(y) + U_{g,y} - \sum_{g' \neq g} R_{g,y'} = U_{g,y} + \sum_{g' \neq g} \psi_{j,r}^{(t)}(g) \pm O(d^{0.01}\sigma_0)$$

However, since we assumed the condition $\mathcal{K}_{\mathsf{neg}}(\mathbf{u}, \delta)$ holds at t, which implies that $|\psi_{j,r}^{(t)}(g) + \psi_{j,r}^{(t)}(y')| \le O(\delta)$ for any $y' \neq y$, and similarly $|\psi_{j,r}^{(t)}(y) + \psi_{j,r}^{(t)}(g')| \le O(\delta)$. Therefore, we can infer that

$$\psi_{j,r}^{(t)}(g) = U_{g,y} + \sum_{y' \neq y} \psi_{j,r}^{(t)}(y') \pm O(d^{0.01}\sigma_0) = U_{g,y} - (n-1)\psi_{j,r}^{(t)}(g) \pm O(n\delta)$$

which implies $\psi_{j,r}^{(t)}(g) = U_{g,y}/n \pm O(\delta)$, we can similarly derive that $\psi_{j,r}^{(t)}(y) = U_{g,y}/n \pm O(\delta)$. Therefore, we have

$$|\psi_{j,r}^{(t)}(g) - \psi_{j,r}^{(t)}(y)| \le O(\delta), \qquad \forall \delta > d^{0.01}\sigma_0$$

This also implies that any $\psi_{j,r}^{(t)}(g')$ for any $g' \neq g$ must be as small as $-\psi_{j,r}^{(t)}(y)$ since it cancels with $\psi_{j,r}^{(t)}(y)$, and similarly $\psi_{j,r}^{(t)}(y') \le -\Omega(\log d)$ for any $y' \neq y$. □

At the end of induction, we have the following theorem as the training result:

Theorem C.1 (learning cyclic group action)**.** *At iteration$t = \mathcal{T}_1$, the following properties hold:*

(A) Optimal loss: The training loss is optimal for $i = 5$:

$$\mathsf{Loss}_5^1(F^{(t)}) \le \frac{1}{\mathsf{poly}(d)}$$

(B) Sparse activations: For $j \in \tau(\mathcal{Y})$, let $\phi = (g, y) \in \mathfrak{F}_j$, then there exists exactly one activated *neuron $r \in [m]$ such that when $g_1 = g, y_0 = y$ happens:*

$$\Lambda_{5,j,r}^{(T_1)} \ge B - O(d^{0.01}\sigma_0) \quad \textit{while} \quad \Lambda_{5,j,r'}^{(\mathcal{T}_1)} \le O(d^{-\Omega(1)}) \quad \forall r' \neq r$$

(C) Cancellation of incorrect features: For $j \in \tau(\mathcal{Y})$, let $\phi = (g, y) \in \mathfrak{F}_j$, and let the $r \in [m]$ be the activated neuron in (B), then for any $g' \neq g \in \mathcal{G}$, and any $y' \in \mathcal{Y}$, we have

$$|\psi_{j,r}^{(T_1)}(g) + \psi_{j,r}^{(\mathcal{T}_1)}(y')| \le O(\delta) \quad \textit{and} \quad |\psi_{j,r}^{(T_1)}(g') + \psi_{j,r}^{(T_1)}(y)| \le O(\delta)$$

for some $\delta = O(((d^{0.01}\sigma_0)^{q-2}d/\boldsymbol{\lambda})^{1/(q-1)})$

Proof. Once all features are learned, that is, the last $\mathbf{u} \in \mathcal{U}^\star$ is learned and $t = T_1$, the correct $\mathbf{logit}_{5,j}(F^{(t)})$ for $j \in \tau(\mathcal{Y})$ will be $1 - O(\frac{e^B}{e^B + d})$ which is optimal, and therefore (A) is correct. (B) is correct following Induction C.3 and (C) is proven in Proposition C.4, for every $\mathbf{u} \in \mathcal{U}^\star$. □

D Learning The Group Actions: Symmetry Group

Similar to the analysis of the cyclic group actions, we first need to define the features for the symmetry group. We first introduce the same set of notations.

Notations. Let $\mathcal{D}^1$ be the LEGO distribution of length 1 under the language $\mathbf{LEGO}(\mathcal{X}, \mathcal{G}, \mathcal{Y})$. We define $\mathcal{D}^1_{\mathcal{X}}$, $\mathcal{D}^1_{\mathcal{G}}$ and $\mathcal{D}^1_{\mathcal{Y}}$ be the distribution of (x_0, x_1), g_0 and (y_0, y_1) in $\mathcal{D}^1$ respectively. That is, given a LEGO sentence

$$Z^{(1,0)} = (Z_{\mathsf{pred},1}, Z_{\mathsf{ans},0}, Z_{\mathsf{ans},1}) \sim \mathcal{D}^1,$$
$$Z_{\mathsf{pred},1} = (x_0, g_1, x_1, \langle\mathsf{blank}\rangle, \langle\mathsf{blank}\rangle), \quad Z_{\mathsf{ans},i} = (\langle\mathsf{blank}\rangle, \langle\mathsf{blank}\rangle, \langle\mathsf{blank}\rangle, x_i, y_i), i \in \{0, 1\}$$

The sampling distribution of (x_0, x_1) is $\mathcal{D}^1_{\mathcal{X}}$, and similarly for g_0 and (y_0, y_1).

We restate the assumption on the symmetry group $\mathcal{G}$ and its action here.

Assumption D.1 (Assumption 4.2, restated)**.** Let $\mathbf{LEGO}(\mathcal{X}, \mathcal{G}, \mathcal{Y})$ be the LEGO language. We assume $\mathcal{Y} = \{0, 1, \ldots, n_y - 1\}$ and $\mathcal{G} = \mathbf{Sym}(\mathcal{Y})$, i.e., the symemtry group of order $n_y!$. We assume $n_y = \Theta(\frac{\log\log d}{\log\log\log d})$ and $n_y! = \Theta(\mathrm{polylog}(d)) \gg \frac{1}{\varrho}$.

We also redefine the feature combinations in the symmetry group case. For symmetric group $\mathcal{G}$, we shall define a new notion called *fiber* of a value.

Definition D.1 (fiber of values)**.** Assuming the group $\mathcal{G}$ follows Assumption D.1. For each $j \in \tau(\mathcal{Y})$ and each $y \in \mathcal{Y}$, we denote the **fiber**

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

denotes all group elements g that sends y to $y' = \tau^{-1}(j)$.

Definition D.2 (feature combinations, symmetric group)**.** Assuming the group $\mathcal{G}$ follows Assumption D.1. Let $y, y' \in \mathcal{Y}$ be a pair of values, we define the following set $G_{y\to y'}$: and we call it the set of **transition from** y **to** y'. Now for each $j \in \tau(\mathcal{Y})$, let

$$\mathfrak{F}_j := \{(\mathrm{Fiber}_{j,y}, y) \in \mathcal{G} \times \mathcal{Y}\},$$

we call the set $\mathfrak{F} = \bigcup_{j\in\tau(\mathcal{Y})} \mathfrak{F}_j$ the set of feature combinations, and the sets $\mathfrak{F}_j$ are called set of **feature combinations** for predicting $y' = \tau^{-1}(j)$. Similar to Definition C.1, for any $\phi = (\mathrm{Fiber}_{j,y}, y) \in \mathfrak{F}$, we write

$$\mathfrak{F}_{\mathsf{conf}}(\phi) := \{\phi' \in \mathfrak{F} \mid \phi' = (\mathrm{Fiber}_{j',y}, y), j \neq j' \text{ or } \phi' = (\mathrm{Fiber}_{j,y}, y'), y' \neq y\}$$

as the set of **confounding features** for $\phi = (\mathrm{Fiber}_{j,y}, y)$.

The ψ and Ψ notations are redefined as follows.

Definition D.3 (ψ, Ψ notations, symmetry group)**.** Using the same notation of ψ in Definition C.3. For each $j \in \tau(\mathcal{Y})$, let $\phi = (g, y) \in \mathfrak{F}_j$, we define the feature magnitude $\Psi_{\mathbf{u}}$ and $\Psi_{\mathbf{u},\max}, \Psi_{\mathbf{u},\min}$ as follows:

$$\Psi_{\mathbf{u}} \equiv \Psi_{j,r}(\phi) := \frac{1}{(n_y - 1)!} \sum_{g\in\mathrm{Fiber}_{j,y}} \frac{1}{2}(\psi_{j,r}(g) + \psi_{j,r}(y))$$

$$\Psi_{\mathbf{u},\max} := \max_{g\in\mathrm{Fiber}_{j,y}} \frac{1}{2}(\psi_{j,r}(g) + \psi_{j,r}(y)), \quad \Psi_{\mathbf{u},\min} := \min_{g\in\mathrm{Fiber}_{j,y}} \frac{1}{2}(\psi_{j,r}(g) + \psi_{j,r}(y))$$

Here $\mathrm{Fiber}_{j,y} = \{g \in \mathcal{G} \mid \tau(g \cdot y) = j\}$ is the fiber of y under the action of $\mathcal{G}$.

The learning order in the symmetry group case is the same as in Definition C.4, but with $\Psi_{\mathbf{u}}$ defined in Definition D.3, which we do not repeat here. The definition of pseudo weights remains the same. We restate the learning order here.

Definition D.4 (learning order)**.** The *learning order* is the ordered set $\mathcal{U}^\star$ that we obtain from the following process: Define a total order on $\mathcal{U}$ as follows:

$$\mathbf{u} \prec \mathbf{u}' \iff \Psi^{(0)}_{\mathbf{u}} \geq \Psi^{(0)}_{\mathbf{u}'} \quad \forall \mathbf{u}, \mathbf{u}' \in \mathcal{U} \tag{32}$$

We construct the sets $\mathcal{U}^\star$ by the following procedure: initialize an empty neuron set $\mathcal{W}^{(0)}_{tmp} = \varnothing$, and an empty feature set $\mathcal{R}^{(0)}_{tmp} = \varnothing$, and the initial index set $\mathcal{U}^{(0)} = \varnothing$. Starting from $k = 1$, we do the following:

(1) Find the index $\mathbf{u} = (j, r, \phi) \in \mathcal{U}$, where $(j, r, \phi) = \arg\max_{j', r', \phi'} \Psi^{(0)}_{j', r'}(\phi')$ such that the feature $\phi \in \mathfrak{F} \setminus \mathcal{R}^{(k-1)}_{tmp}$ and $(j, r) \in \tau(\mathcal{Y}) \times [m] \setminus \mathcal{W}^{(k-1)}_{tmp}$.

(2) Update $\mathcal{R}^{(k)}_{tmp} \leftarrow \mathcal{R}^{(k-1)}_{tmp} \cup \{\phi\}$, $\mathcal{W}^{(k)}_{tmp} \leftarrow \mathcal{W}^{(k-1)}_{tmp} \cup \{(j, r)\}$, and $\mathcal{U}^{(k)} \leftarrow \mathcal{U}^{(k-1)} \cup \{\mathbf{u}\}$.

(3) Iterate the (1) and (2) steps until $k = n_y^2$, then yield $\mathcal{U}^\star \equiv \mathcal{U}^{(n_y^2)}$.

This process yields the ordered set $\mathcal{U}^\star$, equipped with the total order $\prec$ defined in (32).

D.1 Induction Hypothesis and Training Phases

Again we define some useful probabilistic events.

Definition D.5 (probability events of feature appearance)**.** Let $j \in \tau(\mathcal{Y})$, $r \in [m]$ and $\phi = (\mathrm{Fiber}_{j,y}, y) \in \mathfrak{F}_j$ be a feature combination that predicts j-th output. Denote events $\mathcal{B}_\phi, \mathcal{B}(g, y), \mathcal{B}_j(g), \mathcal{B}_j(y)$ and $\widetilde{\mathcal{B}}_\phi, \widetilde{\mathcal{B}}_j(g), \widetilde{\mathcal{B}}_j(y)$ as follows:

1. $\mathcal{B}_\phi \equiv \mathcal{B}(\mathrm{Fiber}_{j,y}, y) \equiv \mathcal{B}_j(y) \equiv \mathcal{B}_j(\mathrm{Fiber}_{j,y})$, which is defined as
$$\mathcal{B}_\phi := \{g_1 \in \mathrm{Fiber}_{j,y}, y_0 = y\}$$
2. For individual $g \in \mathrm{Fiber}_{j,y}$, we let $\mathcal{B}_{g,y} := \{g_1 = g, y_0 = y\}$;
3. $\widetilde{\mathcal{B}}_j(g) := \{g_1 \in \mathrm{Fiber}_{j,y}, y_0 \neq y\}$, the event that $g \in \mathrm{Fiber}_{j,y}$ did not appear together with y for predicting j-th output. Moreover, we define $\widetilde{\mathcal{B}}_j(\mathrm{Fiber}_{j,y}) = \bigcup_{g \in \mathrm{Fiber}_{j,y}} \widetilde{\mathcal{B}}_j(g)$;
4. $\widetilde{\mathcal{B}}_j(y) := \{g_1 \notin \mathrm{Fiber}_{j,y}, y_0 = y\}$, the event that y did not appear together with $g \in \mathrm{Fiber}_{j,y}$ for predicting j-th output;
5. $\widetilde{\mathcal{B}}_\phi := \widetilde{\mathcal{B}}_j(\mathrm{Fiber}_{j,y}) \cup \widetilde{\mathcal{B}}_j(y)$, the event that the appeared feature combination ϕ' is wrong for predicting j-th output.

Similar to above, we give several induction hypoetheses, each characterize different aspects of the process.

We then define the notion of the gradient conditions similar to the cyclic group case.

Definition D.6 (gradient criterion, symmetry group)**.** Let $\mathbf{u} = (j, r, \phi) \in \mathcal{O}^\star$ and $t \leq T_1$, we define the following two conditions:

- Given $\delta > 0$, the positive gradient criterion $\mathcal{K}_{\mathsf{pos}}(\mathbf{u}, \delta)$:
$$\mathcal{K}_{\mathsf{pos}}(\mathbf{u}, \delta) \text{ is true } \iff \left|\mathbb{E}\left[(1 - \mathbf{logit}_{5,j}) \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\mathcal{B}_\phi} \mathbb{1}_{F_{5,j} \leq B}\right]\right| \leq \delta \tag{33}$$
- Given $\delta > 0$, the negative gradient criterion $\mathcal{K}_{\mathsf{neg}}(\mathbf{u}, \delta)$:
$$\mathcal{K}_{\mathsf{neg}}(\mathbf{u}, \delta) \text{ is true } \iff \left|\mathbb{E}\left[\mathbf{logit}_{5,j} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\widetilde{\mathcal{B}}_\phi} \mathbb{1}_{F_{5,j} \leq B}\right]\right| \leq \delta \tag{34}$$

These conditions control the magnitude of the gradient of the feature $\mathbf{u}$ at iteration t. Now we define the following intermediate time-steps:

Definition D.7 (phase decomposition, symmetry group)**.** Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$, we define the following intermediate time-steps:

$$\begin{aligned}
T_{\mathbf{u},1a} &:= \min\{t \geq 0 \mid \Psi^{(t)}_{\mathbf{u},\min} \geq d^{0.01}\sigma_0\} \\
T_{\mathbf{u},1} &:= \min\{t \geq 0 \mid \mathcal{K}_{\mathsf{pos}}(\mathbf{u}, \delta_1), \text{ where } \delta_1 := 1 - d^{-0.1}\} \\
T_{\mathbf{u},2} &:= \min\left\{t \geq 0 \mid \mathcal{K}_{\mathsf{pos}}(\mathbf{u}, \delta_3) \wedge (\Psi^{(t)}_{\mathbf{u}} \geq B - O(d^{0.01}\sigma_0)), \text{ where } \delta_3 := d^{0.01}\sigma_0^{q-2}\right\}
\end{aligned}$$

Below we shall introduce some induction hypotheses that will be used in the proof. We first introduce a induction hypothesis for the pseudo weights.

Induction D.1 (induction on pseudo weight bounds)**.** *Let $j \in \tau(\mathcal{Y})$ and $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$. Let $\mathbf{u}' \prec \mathbf{u} \in \mathcal{U}^\star$ be the immediate predecessor of $\mathbf{u}$ in $\mathcal{U}^\star$. Then at $t = T_{\mathbf{u}',2}$, it holds that the pseudo weights $\widetilde{\mathbf{W}}_{5,j,r}^{(t)}(\mathbf{u})$ defined in Definition C.5 satisfies*

$$\left|\langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t)}(\mathbf{u}), e_v\rangle - \langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v\rangle\right| \le \widetilde{O}(\sigma_0/d^{\Omega(1)}),$$

where $p = 2, v \in \mathcal{G}$ or $p = 5, v \in \mathcal{Y}$.

We maintain the following induction hypotheses for the case of Assumption D.1.

Induction D.2 (induction on weight bounds)**.** *Assuming Assumption 4.1, for $t \le T_1$, the following properties hold:*

(A) $|\langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v\rangle - \langle \mathbf{W}_{5,j,r,p}^{(0)}, e_v\rangle| \le \widetilde{O}(1/d)$ *for all $v \in \mathcal{X}$ and $p \in \{1, 3, 4\}$;*

(B) $|\langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v\rangle - \langle \mathbf{W}_{5,j,r,p}^{(0)}, e_v\rangle| \le \widetilde{O}(\sigma_0/d)$ *for all $v \in \mathcal{V}$ and $p \in [5], r \in [m]$ if $j \notin \tau(\mathcal{Y})$;*

The last induction hypothesis is about the checkpoints defined in Definition D.7.

Induction D.3 (induction on cyclic group actions)**.** *Under Assumption D.1, for $t \le T_1$, in addition to the induction hypotheses in Induction D.1 and D.2, the following properties hold:*

(A) For any $\mathbf{u}' \prec \mathbf{u} \in \mathcal{U}^\star$, it holds that $T_{\mathbf{u},1a} \ge T_{\mathbf{u}',2} + \widetilde{\Omega}(1/\eta\sigma_0^{q-2})$;

(B) Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$, for any $t \in [T_{\mathbf{u},2}, T_1]$, it holds that $\Psi_{j,r}^{(t)}(\phi) \ge B - O(d^{0.01}\sigma_0)$ and

$$\max_{\phi' \in \mathfrak{F}_{\mathsf{conf}}(\phi)} \Psi_{j,r}^{(t)}(\phi') \le ((d^{0.01}\sigma_0)^{q-2} d/\boldsymbol{\lambda})^{1/(q-1)}$$

The proof will proceed in a similar way as in the cyclic group case, with some additional technical details. We first introduce some lemmas that will be used in the proof, many of them are similar to the ones in the cyclic group case.

Lemma D.1 (initialization gap between features)**.** *Assuming Assumption 4.1. Let $j \in \tau(\mathcal{Y})$, for all $r \in [m]$ and any two $\mathbf{u}, \mathbf{u}' \in \mathcal{U}$, we have with prob $\ge 1 - o(1)$ over the randomness at initialization that*

$$|\Psi_{\mathbf{u}}^{(0)} - \Psi_{\mathbf{u}'}^{(0)}| \gtrsim \frac{\sigma_0}{n_y^4 m^2 \log d}$$

Proof. the proof is similar to that in Lemma C.2, one simply needs to modify the feature g that appears in the proof. □

Lemma D.2 (gradient bounds)**.** *Let $j \in \tau(\mathcal{Y})$, and $\phi = (\mathrm{Fiber}_{j,y}, y) \in \mathfrak{F}$. Suppose Induction D.3 is satisfied at $t \le T_1$, then for any $\delta \in [0, 0.4]$, if $\sum_{r\in[m]} \Psi_{j,r,\max}^{(t)}(\phi) \le (0.5 + \delta)\log d$, it holds that*

(a) $\mathbb{E}[(1 - \mathbf{logit}_{5,j}^{(t)}) \mid \mathcal{B}_\phi] \ge 1 - d^{-0.49+\delta}$

(b) $\mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mid \widetilde{\mathcal{B}}_\phi] \le d^{-0.49+\delta}$

(c) for any $r \in [m]$, for any $g \in \mathrm{Fiber}_{j,y}$, it holds that with $v = g$ or $v = y$,

$$\nabla_{\psi_{j,r}^{(t)}(v)} \mathsf{Loss}^{(t)} \ge (1 - \widetilde{O}(\frac{1}{d^{0.49-\delta}}))\mathbb{E}[\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})\mathbb{1}_{\mathcal{B}_{g,y}}\mathbb{1}_{F_{5,j}\le B}]$$

Also, we can control the irrelavant features in the same way as in the cyclic group case.

Lemma D.3 (irrelavant features)**.** *Suppose Induction D.3 holds for all $t < T_1$, then Induction D.2a holds at iteration $t + 1$. Moreover, let $\mathcal{A}_x = \{x \in \{x_0, x_1\}, (x_0, x_1) \sim \mathcal{D}_{\mathcal{X}}^1\}$, at each step the following holds:*

$$|\langle \mathbf{W}_{5,j,r,p}^{(t+1)}, e_x\rangle - \langle \mathbf{W}_{5,j,r,p}^{(t)}, e_x\rangle| \le \sum_{g\in\mathcal{G}} \eta \mathbf{Pr}(\mathcal{A}_x) \cdot |\langle \nabla_{\mathbf{W}_{5,j,r,2}} \mathsf{Loss}^{(t)}, e_g\rangle|$$

The proofs of the above lemmas are similar to that in the cyclic group case, we omit them here.

D.2 Phase I: Emergence of the Feature

We need to show that the feature growth in Phase I is again following the pseudo weight trajectory. We give the same version of the proposition as in the cyclic group case.

Proposition D.1 (Phase I.a)**.** Let $\mathbf{u}'$ be the immediate predecessor of $\mathbf{u}$ in $\mathcal{U}^\star$, then Induction D.1 is satisfied for all iterations $t \le T_{\mathbf{u}',2}$. Moreover, (25) holds at $t = T_{\mathbf{u}',2}$ for the symmetry group case.

The proof of Proposition D.1 follows the same line as in the cyclic group case, we omit it here. This proposition guarantees that the feature growth before $T_{\mathbf{u}',2}$ is small and rather negligible.

Proposition D.2 (Phase I)**.** Let $\mathbf{u}'$ be the immediate predecessor of $\mathbf{u}$ in $\mathcal{U}^\star$, then Induction D.3 is satisfied for all iterations $t \le T_{\mathbf{u}',2}$.

Now we analyze the feature competition in Phase Ia.

Proposition D.3 (Phase I.a, feature competition)**.** Let $\mathbf{u}' \prec \mathbf{u} \in \mathcal{U}^\star$, then Induction D.3 holds for all iterations $t \in [0, T_{\mathbf{u},1a}]$. Moreover, we have the following properties:

$$\Psi_{\mathbf{u},\min}^{(T_{\mathbf{u},1a})} \ge d^{0.01}\sigma_0, \quad \text{while } \Psi_{\widetilde{\mathbf{u}},\max}^{(T_{\mathbf{u},1a})} = \widetilde{O}(\sigma_0), \quad \forall \widetilde{\mathbf{u}} \succ \mathbf{u} \in \mathcal{U}$$

Proof. The proof is similar to that in the cyclic group case. The caveat here is that we need to consider a different version of TPM by applying Lemma E.2 with Lemma D.1. In fact, since each $\mathrm{Fiber}_{j,y}$ contains $(n_y - 1)!$ elements, we can transform the problem into the tensor power method of the growth of feature $\psi_{j,r}(y)$, averaged over all all combinations with $g \in \mathrm{Fiber}_{j,y}$. Since the constant is much smaller for group feature g, we only need to control the growth of the feature $\psi_{j,r}(y)$ for comparison. □

Finally, we show the result for Phase I.

Proposition D.4 (Phase I, symmetry group)**.** Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$, then Induction D.3 holds for all $t \le T_{\mathbf{u},1}$, and we have the following results at $t = T_{\mathbf{u},1}$:

(A) $\Psi_{\mathbf{u},\min}^{(T_{\mathbf{u},1})}, \Psi_{\mathbf{u},\max}^{(T_{\mathbf{u},1})} \ge \Omega(\log d)$;

(B) For any $\widetilde{\mathbf{u}} \in \mathcal{U}$ such that $\widetilde{\mathbf{u}} \succ \mathbf{u}$, we have $\Psi_{\widetilde{\mathbf{u}}}^{(T_{\mathbf{u},1})} \le \widetilde{O}(\sigma_0)$.

Proof. The proof is similar to that in the cyclic group case but with some caveats. We still mainly employ Lemma E.2 along with gradient approximation in Lemma D.2. □

D.3 Phase II: Convergence of the Feature

Beginning with Phase II, we need to show that the feature $\Psi_{\mathbf{u}}^{(t)}$ continues to grow despite very complex landscape.

Firstly, we have some similar results as in the cyclic group case.

Lemma D.4 (activeness of a neuron)**.** *Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$ and $\phi = (\mathrm{Fiber}_{j,y}, y)$. If Induction D.3 holds, then for all $t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$, it holds that $\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) = 1$ conditioned on $\mathcal{B}_\phi$.*

Proof. The proof is similar to that in the cyclic group case but relies on Lemma D.2 in a slightly different way. Here we need to consider the gradient lower bound when $\Psi_{\mathbf{u},\min}^{(t)} \le \frac{1}{2}\log d$. This already provides enough gradient to keep the neuron active for specific feature combination (g, y) where $g \in \mathrm{Fiber}_{j,y}$. □

also, we have the following lemma for the same feature in different neurons.

Lemma D.5 (bounds on same feature in different neurons)**.** *Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}$, where $\phi = (g, y)$, and $\mathbf{u}' = (j, r', \phi) \in \mathcal{U}$ for some $r' \ne r$ such that $\mathbf{u}' \succ \mathbf{u}$, we have that $\Psi_{\mathbf{u}'}^{(t)} \le \widetilde{O}(\sigma_0)$ for all $t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$.*

Proof. The proof is similar, only requiring some bounds on the gradient trajctory for $t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$. □

Similarly, we classify the logits in phase II into two categories.

Lemma D.6 (logits in phase II)**.** *Consider a $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$, let $\mathfrak{F}^{(t)}$ contains all the feature combinations from $\mathfrak{F}$ that are learned at iteration t, that is, for each $\phi' \in \mathfrak{F}^{(t)}$, there exists a $\mathbf{u}' \in \mathcal{U}^\star$ such that $\mathbf{u}' \preceq \mathbf{u}$, which also means $t \geq T_{\mathbf{u}',2}$. Then for any $\phi' \in \mathfrak{F}^{(t)} \cap \mathfrak{F}_{\mathsf{conf}}(\phi)$, we have that for all $t \geq T_{\mathbf{u},2}$,*

$$\mathbf{logit}_{5,j}^{(t)} \mathbb{1}_{\mathcal{B}_{\phi'}} \leq O(\exp(\Psi_{j,r,\max}^{(t)}(\phi') - B)) = \Theta(\exp(\Psi_{j,r,\max}^{(t)}(\phi')\boldsymbol{\lambda}/d))$$

otherwise, we have that $\mathbf{logit}_{5,j}^{(t)} \mathbb{1}_{\mathcal{B}_{\phi'}} \geq \Omega(\exp(\Psi_{j,r,\min}^{(t)}(\phi')/d))$.

Now based on the classification of logits, we can derive the following lemma, which concerns with different cancellation conditions.

Lemma D.7 (different cancellation conditions)**.** *Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$ and $\phi = (\mathrm{Fiber}_{j,y}, y)$, let $\mathfrak{F}^{(t)}$ contains all the feature combinations from $\mathfrak{F}$ that are learned at iteration t as in Lemma D.6. Let $g \in \mathrm{Fiber}_{j,y}$, then the following holds: Here we abuse the notation and revert to using those in Definition C.1.*

- *let $g \in \mathrm{Fiber}_{j,y}$, then before $\psi_{j,r}^{(t)}(g) + \psi_{j,r}^{(t)}(y)$ exceeds $B/2$, $\Psi_{j,r,\min}^{(t)}(\phi') \geq \Omega(\log d)$ for $\phi' = (g', y), g' \notin \mathrm{Fiber}_{j,y}$, or $\phi' = (g, y')$ for any $y' \neq y$.*
- *Otherwise, we have that $\Psi_{j,r,\max}^{(t)}(\phi') \leq \frac{1}{d^{\Omega(1)}}$ once $\Psi_{\mathbf{u},\min}^{(t)} \geq 2.01 \log d$ for $\phi' = (g', y), g' \notin \mathrm{Fiber}_{j,y}$, or $\phi' = (g, y')$ for any $y' \neq y$.*

Proof. The proof of this lemma requires slightly more delicate analysis. Suppose let $g \in \mathrm{Fiber}_{j,y}$ and $y' \neq y$ we have that $\psi_{j,r}^{(t)}(g) + \psi_{j,r}^{(t)}(y') \leq B/2$. Then by Lemma D.6, we have that the gradient of $\psi_{j,r}^{(t)}(y')$ remains small, and therefore the feature $\psi_{j,r}^{(t)}(y')$ cannot cancel out the feature $\psi_{j,r}^{(t)}(g)$. □

Now we show how the correct features can grow to maximum value possible, as stipulated by the upper bound assumption Assumption A.1.

Lemma D.8 (convergence of positive gradient)**.** *Let $j \in \tau(\mathcal{Y})$ and $\phi \in \mathfrak{F}_j$. For any level $\delta > d\sigma_0^{q-2}$, the total number of iterations of which the condition $\mathcal{K}_{\mathsf{pos}}(\phi, \delta)$ does not hold must be smaller than $O(\frac{n_y^3 \log^2 d}{\eta\delta})$.*

Proof. The proof is similar to that in the cyclic group case. Instead of analyzing a single pair of g, y, we consider the growth of feature $\psi_{j,r}^{(t)}(g)$ and $\psi_{j,r}^{(t)}(y)$ for all $g \in \mathrm{Fiber}_{j,y}$ and $y \in \mathcal{Y}$. By using the same analysis as in Lemma C.10, we can show that the following difference

$$\sum_{g \in \mathrm{Fiber}_{j,y}} \psi_{j,r}^{(t)}(g) - \sum_{y' \neq y} \psi_{j,r}^{(t)}(y')$$

is non-decreasing. Same technique can be applied another difference:

$$\psi_{j,r}^{(t)}(y) - \sum_{g \in \bigcup_{y' \neq y} \mathrm{Fiber}_{j,y'}} \psi_{j,r}^{(t)}(g)$$

which is also non-increasing. By computing their possible growth upper bounds, we can show that they can grow for at most $O(\frac{n_y^3 \log^2 d}{\eta\delta})$ iterations, assuming $\mathcal{K}_{\mathsf{pos}}(\phi, \delta)$ does not hold. □

Similarly, we have the following lemma for the negative gradient.

Using the above corollary, we can now prove the following lemma:

Lemma D.9 (convergence of negative gradient)**.** *Let $j \in \tau(\mathcal{Y})$ and $\phi \in \mathfrak{F}_j$ and $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$. We have that*

(a) For all iterations $t \geq T_{\mathbf{u},1}$, it holds that $\Lambda_{5,j,r}^{(t)}(\mathbf{Z}) \mathbb{1}_{\mathcal{B}_{\widetilde{\phi}}} \in (-\frac{\varrho}{2}, \frac{\varrho}{2})$ for any $\widetilde{\phi} \in \mathfrak{F}_{\mathsf{conf}}(\phi)$.

(b) For any level $\delta \in (\frac{\boldsymbol{\lambda}}{d^{1.1}}, (d^{0.01}\sigma_0)^{q-2})$, there exists an iteration $t \geq T_{\mathbf{u},1} + O(\frac{n_y^3 \log^2 d}{\delta})$ such that $\mathcal{K}_{\mathsf{neg}}(\phi, \delta)$ holds.

Proof. The first claim is from the fact that for any $\psi_{j,r}^{(t)}(v)$ where $v \in \mathcal{G} \cup \mathcal{Y}$ such that they are either $y' \neq y$ or $g \notin \mathrm{Fiber}_{j,y}$, their positive gradient is upper bounded by $O(\sigma_0^{q-2})$ while lower bounded by $-\Theta(\varrho)$. Any such feature will be pushed to drop below $-\varrho$ after cancelling out (i.e., if δ is as small as $o(\frac{1}{d})$ or $o(\boldsymbol{\lambda}/d)$). Thus we only need to consider the negative gradient of $\psi_{j,r}^{(t)}(y'), y' \neq y$ and $\psi_{j,r}^{(t)}(g)$ for $g \notin \mathrm{Fiber}_{j,y}$. Then the logic is the same as in Lemma C.10: we consider when the growth of correct feature $\Psi_{\mathbf{u},\min}^{(t)}$ exceeds $B - O(d^{0.001}\sigma_0)$, then by applying Lemma C.10 and showing the contradiction that $\mathcal{K}_{\mathsf{neg}}(\phi, \delta)$ does not hold we shall have that for every incorrect pair $(g, y'), y' \neq y, g \in \mathrm{Fiber}_{j,y}$ or $(g', y), g' \notin \mathrm{Fiber}_{j,y}$, the feature $\psi_{j,r}^{(t)}(g')$ or $\psi_{j,r}^{(t)}(y')$ will grow below $-\psi_{j,r}^{(t)}(y)$ (or $\max_{g \in \mathrm{Fiber}_{j,y}} \psi_{j,r}^{(t)}(g)$) after $O(\frac{n_y^3 \log^2 d}{\eta\delta})$ iterations, when $\delta < \frac{1}{d}$ (if the corresponding logit is unsuppressed) or when $\delta < o(\boldsymbol{\lambda}/n_y d)$. □

By inserting $\delta = O((d^{0.01}\sigma_0)^{q-2} d/\boldsymbol{\lambda})^{1/(q-1)}$ into Lemma C.11, we have the a similar corollary to Corollary C.2:

Finally, we have the following proposition for the convergence of the feature.

Proposition D.5. Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$, then the following holds:

- At $t \geq T_{\mathbf{u},2} = T_{\mathbf{u},1} + O(\frac{1}{\eta d^{0.01}\sigma_0^{q-2}})$, we have $\Psi_{\mathbf{u},\min}^{(t)} \geq B - O(d^{0.01}\sigma_0)$.
- For any $\phi' \in \mathfrak{F}_{\mathsf{conf}}(\phi)$, we have both $|\Psi_{j,r,\min}^{(t)}(\phi')|$ and $|\Psi_{j,r,\max}^{(t)}(\phi')|$ bounded by $O((d^{0.01}\sigma_0)^{q-2} d/\boldsymbol{\lambda})^{1/(q-1)}$ if $t \geq T_{\mathbf{u},2}$.

D.3.1 Phase II: Shape of the Feature

A similar analysis as in Lemma C.12 gives the following lemma. We do not repeat the proof here.

Lemma D.10 (symmetry of ϕ in the end)**.** *Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^\star$ where $\phi = (\mathrm{Fiber}_{j,y}, y) \in \mathfrak{F}_j$. Suppose $\Psi_{\mathbf{u}}^{(t)} \geq B - O(d^{0.01}\sigma_0)$ and $\max_{\phi' \in \mathfrak{F}_{\mathsf{conf}}(\phi)} |\Psi_{j,r}^{(t)}(\phi')| \leq O(\delta)$ for all $t \geq T_{\mathbf{u},2}$, then we have*

- $\psi_{j,r}^{(t)}(g) + \psi_{j,r}^{(t)}(y) \geq B - O(d^{0.01}\sigma_0)$ *for all* $g \in \mathrm{Fiber}_{j,y}$.
- $\mathcal{J}_\phi^{(t)} := |n_y \psi_{j,r}^{(t)}(y) - \psi_{j,r}^{(t)}(g)| \leq O(\delta)$ *for all* $g \in \mathrm{Fiber}_{j,y}$ *and* $t \geq T_{\mathbf{u},2}$.

plugging in $\delta = O((d^{0.01}\sigma_0)^{q-2} d/\boldsymbol{\lambda})^{1/(q-1)}$ as in Proposition C.4, we have that $\mathcal{J}_\phi^{(t)} \leq O((d^{0.01}\sigma_0)^{q-2} d/\boldsymbol{\lambda})^{1/(q-1)}$.

Thus we arrive at the following theorem.

Theorem D.1 (learning symemtry group action)**.** *For group structure Assumption D.1, at iteration $t = \mathcal{T}_1$, the following properties of $F^{(t)}$ hold:*

(A) Optimal loss: The training loss is optimal for $i = 5$:

$$\mathsf{Loss}_5^1(F^{(t)}) \leq \frac{1}{\mathsf{poly}(d)}$$

(B) Sparse activations: For $j \in \tau(\mathcal{Y})$, let $\phi = (\mathrm{Fiber}_{j,y}, y) \in \mathfrak{F}_j$, then there exists exactly one activated *neuron $r \in [m]$ such that when $g_1 \in \mathrm{Fiber}_{j,y}, y_0 = y$ happens:*

$$\Lambda_{5,j,r}^{(\mathcal{T}_1)} \geq B - O(d^{0.01}\sigma_0) \quad \textit{while} \quad \Lambda_{5,j,r'}^{(\mathcal{T}_1)} \leq O(d^{-\Omega(1)}) \quad \forall r' \neq r$$

(C) Cancellation of incorrect features: For $j \in \tau(\mathcal{Y})$, let $\phi = (\mathrm{Fiber}_{j,y}, y) \in \mathfrak{F}_j$, and let the $r \in [m]$ be the activated neuron in (B), then for any $g' \notin \mathrm{Fiber}_{j,y}$, and any $y' \neq y \in \mathcal{Y}$, we have

$$|\psi_{j,r}^{(T_1)}(g) + \psi_{j,r}^{(T_1)}(y')| \leq O(\delta) \quad \textit{and} \quad |\psi_{j,r}^{(T_1)}(g') + \psi_{j,r}^{(T_1)}(y)| \leq O(\delta)$$

for some $\delta = O(((d^{0.01}\sigma_0)^{q-2} d/\boldsymbol{\lambda})^{1/(q-1)})$

E Auxiliary Technical Tools

E.1 Probability

First we need a Bernstein inequality for U-statistics

Lemma E.1 (concentration inequality for pseudo-U-statistics)**.** *Let $x_1, \ldots, x_n$ be different symbols, and let $m \ll n$ be such that $n \equiv 0 \pmod{n}$. Suppose for some function h with $|h| \leq M$ the random variables $h(x_{i_1}, x_{i_2}, \ldots, x_{i_m})$ and $h(x_{i'_1}, x_{i'_2}, \ldots, x_{i'_m})$ are independent and identically distributed as long as $\{x_{i_1}, x_{i_2}, \ldots, x_{i_m}\} \cap \{x_{i'_1}, x_{i'_2}, \ldots, x_{i'_m}\} = \emptyset$, then the pseudo-U-statistic*

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

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

Proof. The proof is the same as in [87]. □

E.2 Tensor Power Method Bounds

We present two lemmas related to the tensor power method.

Lemma E.2 (TPM, adapted from [73])**.** *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 $\eta \in (0, 1)$:*

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

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

This lemma has a corollary:

Corollary E.1 (TPM, from [73])**.** *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 $x_0 \geq y_0 S^{\frac{1}{q-2}}(1 + \frac{1}{\mathrm{polylog}(d)})$, letting T_x be the first iteration s.t., $x_t \geq A$, then

$$y_{T_x} \leq \widetilde{O}(y_0).$$

We also have a generalized version:

Corollary E.2. *Let $q \geq 3$ be a constant and $x_0, y_0, u_0, v_0 = o(1)$ and $A = O(1)$. Let $\{x_t, y_t\}_{t \geq 0}$ be two positive sequences updated as*

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

where $u_t = \Theta(x_t)$ *and* $v_t = \Theta(y_t)$ *for some constants* $c, d > 0$*. Suppose* $u_0 \geq v_0 S^{\frac{1}{q-2}}(1 + \frac{1}{\text{polylog}(d)})$*, letting* T_x *be the first iteration s.t.,* $x_t \geq A$*, then*

$$y_{T_x} \leq \widetilde{O}(y_0).$$

F 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 $\mathcal{Y} = \{0, 1, \ldots, n_y - 1\}$, where $n_y \in [\Omega(1), \log d]$.

We focus on updating only $\mathbf{Q}$, while keeping $\mathbf{W}$ fixed. Combined with the attention structure specified in Assumption A.2, 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 $\sum_{\ell=1}^{2} \mathsf{Loss}_5^{2,\ell}$. 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., $\psi_{j,r}$) for notational simplicity.

F.1 Gradient Computations

Notations for gradient expressions. We firsr introduce some notations for the gradients of the attention layer. For $1 \le \ell \le L$, given $\mathbf{Z}^{L,\ell-1}$ and $\mathbf{k} \in \mathcal{I}^{L,\ell-1}$, define

$$\Xi^L_{\ell,i,\mathbf{k}}(\mathbf{Z}^{L,\ell-1}) \triangleq \sum_{j\in[d]} \mathcal{E}_{i,j}(\mathbf{Z}^{L,\ell-1}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{i,j,r}(\mathbf{Z}^{L,\ell-1})\big)\langle \mathbf{W}_{i,j,r}, \mathbf{Z}_{\mathbf{k}}\rangle, \quad i \in [5]. \tag{35}$$

For simplicity of notation, we will henceforth omit the dependence on $\mathbf{Z}^{L,\ell-1}$ in the notation of $\Xi^L_{\ell,i,\mathbf{k}}$ when it is clear from the context.

Fact F.1 (Gradients of $\mathbf{Q}$)**.** For any $p, q \in [5]$, we have

$$-\nabla_{\mathbf{Q}_{p,q}} \mathsf{Loss}^L = \sum_{\ell=1}^{L} \sum_{i\in[5]} -\nabla_{\mathbf{Q}_{p,q}} \mathsf{Loss}_i^{L,\ell}, \quad \text{where}$$

$$-\nabla_{\mathbf{Q}_{p,q}} \mathsf{Loss}_i^{L,\ell} =$$

$$\mathbb{E}\left[\sum_{\mathbf{k}\in\mathcal{I}^{L,\ell-1}} \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathbf{k}} \cdot \left(\Xi^L_{\ell,i,\mathbf{k}} - \sum_{\mathbf{k}'\in\mathcal{I}^{L,\ell-1}} \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathbf{k}'} \Xi^L_{\ell,i,\mathbf{k}'}\right) \mathbf{Z}_{\mathsf{ans},\ell-1,p}\mathbf{Z}^\top_{\mathbf{k},q}\right].$$

Lemma F.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*

$$\Big[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}_5^{2,1}\Big]_{s,s} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,0}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot \Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},1}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Bigg)\mathbb{1}_{s=\tau(x_0)}\Bigg],$$

$$\Big[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}_5^{2,2}\Big]_{s,s} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot \Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},2}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Bigg)\mathbb{1}_{s=\tau(x_1)}\Bigg].$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\Big[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}_5^{2,1}\Big]_{s,s'} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},2} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,0}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot \Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},2}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Bigg)\mathbb{1}_{s=\tau(x_0),s'=\tau(x_1)}\Bigg],$$

$$\Big[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}_5^{2,2}\Big]_{s,s'} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot$$

$$\Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},1}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Big)\mathbb{1}_{s=\tau(x_1),s'=\tau(x_0)}\Big].$$

Proof. For $\ell = 1$, due to Fact F.1, the diagonal entry $[\mathbf{Q}_{4,3}]_{s,s}$ with $s \in \tau(\mathcal{X})$, the expected gradient contribution from $-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}_5^{2,1}$ takes the form

$$\mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1} \cdot \Big(\Xi^2_{\ell,5,\mathsf{pred},1} - \sum_{\mathbf{k}'\in\mathcal{I}^{2,\ell-1}} \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathbf{k}'}\, \Xi^2_{\ell,5,\mathbf{k}'}\Big) \cdot \mathbb{1}_{s=\tau(x_0)}\Big],$$

which is nonzero in expectation only when $s = \tau(x_0)$. Therefore, combined the definition of Ξ in (35) we have:

$$\begin{aligned}
&\Big[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}_5^{2,1}\Big]_{s,s} \\
&= \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1} \cdot \Big(\Xi^2_{\ell,5,\mathsf{pred},1} - \sum_{\mathbf{k}'\in\mathcal{I}^{2,\ell-1}} \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathbf{k}'} \Xi^2_{\ell,5,\mathbf{k}'}\Big)\mathbb{1}_{s=\tau(x_0)}\Bigg] \\
&= \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j} \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot \\
&\qquad \Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},1}\rangle - \sum_{\mathbf{k}'\in\mathcal{I}^{2,\ell-1}} \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathbf{k}'} \langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathbf{k}'}\rangle\Big)\Bigg)\mathbb{1}_{s=\tau(x_0)}\Bigg] \\
&= \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j} \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot \\
&\qquad \Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},1}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Bigg)\mathbb{1}_{s=\tau(x_0)}\Bigg].
\end{aligned}$$

Other quantities are computed similarly, and thus is omitted here. □

Lemma F.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*

$$\begin{aligned}
\Big[-\nabla_{\mathbf{Q}_{4,4}} \mathsf{Loss}_5^{2,1}\Big]_{s,s} &= \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{ans},0} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,0}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot \\
&\qquad \Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{ans},0}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Bigg)\mathbb{1}_{s=\tau(x_0)}\Bigg], \\
\Big[-\nabla_{\mathbf{Q}_{4,4}} \mathsf{Loss}_5^{2,2}\Big]_{s,s} &= \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot \\
&\qquad \Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{ans},1}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Bigg)\mathbb{1}_{s=\tau(x_1)}\Bigg].
\end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,4}]_{s,s'}$ with $s \neq s'$, we have $[-\nabla_{\mathbf{Q}_{4,4}} \mathsf{Loss}_5^{2,1}]_{s,s'} = 0$, and

$$\Big[-\nabla_{\mathbf{Q}_{4,4}} \mathsf{Loss}_5^{2,2}\Big]_{s,s'} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot \Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{ans},0}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Bigg)\mathbb{1}_{s=\tau(x_1),s'=\tau(x_0)}\Bigg].$$

Proof. The analysis is similar to that of Lemma F.1, and we omit the details here. □

F.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 C.12 and Theorem C.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:

$$\Lambda_{5,j,r}^{(T_1)} \geq B - O(d^{0.01}\sigma_0), \quad |\psi_{j,r}^{(T_1)}(g) - \psi_{j,r}^{(T_1)}(y)| \leq O(\delta)$$
$$\Lambda_{5,j,r'}^{(T_1)} \leq O(\delta^{q-1}) \quad \text{for all } r' \neq r,$$

  where $\delta = O\left(\left(\frac{(d^{0.01}\sigma_0)^{q-2}d}{\lambda}\right)^{\frac{1}{q-1}}\right)$.

- **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:

$$\left|\psi_{j,r}^{(T_1)}(g) + \psi_{j,r}^{(T_1)}(y')\right| \leq O(\delta),$$
$$\left|\psi_{j,r}^{(T_1)}(g') + \psi_{j,r}^{(T_1)}(y)\right| \leq O(\delta),$$

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 $r_{g\cdot y}$. Moreover, let

$$\mathfrak{A} \triangleq \cup_{j\in\tau(\mathcal{Y})}\mathfrak{A}_j, \text{ where } \mathfrak{A}_j \triangleq \{r \mid \exists (g, y) \in \mathfrak{F}_j, r = r_{g\cdot y}\}.$$

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 cluases. Similarly $\widehat{\mathcal{Y}} = \cup\{y_{\ell'}\}_{\ell'=0}^{\ell-1}$. Then define $\widehat{\mathfrak{A}_j}(\mathbf{Z}^{L,\ell-1}) = \{r_{g\cdot y} \mid (g, y) \in \mathfrak{F}_j \wedge \big(g \in \widehat{\mathcal{G}}(\mathbf{Z}^{L,\ell-1}) \vee y \in \widehat{\mathcal{Y}}\big)\}$. 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 F.3 (Properties of target feature magnitude)**.** *Given $(g, y) \in \mathfrak{F}_j$ with $j \in \tau(\mathcal{Y})$, then, the following properties hold.*

$$\frac{1}{2}\left(\psi_{j,r_{g\cdot y}}(g) + \psi_{j,r_{g\cdot y}}(y)\right) \geq B - O(d^{0.01}\sigma_0), \quad \left|\psi_{j,r_{g\cdot y}}(g) - \psi_{j,r_{g\cdot y}}(y)\right| \leq O(\delta), \tag{36}$$
$$\left|\psi_{j,r_{g\cdot y}}(g) + \psi_{j,r_{g\cdot y}}(y')\right| \leq O(\delta), \psi_{j,r_{g\cdot y}}(y') < 0 \quad \textit{for all } y' \neq y, \tag{37}$$
$$\left|\psi_{j,r_{g\cdot y}}(g') + \psi_{j,r_{g\cdot y}}(y)\right| \leq O(\delta), \psi_{j,r_{g\cdot y}}(g') < 0 \quad \textit{for all } g' \neq g. \tag{38}$$
$$|\psi_{j,r}(g)|, |\psi_{j,r}(y)| \leq O(\delta) \quad \textit{for all } r \notin \mathfrak{A}_j. \tag{39}$$

Lemma F.4 (Properties of irrelavant 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*

$$\left|\langle \mathbf{W}_{5,j,r,p}, e_v\rangle\right| \leq \widetilde{O}(\sigma_0). \tag{40}$$

The above lemmas give us some direct computations of the inner products between the weight matrices and input embedding vectors.

Lemma F.5. *Let $j \in \tau(\mathcal{Y})$ and $\ell \in [2]$. Then for any $r \in [m]$, the following holds:*

$$\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},\ell}\rangle = \Psi_{j,r}(g_\ell) \pm \widetilde{O}(\sigma_0), \tag{41}$$
$$\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{ans},\ell-1}\rangle = \Psi_{j,r}(y_{\ell-1}) \pm \widetilde{O}(\sigma_0). \tag{42}$$

Moreover, for $j \notin \tau(\mathcal{Y})$ and any $\mathbf{k} \in \mathcal{I}^{2,1}$ and $r \in [m]$, we have

$$\left|\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathbf{k}}\rangle\right| = \widetilde{O}(\sigma_0). \tag{43}$$

Proof. By direct computations and the definition of $\mathbf{Z}_{\mathsf{pred},\ell}$, we have

$$\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},\ell} \rangle = \langle \mathbf{W}_{5,j,r,1}, e_{x_\ell} \rangle + \langle \mathbf{W}_{5,j,r,2}, e_{g_\ell} \rangle + \langle \mathbf{W}_{5,j,r,3}, e_{x_{\ell-1}} \rangle$$

Plug in Lemma F.4 and the definition of $\psi_{j,r}(g_\ell)$, we obtain (41). The proof of (42) and (43) is similar, and we omit the details here. □

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 F.6 (Characterizations of Lambda)**.** *Given $\mathbf{Z}^{2,\ell-1}$ with $\ell \in [2]$, with an attention structure $\{\mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathbf{k}}\}_{\mathbf{k}\in\mathcal{I}^{2,\ell-1}}$,*

(a) for $j \in \tau(\mathcal{Y})$, for activated neuron $r \in \mathfrak{A}_j$, we have

$$\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1}) = \sum_{\ell'=1}^{2} \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{pred},\ell'} \psi_{j,r}(g_{\ell'}) + \sum_{\ell'=1}^{\ell} \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{ans},\ell'-1} \psi_{j,r}(y_{\ell'-1}) \pm \widetilde{O}(\sigma_0).$$

(b) for $j \in \tau(\mathcal{Y})$, for any non-activated neuron $r \notin \mathfrak{A}_j$ we have

$$\left|\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1})\right| \le O(\delta).$$

(c) for $j \notin \tau(\mathcal{Y})$, for any $r \in [m]$, we have

$$\left|\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1})\right| \le \widetilde{O}(\sigma_0).$$

Proof. Recall the definition of $\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1})$ in (13), we have

$$\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1}) = \sum_{\mathbf{k}\in\mathcal{I}^{2,\ell-1}} \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathbf{k}} \cdot \langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathbf{k}} \rangle + b_{5,j,r}.$$

Thus, the first part follows directly from (41) and (42); similarly, the second part holds by plugging (39) into (41) and (42); the last part is a direct consequence of (43) and the fact $b_{i,j,r} = \sigma_0 \log d$.

□

A direct consequence of the above lemma is the following finer characterization of the activated neurons.

Lemma F.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 F.6, we have

$$\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1}) = \sum_{\ell'=1}^{2} \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{pred},\ell'} \psi_{j,r}(g_{\ell'}) + \sum_{\ell'=1}^{\ell} \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{ans},\ell'-1} \psi_{j,r}(y_{\ell'-1}) \pm \widetilde{O}(\sigma_0).$$

By (39), for for $r \in \mathfrak{A}_j \setminus \widehat{\mathfrak{A}}_j$, we have $\psi_{j,r}(g_{\ell'}), \psi_{j,r}(y_{\ell'-1}) \le -B - O(\delta)$. Hence

$$\Lambda_{5,j,r} = -\Omega(B) \pm \widetilde{O}(\sigma_0) \ll -\varrho,$$

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 F.1 and F.2 and the properties established above.

Lemma F.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*

$$\left[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}_5^{2,1} \right]_{s,s} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1} \cdot$$

$$\Big((1-\mathbf{logit}_{5,j_1})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_1}}\mathbf{sReLU}'(\Lambda_{5,j_1,r})\cdot\Big(\psi_{j_1,r}(g_1)-\Lambda_{5,j_1,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$-\sum_{j\neq j_1\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(g_1)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$\pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_0)=s}\Big]$$

$$\Big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}_5^{2,2}\Big]_{s,s}=\mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot$$
$$\Big((1-\mathbf{logit}_{5,j_2})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot\Big(\psi_{j_2,r}(g_2)-\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$-\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(g_2)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$\pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_1)=s}\Big]$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s\neq s'$, we have

$$\Big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}_5^{2,1}\Big]_{s,s'}=\mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},2}\cdot$$
$$\Big((1-\mathbf{logit}_{5,j_1})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_1}}\mathbf{sReLU}'(\Lambda_{5,j_1,r})\cdot\Big(\psi_{j_1,r}(g_2)-\Lambda_{5,j_1,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$-\sum_{j\neq j_1\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(g_2)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$\pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_0)=s,\tau(x_1)=s'}\Big]$$

$$\Big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}_5^{2,2}\Big]_{s,s'}=\mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot$$
$$\Big((1-\mathbf{logit}_{5,j_2})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot\Big(\psi_{j_2,r}(g_1)-\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$-\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(g_1)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$\pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_1)=s,\tau(x_0)=s'}\Big]$$

Proof. For $\ell=1$,

- for the diagonal entry $[\mathbf{Q}_{4,3}]_{s,s}$ with $s\in\tau(\mathcal{X})$,

$$\Big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}_5^{2,1}\Big]_{s,s}=\mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}\cdot\Big(\sum_{j\in[d]}\mathcal{E}_{5,j}(\mathbf{Z}^{2,0})\sum_{r\in[m]}\mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot$$

$$\Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},1}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Big)\mathbb{1}_{s=\tau(x_0)}\Big]$$

$$= \mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}\cdot\Big((1-\mathbf{logit}_{5,j_1})\cdot\sum_{r\in[m]}\mathbf{sReLU}'(\Lambda_{5,j_1,r})\Big(\psi_{j_1,r}(g_1)-\Lambda_{5,j_1,r}\pm\widetilde{O}(\sigma_0)\Big)$$
$$-\sum_{j\neq j_1\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\sum_{r\in[m]}\mathbf{sReLU}'(\Lambda_{5,j,r})\Big(\psi_{j,r}(g_1)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)$$
$$-\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\sum_{r\in[m]}\mathbf{sReLU}'(\Lambda_{5,j,r})\Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},1}\rangle-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\Big)\mathbb{1}_{\tau(x_0)=s}\Big]$$

$$= \mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}\cdot$$
$$\Big((1-\mathbf{logit}_{5,j_1})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_1}}\mathbf{sReLU}'(\Lambda_{5,j_1,r})\cdot\Big(\psi_{j_1,r}(g_1)-\Lambda_{5,j_1,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$-\sum_{j\neq j_1\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j}}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(g_1)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$\pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_0)=s}\Big]$$

where the last equality follows from Lemma F.6 and Lemma F.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

$$\Big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}_5^{2,1}\Big]_{s,s'} = \mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},2}\cdot\Big(\sum_{j\in[d]}\mathcal{E}_{5,j}(\mathbf{Z}^{2,0})\sum_{r\in[m]}\mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot$$
$$\Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},2}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Big)\mathbb{1}_{s=\tau(x_0),s'=\tau(x_1)}\Big]$$

$$= \mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},2}\cdot$$
$$\Big((1-\mathbf{logit}_{5,j_1})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_1}}\mathbf{sReLU}'(\Lambda_{5,j_1,r})\cdot\Big(\psi_{j_1,r}(g_2)-\Lambda_{5,j_1,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$-\sum_{j\neq j_1\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j}}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(g_2)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$\pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_0)=s,\tau(x_1)=s'}\Big]$$

For $\ell=2$,

- for the diagonal entry $[\mathbf{Q}_{4,3}]_{s,s}$ with $s\in\tau(\mathcal{X})$,

$$\Big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}_5^{2,2}\Big]_{s,s} = \mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\Big(\sum_{j\in[d]}\mathcal{E}_{5,j}(\mathbf{Z}^{2,1})\sum_{r\in[m]}\mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot$$
$$\Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},2}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Big)\mathbb{1}_{s=\tau(x_1)}\Big]$$

$$= \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\Bigg((1-\mathbf{logit}_{5,j_2})\cdot\sum_{r\in[m]}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\Big(\psi_{j_2,r}(g_2)-\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)$$
$$-\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\sum_{r\in[m]}\mathbf{sReLU}'(\Lambda_{5,j,r})\Big(\psi_{j,r}(g_2)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)$$
$$-\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\sum_{r\in[m]}\mathbf{sReLU}'(\Lambda_{5,j,r})\Big(\langle\mathbf{W}_{5,j,r},\mathbf{Z}_{\mathsf{pred},2}\rangle-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\Bigg)\mathbb{1}_{\tau(x_1)=s}\Bigg]$$

$$= \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot$$
$$\Big((1-\mathbf{logit}_{5,j_2})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot\Big(\psi_{j_2,r}(g_2)-\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$-\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(g_2)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$\pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_1)=s}\Bigg]$$

where the last equality follows from Lemma F.6 and Lemma F.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

$$\Big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}_5^{2,2}\Big]_{s,s'}=\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\Big(\sum_{j\in[d]}\mathcal{E}_{5,j}(\mathbf{Z}^{2,1})\sum_{r\in[m]}\mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot$$
$$\Big(\langle\mathbf{W}_{5,j,r},\mathbf{Z}_{\mathsf{pred},1}\rangle-\Lambda_{5,j,r}+b_{5,j,r}\Big)\Big)\mathbb{1}_{s=\tau(x_1),s'=\tau(x_0)}\Bigg]$$

$$= \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot$$
$$\Big((1-\mathbf{logit}_{5,j_2})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot\Big(\psi_{j_2,r}(g_1)-\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$-\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(g_1)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$\pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_1)=s,\tau(x_0)=s'}\Bigg]$$

Therefore, we complete the proof. □

Lemma F.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*

$$\Big[-\nabla_{\mathbf{Q}_{4,4}}\mathsf{Loss}_5^{2,1}\Big]_{s,s}=\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{ans},0}\cdot$$
$$\Big((1-\mathbf{logit}_{5,j_1})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_1}}\mathbf{sReLU}'(\Lambda_{5,j_1,r})\cdot\Big(\psi_{j_1,r}(y_0)-\Lambda_{5,j_1,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$-\sum_{j\neq j_1\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(y_0)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$

$$\pm \sum_{j\notin\tau(\mathcal{Y})} \mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_0)=s}\Big].$$

$$\Big[-\nabla_{\mathbf{Q}_{4,4}}\mathsf{Loss}_5^{2,2}\Big]_{s,s} = \mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot$$
$$\Big((1-\mathbf{logit}_{5,j_2})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot\Big(\psi_{j_2,r}(y_1)-\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$-\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(y_1)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$\pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_1)=s}\Big].$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,4}]_{s,s'}$ *with* $s\neq s'$*, we have* $[-\nabla_{\mathbf{Q}_{4,4}}\mathsf{Loss}_5^{2,1}]_{s,s'}=0$*, and*

$$\Big[-\nabla_{\mathbf{Q}_{4,4}}\mathsf{Loss}_5^{2,2}\Big]_{s,s'} = \mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot$$
$$\Big((1-\mathbf{logit}_{5,j_2})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot\Big(\psi_{j_2,r}(y_0)-\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$-\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(y_0)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$\pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_1)=s,\tau(x_0)=s'}\Big].$$

The proof is similar to Lemma F.8 and we omit it here.

Notations for gradient decompositions. We shall define some useful notations to further simplify the expressions of gradient.

- for $\ell=1$,
  - for $[\mathbf{Q}_{4,3}]_{s,s}$ with $s\in\tau(\mathcal{X})$, we have $[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}_5^{2,1}]_{s,s}=\mathcal{N}_{s,3,1,i}+\mathcal{N}_{s,3,1,ii}+\mathcal{N}_{s,3,1,iii}$, where

$$\mathcal{N}_{s,3,1,i} = \mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}\cdot(1-\mathbf{logit}_{5,j_1})\cdot \tag{44}$$
$$\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_1}}\mathbf{sReLU}'(\Lambda_{5,j_1,r})\cdot\Big(\psi_{j_1,r}(g_1)-\Lambda_{5,j_1,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_0)=s}\Big]$$

$$\mathcal{N}_{s,3,1,ii} = -\mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}\cdot\sum_{j\neq j_1\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot \tag{45}$$
$$\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(g_1)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_0)=s}\Big]$$

$$\mathcal{N}_{s,3,1,iii} = \pm\mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}\cdot\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\mathbb{1}_{\tau(x_0)=s}\Big] \tag{46}$$

– for $[\mathbf{Q}_{4,4}]_{s,s}$ with $s \in \tau(\mathcal{X})$, we have $[-\nabla_{\mathbf{Q}_{4,4}} \mathsf{Loss}_5^{2,1}]_{s,s} = \mathcal{N}_{s,4,1,i} + \mathcal{N}_{s,4,1,ii} + \mathcal{N}_{s,4,1,iii}$, where

$$\mathcal{N}_{s,4,1,i} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{ans},0} \cdot (1 - \mathbf{logit}_{5,j_1}) \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_1}} \mathbf{sReLU}'(\Lambda_{5,j_1,r}) \cdot \Big(\psi_{j_1,r}(y_0) - \Lambda_{5,j_1,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big) \mathbb{1}_{\tau(x_0)=s}\Bigg] \tag{47}$$

$$\mathcal{N}_{s,4,1,ii} = -\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{ans},0} \cdot \sum_{j\neq j_1\in\tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \Big(\psi_{j,r}(y_0) - \Lambda_{5,j,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big) \mathbb{1}_{\tau(x_0)=s}\Bigg] \tag{48}$$

$$\mathcal{N}_{s,4,1,iii} = \pm\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{ans},0} \cdot \sum_{j\notin\tau(\mathcal{Y})} \mathbf{logit}_{5,j} \widetilde{O}(\sigma_0^q) \mathbb{1}_{\tau(x_0)=s}\Bigg] \tag{49}$$

- for $\ell = 2$,
  – for $[\mathbf{Q}_{4,3}]_{s,s}$ with $s \in \tau(\mathcal{X})$, we have $[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}_5^{2,2}]_{s,s} = \mathcal{N}_{s,3,2,i} + \mathcal{N}_{s,3,2,ii} + \mathcal{N}_{s,3,2,iii}$, where

$$\mathcal{N}_{s,3,2,i} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2} \cdot (1 - \mathbf{logit}_{5,j_2}) \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \Big(\psi_{j_2,r}(g_2) - \Lambda_{5,j_2,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big) \mathbb{1}_{\tau(x_1)=s}\Bigg] \tag{50}$$

$$\mathcal{N}_{s,3,2,ii} = -\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2} \cdot \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \Big(\psi_{j,r}(g_2) - \Lambda_{5,j,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big) \mathbb{1}_{\tau(x_1)=s}\Bigg] \tag{51}$$

$$\mathcal{N}_{s,3,2,iii} = \pm\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2} \cdot \sum_{j\notin\tau(\mathcal{Y})} \mathbf{logit}_{5,j} \widetilde{O}(\sigma_0^q) \mathbb{1}_{\tau(x_1)=s}\Bigg] \tag{52}$$

  – for $[\mathbf{Q}_{4,4}]_{s,s}$ with $s \in \tau(\mathcal{X})$, we have $[-\nabla_{\mathbf{Q}_{4,4}} \mathsf{Loss}_5^{2,2}]_{s,s} = \mathcal{N}_{s,4,2,i} + \mathcal{N}_{s,4,2,ii} + \mathcal{N}_{s,4,2,iii}$, where

$$\mathcal{N}_{s,4,2,i} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1} \cdot (1 - \mathbf{logit}_{5,j_2}) \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \Big(\psi_{j_2,r}(y_1) - \Lambda_{5,j_2,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big) \mathbb{1}_{\tau(x_1)=s}\Bigg] \tag{53}$$

$$\mathcal{N}_{s,4,2,ii} = -\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1} \cdot \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \Big(\psi_{j,r}(y_1) - \Lambda_{5,j,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big) \mathbb{1}_{\tau(x_1)=s}\Bigg] \tag{54}$$

$$\mathcal{N}_{s,4,2,iii} = \pm\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1} \cdot \sum_{j\notin\tau(\mathcal{Y})} \mathbf{logit}_{5,j} \widetilde{O}(\sigma_0^q) \mathbb{1}_{\tau(x_1)=s}\Bigg] \tag{55}$$

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.

$$\mathcal{E}_1 \triangleq \Big\{ g_1 \neq g_2 \Big\}, \tag{56}$$

$$\mathcal{E}_2 \triangleq \Big\{ g_{\ell_1}(y_{\ell_1'}) \neq g_{\ell_2}(y_{\ell_2'}),\ \text{for any } (\ell_1,\ell_1') \neq (\ell_2,\ell_2'),\ \text{where } \ell_k \in [2], \ell_k' \in \{0,1\} \Big\}. \tag{57}$$

It is easy to see that $\mathcal{E}_1$ and $\mathcal{E}_2$ hold with high probability $1 - O\big(\frac{1}{\log d}\big)$.

F.3 Stage 2.1: Initial Growth of Gap

At the begining 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

$$\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1}) \approx \frac{1}{3}\psi_{j,r}(g_1) + \frac{1}{3}\psi_{j,r}(g_2) + \frac{1}{3}\psi_{j,r}(y_0) \pm \widetilde{O}(\sigma_0).$$

If $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_1$,

- for $j = j_1$, for $r \in \widehat{\mathfrak{A}}_{j_1}$, 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 $j = j_1' \triangleq \tau\big(g_2(y_0)\big)$, 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 $r \in \widehat{\mathfrak{A}}_j$, thus no activation.

Morover, for $\ell = 2$, we have

$$\Lambda_{5,j,r}(\mathbf{Z}^{2,\ell-1}) \approx \frac{1}{4}\psi_{j,r}(g_1) + \frac{1}{4}\psi_{j,r}(g_2) + \frac{1}{4}\psi_{j,r}(y_0) + \frac{1}{4}\psi_{j,r}(y_1) \pm \widetilde{O}(\sigma_0).$$

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 correspnding $r_{g_\ell\cdot y_{\ell'}}$ 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 $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_\ell\cdot y_{\ell'}}\}$,
- for other $j \in \tau(\mathcal{Y})$, we have $\Lambda_{5,j,r} \leq -\frac{1}{2}B$ for all $r \in \widehat{\mathfrak{A}}_j$, 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, initally, the learning process is domintaed by $\nabla_{\mathbf{Q}}\mathsf{Loss}_5^{2,1}$. Moreover, if we take a closer look at the gradient from $\ell = 1$, we have

$$\mathcal{N}_{s,3,1,ii} \approx -\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}\cdot\mathbf{logit}_{5,j_1'}\cdot \Big(\mathbf{sReLU}'(\Lambda_{5,j_1',r_{g_2\cdot y_0}})\cdot\Big(\psi_{j_1',r_{g_2\cdot y_0}}(g_1) - \Lambda_{5,j_1',r_{g_2\cdot y_0}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_0)=s}\Bigg] \geq 0$$

$$\mathcal{N}_{s,4,1,ii} \approx -\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{ans},0}\cdot\mathbf{logit}_{5,j_1'}\cdot \Big(\mathbf{sReLU}'(\Lambda_{5,j_1',r_{g_2\cdot y_0}})\cdot\Big(\psi_{j_1',r_{g_2\cdot y_0}}(y_0) - \Lambda_{5,j_1',r_{g_2\cdot y_0}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\Bigg] \leq 0$$

since $\psi_{j_1',r_{g_2\cdot y_0}}(y_0) \geq \Omega(B)$ while $\psi_{j_1',r_{g_2\cdot y_0}}(g_1) \leq -\Omega(B)$. 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 subsatge, 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 substage. 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 F.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 $\Omega\Big(\frac{1}{\log d}\Big)$. For all iterations $t < T_{2,1,s}$, we have the following holds*

(a) $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$ *monotonically increases;*

(b) $\big|[\mathbf{Q}_{4,4}^{(t)}]_{s,s}\big| \le [\mathbf{Q}_{4,3}^{(t)}]_{s,s}$ *and* $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} - [\mathbf{Q}_{4,4}^{(t)}]_{s,s} = \Theta\Big([\mathbf{Q}_{4,3}^{(t)}]_{s,s}\Big)$;

(c) for $p \in \{3,4\}$, for $s' \in \tau(\mathcal{X}) \neq s$, $|[\mathbf{Q}_{4,p}^{(t)}]_{s,s'}| \le O\Big(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,s}}{d}\Big)$.

F.3.1 Attention and Logit Preliminaries

We first introduce several properties of the attention scores and logits if Induction F.1 holds.

Lemma F.10. *If Induction F.1 holds for all iterations $< t$, given input $\mathbf{Z}^{2,\ell-1}$, then we have*

1. *for $\ell = 1$,*
   - *(a)* $\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}^{(t)} \in \Big[\frac{1}{3}, \frac{1}{3} + O\big(\frac{1}{\log d}\big)\Big]$;
   - *(b)* $\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},2}^{(t)}, \mathbf{Attn}_{\mathsf{ans},0\to\mathsf{ans},0}^{(t)} \le \mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}^{(t)}$;
   - *(c)* $\big|\mathbf{Attn}_{\mathsf{ans},0\to\mathbf{k}}^{(t)} - \mathbf{Attn}_{\mathsf{ans},0\to\mathbf{k}'}^{(t)}\big| \le O(\frac{1}{\log d})$ *for* $\mathbf{k} \neq \mathbf{k}' \in \mathcal{I}^{(2,0)}$.
2. *for $\ell = 2$,*
   - *(a)* $\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} \in \Big[\frac{1}{4}, \frac{1}{4} + O\big(\frac{1}{\log d}\big)\Big]$;
   - *(b)* $\mathbf{Attn}_{\mathsf{ans},1\to\mathbf{k}}^{(t)} \le \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}$ *for* $\mathbf{k} \neq (\mathsf{pred}, 2)$;
   - *(c) for* $\mathbf{k} \in \big\{(\mathsf{ans},0),(\mathsf{pred},1)\big\}$, $\mathbf{k}' \in \big\{(\mathsf{ans},1),(\mathsf{pred},2)\big\}$,
   $$\big|\mathbf{Attn}_{\mathsf{ans},1\to\mathbf{k}}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathbf{k}'}^{(t)}\big| \le O\Big(\frac{1}{\log d}\Big);$$
   - *(d)* $\big|\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\big| \le O\big(\frac{1}{d}\big)$.

Proof. For $\ell = 1$, given $\mathbf{Z}^{2,\ell-1}$, according to Assumption A.2, we have

$$\begin{aligned}\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}^{(t)} &= \frac{e^{[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0),\tau(x_0)}}}{e^{[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0),\tau(x_0)}} + e^{[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0),\tau(x_1)}} + e^{[\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_0),\tau(x_0)}}} \\ &= \frac{1}{1 + e^{[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0),\tau(x_1)} - [\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0),\tau(x_0)}} + e^{[\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_0),\tau(x_0)} - [\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0),\tau(x_0)}}}\end{aligned}$$

Thus, by Induction F.1,

$$-O\Big(\frac{1}{\log d}\Big) \le [\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0),\tau(x_1)} - [\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0),\tau(x_0)}, [\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_0),\tau(x_0)} - [\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0),\tau(x_0)} \le 0,$$

which implies that $0 \le \mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}^{(t)} \le \frac{1}{3} + O\big(\frac{1}{\log d}\big)$. (b) is straightforward since

$$[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0),\tau(x_1)}, [\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_0),\tau(x_0)} \le [\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_0),\tau(x_0)}.$$

(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

$$\begin{aligned}&\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)} \\ &= \frac{e^{[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_1),\tau(x_0)}} - e^{[\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_1),\tau(x_0)}}}{e^{[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_1),\tau(x_0)}} + e^{[\mathbf{Q}_{4,3}^{(t)}]_{\tau(x_1),\tau(x_1)}} + e^{[\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_1),\tau(x_0)}} + e^{[\mathbf{Q}_{4,4}^{(t)}]_{\tau(x_1),\tau(x_1)}}}\end{aligned}$$

$$\overset{(i)}{\leq} O\Big([\mathbf{Q}^{(t)}_{4,3}]_{\tau(x_1),\tau(x_0)} - [\mathbf{Q}^{(t)}_{4,4}]_{\tau(x_1),\tau(x_0)}\Big) \overset{(ii)}{\leq} O(\frac{1}{d}),$$

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 F.1 (b). □

Lemma F.11. *If Induction F.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 $j = j_1$, $\Lambda^{(t)}_{5,j_1,r} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_{j_1} \setminus \{r_{g_1 \cdot y_0}\}$;*
   (b) *for $j = j_1' \triangleq \tau\big(g_2(y_0)\big)$, $\Lambda^{(t)}_{5,j_1',r} \ll -\varrho$ 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 activted for all $r \in \widehat{\mathfrak{A}}_j$, i.e., $\Lambda^{(t)}_{5,j,r} \ll -\varrho$.*
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 correspnding $r_{g_\ell \cdot y_{\ell'}}$ may be activated, with $|\Lambda^{(t)}_{5,j,r_{g_\ell \cdot y_{\ell'}}}| \leq O(1)$, while all other $\Lambda^{(t)}_{5,j,r} \ll -\varrho$ 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 activted for all $r \in \widehat{\mathfrak{A}}_j$, i.e., $\Lambda^{(t)}_{5,j,r} \ll -\varrho$.*

Proof. We only prove (a) for $\ell = 2$ since the other cases are straightforward. $j = \tau(g_\ell(y_{\ell'}))$, by Lemma F.6 we have

$$\begin{aligned}\Lambda_{5,j,r_{g_\ell \cdot y_{\ell'}}} &= \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\psi_{j,r_{g_\ell \cdot y_{\ell'}}}(g_1) + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\psi_{j,r_{g_\ell \cdot y_{\ell'}}}(g_2) \\ &\quad + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\psi_{j,r_{g_\ell \cdot y_{\ell'}}}(y_0) + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\psi_{j,r_{g_\ell \cdot y_{\ell'}}}(y_1) \pm \widetilde{O}(\sigma_0).\end{aligned}$$

Notice that since $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_2$, we have two ψ terms are postive and two are negative with magnitude $B \pm O(\delta)$. Therefore, $|\Lambda_{5,j,r_{g_\ell \cdot y_{\ell'}}}| \leq O\Big(\frac{1}{\log d}\Big) \cdot B = O(1)$. □

Lemma F.12. *If Induction F.1 holds for all iterations $< t$, given input $\mathbf{Z}^{2,0} \in \mathcal{E}_1$, then we have*

(a) $\Lambda^{(t)}_{5,j_1,r_{g_1 \cdot y_0}} \geq \frac{1}{3}B - O(1)$;

(b) $-O(\delta) \leq \Lambda^{(t)}_{5,j_1,r_{g_1 \cdot y_0}} - \Lambda^{(t)}_{5,j_1',r_{g_2 \cdot y_0}} \leq O(1)$;

Proof. By Lemma F.6 we have

$$\begin{aligned}&\Lambda_{5,j_1,r_{g_1 \cdot y_0}} = \\ &\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1}\psi_{j_1,r_{g_1 \cdot y_0}}(g_1) + \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},2}\psi_{j_1,r_{g_1 \cdot y_0}}(g_2) + \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0}\psi_{j,r_{g_1 \cdot y_0}}(y_0) \pm \widetilde{O}(\sigma_0).\end{aligned}$$

By Lemma F.10 and the cancellation in (38), we have

$$\begin{aligned}&\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},2}\psi_{j_1,r_{g_1 \cdot y_0}}(g_2) + \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0}\psi_{j_1,r_{g_1 \cdot y_0}}(y_0) \\ &\quad \geq \Big(\frac{1}{3} - O\Big(\frac{1}{\log d}\Big)\Big) \cdot O(\delta) - O\Big(\frac{1}{\log d}\Big) \cdot B.\end{aligned}$$

Putting it back, and using the fact that $\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1}\psi_{j_1,r_{g_1 \cdot y_0}}(g_1) \geq \frac{1}{3} \cdot (B - O(\delta))$, we have

$$\begin{aligned}\Lambda_{5,j_1,r_{g_1 \cdot y_0}} &\geq \frac{1}{3} \cdot (B - O(\delta)) + \Big(\frac{1}{3} - O\Big(\frac{1}{\log d}\Big)\Big) \cdot O(\delta) - O\Big(\frac{1}{\log d}\Big) \cdot B \pm \widetilde{O}(\sigma_0) \\ &\geq \frac{1}{3}B - O(1).\end{aligned}$$

Moving on to (b), we have

$$\Lambda_{5,j_1,r_{g_1 \cdot y_0}} - \Lambda_{5,j_1',r_{g_2 \cdot y_0}} = \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1}(\psi_{j_1,r_{g_1 \cdot y_0}}(g_1) - \psi_{j_1',r_{g_2 \cdot y_0}}(g_1)) \pm \widetilde{O}(\sigma_0)$$

$$
\begin{aligned}
&+ \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},2}(\psi_{j_1,r_{g_1\cdot y_0}}(g_2) - \psi_{j_1',r_{g_2\cdot y_0}}(g_2)) + \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0}(\psi_{j_1,r_{g_1\cdot y_0}}(y_0) - \psi_{j_1',r_{g_2\cdot y_0}}(y_0)) \\
&\le \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1}\cdot(2B + O(\delta)) - \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},2}\cdot(2B - O(\delta)) + \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0}\cdot O(\delta) \\
&\le (2B - O(\delta))\cdot O\Big(\frac{1}{\log d}\Big) + O(\delta) \le O(1).
\end{aligned}
$$

Similarly, we have

$$
\begin{aligned}
&\Lambda_{5,j_1,r_{g_1\cdot y_0}} - \Lambda_{5,j_1',r_{g_2\cdot y_0}} \\
&\ge \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1}\cdot(2B - O(\delta)) - \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},2}\cdot(2B + O(\delta)) - \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0}\cdot O(\delta) \\
&\ge -\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},2}\cdot O(\delta) - \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0}\cdot O(\delta) \ge -O(\delta).
\end{aligned}
$$

□

Lemma F.13. *If Induction F.1 holds for all iterations $< t$, given input $\mathbf{Z}^{2,\ell-1}$, then we have*

1. *for $\ell = 1$, if $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_1$, $\mathbf{logit}^{(t)}_{5,j} = \Omega(1)$ for $j \in \{j_1, j_1'\}$, $1 - \mathbf{logit}^{(t)}_{5,j_1} - \mathbf{logit}^{(t)}_{5,j_1'} = \frac{1}{\mathsf{poly}d}$.*
2. *for $\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_2$, $\mathbf{logit}^{(t)}_{5,j} = O(\frac{1}{d})$ for all j.*

Proof. • For $\ell = 1$, by Lemma F.11 and Lemma F.6, we have

$$
\begin{aligned}
&F^{(t)}_{5,j_1}(\mathbf{Z}^{2,\ell-1}) = \sum_{r\in[m]} \mathbf{sReLU}\big(\Lambda^{(t)}_{5,j_1,r}\big) = \Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} + \varrho(m/q-1) = \Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} + O\Big(\frac{1}{\mathrm{polylog}d}\Big) \\
&F^{(t)}_{5,j_1'}(\mathbf{Z}^{2,\ell-1}) = \sum_{r\in[m]} \mathbf{sReLU}\big(\Lambda^{(t)}_{5,j_1',r}\big) = \Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}} + \varrho(m/q-1) \\
&F^{(t)}_{5,j}(\mathbf{Z}^{2,\ell-1}) = \sum_{r\in[m]} \mathbf{sReLU}\big(\Lambda^{(t)}_{5,j,r}\big) \le O\Big(\frac{1}{\mathrm{polylog}d}\Big) \text{ for } j \ne j_1, j_1' \in \tau(\mathcal{Y}) \\
&F^{(t)}_{5,j}(\mathbf{Z}^{2,\ell-1}) \le m\cdot\widetilde{O}(\sigma_0^q) \text{ for } j \notin \tau(\mathcal{Y}).
\end{aligned}
$$

Putting it together, we obtain

$$
\begin{aligned}
\mathbf{logit}^{(t)}_{5,j_1} &= \frac{1}{1 + e^{F^{(t)}_{5,j_1'} - F^{(t)}_{5,j_1}} + \Big(\sum_{j\ne j_1,j_1'\in\tau(\mathcal{Y})} e^{F^{(t)}_{5,j}} + \sum_{j\notin\tau(\mathcal{Y})} e^{F^{(t)}_{5,j}}\Big)\cdot e^{-F^{(t)}_{5,j_1}}} \\
&= \frac{1}{1 + e^{\Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}} - \Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}}} + \Big(O(\log d)\cdot e^{O(\frac{1}{\mathrm{polylog}d})} + O(d)\cdot e^{\widetilde{O}(m\sigma_0^q)}\Big)\cdot e^{-F^{(t)}_{5,j_1}}}.
\end{aligned}
$$

Thus, by Lemma F.12, and the fact that $B = C_B \log d$ for some sufficiently large constant $C_B > 0$, we have $\mathbf{logit}^{(t)}_{5,j_1} = \frac{1}{1+e^{-O(\delta)}+O(1)\cdot e^{-(C_B/3-1)\log d}} = \Omega(1)$. Similarly, we have

$$
\begin{aligned}
\mathbf{logit}^{(t)}_{5,j_1'} &= \frac{1}{1 + e^{-\Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}} + \Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}}} + \Big(O(\log d)\cdot e^{O(\frac{1}{\mathrm{polylog}d})} + O(d)\cdot e^{\widetilde{O}(m\sigma_0^q)}\Big)\cdot e^{-F^{(t)}_{5,j_1'}}} \\
&= \frac{1}{1 + e^{O(1)} + O(1)\cdot e^{-(C_B/3-1)\log d}} = \Omega(1).
\end{aligned}
$$

From the above analysis, it is easy to see that $1 - \mathbf{logit}^{(t)}_{5,j_1} - \mathbf{logit}^{(t)}_{5,j_1'} \le O\Big(\frac{1}{e^{(C_B/3-1)\log d}}\Big) = O(\frac{1}{\mathsf{poly}d})$.

• For $\ell = 2$, by Lemma F.11 and Lemma F.6, we have

$$
F^{(t)}_{5,j}(\mathbf{Z}^{2,\ell-1}) = \sum_{r\in[m]} \mathbf{sReLU}\big(\Lambda^{(t)}_{5,j_1,r}\big) \in [\varrho m/q, O(1) + \varrho(m/q-1)] \text{ for } j \in \{\tau(g_\ell(y_{\ell'}))\}_{\ell\in[2],\ell'\in\{0,1\}}
$$

$F_{5,j}^{(t)}(\mathbf{Z}^{2,\ell-1}) = \sum_{r\in[m]} \mathbf{sReLU}(\Lambda_{5,j,r}^{(t)}) \le O\Big(\frac{1}{\mathrm{polylog} d}\Big)$ for $j \in \tau(\mathcal{Y}) \setminus \{\tau(g_\ell(y_{\ell'}))\}_{\ell\in[2],\ell'\in\{0,1\}}$

$F_{5,j}^{(t)}(\mathbf{Z}^{2,\ell-1}) \le m \cdot \widetilde{O}(\sigma_0^q)$ for $j \notin \tau(\mathcal{Y})$.

Therefore, for any j, we have $\mathbf{logit}_j^{(t)} = O\big(\frac{1}{d}\big)$ since $F_{5,j'}^{(t)} \le O(1)$ for all j'.

□

In the following, we illustrate the activations on the non-high probability event.

Lemma F.14. *If Induction F.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 $j = \tau(g_2(y))$, $|\Lambda_{5,j,r_{g_2\cdot y}}^{(t)}| \le O(1)$ 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 $j = \tau(g(y_1))$, $|\Lambda_{5,j,r_{g\cdot y_1}}^{(t)}| \le O(1)$ 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 $j \neq j_2 \in \tau(\mathcal{Y})$, $|\Lambda_{5,j,r}^{(t)}| \le O(1)$ for all $r \in \widehat{\mathfrak{A}}_j$.*
   - (d) *$g_1 \neq g_2 \wedge y_0 \neq y_1 \wedge \Big(g_1(y_0) = g_2(y_1) \vee g_2(y_0) = g_1(y_1)\Big)$, $|\Lambda_{5,j,r}^{(t)}| \le O(1)$ 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 F.15. *If Induction F.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 $1 - \mathbf{logit}_{5,j_1}^{(t)} = O\Big(\frac{1}{\mathsf{poly} d}\Big)$.*
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 $j \in \{j_2, \tau(g_2(y_0))\}$, $1 - \mathbf{logit}_{5,j_2}^{(t)} - \mathbf{logit}_{5,\tau(g_2(y_0))}^{(t)} = \frac{1}{\mathsf{poly} d}$.*
   - (b) *if $g_1 \neq g_2 \wedge y_0 = y_1$, $\mathbf{logit}_{5,j}^{(t)} = \Omega(1)$ for $j \in \{j_2, \tau(g_1(y_1))\}$, $1 - \mathbf{logit}_{5,j_2}^{(t)} - \mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} = \frac{1}{\mathsf{poly} d}$.*

(c) *if* $g_1 = g_2 \wedge y_0 = y_1$, $1 - \mathbf{logit}^{(t)}_{5,j_2} = O\Big(\frac{1}{\mathsf{poly}d}\Big)$.

(d) $g_1 \neq g_2 \wedge y_0 \neq y_1 \wedge \Big(g_1(y_0) = g_2(y_1) \vee g_2(y_0) = g_1(y_1)\Big)$, $\mathbf{logit}^{(t)}_{5,j_2} = O(\frac{1}{d})$ *for all* j.

F.3.2 Gradient Lemma

Lemma F.16. *If Induction F.1 holds for all iterations* $< t$, *given* $s \in \tau(\mathcal{X})$, *for* $[\mathbf{Q}^{(t)}_{4,3}]_{s,s}$, *we have*

$$\sum_{\ell=1}^{2} \Big[-\nabla_{\mathbf{Q}^{(t)}_{4,3}} \mathsf{Loss}^{2,\ell}_5 \Big]_{s,s} = \Theta\Big(\frac{\log d}{d}\Big).$$

Proof. By gradient decompositions, we have

$$\sum_{\ell=1}^{2} \Big[-\nabla_{\mathbf{Q}^{(t)}_{4,3}} \mathsf{Loss}^{2,\ell}_5 \Big]_{s,s} = \sum_{\ell\in[2]} \sum_{\kappa\in\{i,ii,iii\}} \mathcal{N}^{(t)}_{s,3,\ell,\kappa}.$$

By Lemma F.13 and Lemma F.15, it is straightforward to see that $|\mathcal{N}^{(t)}_{s,3,1,iii}|, |\mathcal{N}^{(t)}_{s,3,2,iii}| = O(\frac{1}{\mathsf{poly}d})$, and thus we can focus on other terms.

By Lemma F.11 and Lemma F.14, we have

$$\begin{aligned}
\mathcal{N}^{(t)}_{s,3,1,i} = \mathbb{E}\Bigg[& \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1} \cdot (1 - \mathbf{logit}^{(t)}_{5,j_1}) \cdot \\
& \Big(\Big(\psi_{j_1,r_{g_1\cdot y_0}}(g_1) - \Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big) \mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1} \Bigg] \\
+ \mathbb{E}\Bigg[& \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1} \cdot (1 - \mathbf{logit}^{(t)}_{5,j_1}) \cdot \\
& \Big(\Big(\psi_{j_1,r_{g_1\cdot y_0}}(g_1) - \Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big) \mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1^c} \Bigg] \\
\overset{(a)}{=} & \Theta(\frac{1}{d}) \cdot \Omega(1) \cdot \Theta(B) + \Theta(\frac{1}{d}) \cdot O(\frac{1}{\mathsf{poly}d}) \cdot \Theta(B) \cdot O(\frac{1}{\log d}) \\
= & \Theta\Big(\frac{\log d}{d}\Big),
\end{aligned}$$

where (a) follows from Lemma F.12, Lemma F.13, Lemma F.14 and Lemma F.15, and the fact that $\tau(x_0) = s$ holds with probability $\frac{1}{d}$. $\mathcal{N}^{(t)}_{s,3,2,i}$ can be upper bounded similarly.

Moving to $\mathcal{N}^{(t)}_{s,3,1,ii}$, noticing that $\psi_{j_1',r_{g_2\cdot y_0}}(g_1) = -B + O(\delta)$ on $\mathcal{E}_1$, we have

$$\begin{aligned}
\mathcal{N}^{(t)}_{s,3,1,ii} = -\mathbb{E}\Bigg[& \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1} \cdot \mathbf{logit}^{(t)}_{5,j_1'} \cdot \\
& \Big(\Big(\psi_{j_1',r_{g_2\cdot y_0}}(g_1) - \Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big) \mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1} \Bigg] \\
- \mathbb{E}\Bigg[& \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1} \cdot \sum_{y\neq y_0\in\mathcal{Y}} \mathbf{logit}^{(t)}_{5,\tau(g_1(y))} \cdot \\
& \Big(\Big(\psi_{\tau(g_1(y)),r_{g_1\cdot y}}(g_1) - \Lambda^{(t)}_{5,\tau(g_1(y)),r_{g_1\cdot y}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big) \mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1^c} \Bigg] \\
= & \Theta(\frac{1}{d}) \cdot \Omega(1) \cdot \Theta(B) - \Theta(\frac{1}{d}) \cdot O(\frac{1}{\mathsf{poly}d}) \cdot \Theta(B) \cdot O(\frac{1}{\log d})
\end{aligned}$$

$$= \Theta\Big(\frac{\log d}{d}\Big).$$

For $\mathcal{N}^{(t)}_{s,3,2,ii}$, we only need to control the negative gradient, since the positive part can be easily upper bounded by $O\Big(\frac{\log d}{d}\Big)$.

$$\begin{aligned}
\mathcal{N}^{(t)}_{s,3,2,ii} \geq\ & -\Theta(\frac{1}{d}) \cdot O(\frac{1}{d}) \cdot \Theta(B) - \mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \cdot \sum_{y\neq y_1\in\mathcal{Y}} \mathbf{logit}^{(t)}_{5,\tau(g_2(y))} \cdot \\
& \Big(\Big(\psi_{\tau(g_2(y)),r_{g_2\cdot y}}(g_2) - \Lambda_{5,j,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s,g_1=g_2,y_0\neq y_1\}}\Bigg] \\
\overset{(a)}{\geq}\ & -\Theta(\frac{1}{d}) \cdot O(\frac{1}{d}) \cdot \Theta(B) - \Theta(\frac{1}{d}) \cdot O(1) \cdot \Theta(B) \cdot O(\frac{1}{\log d}) \\
\geq\ & -O\Big(\frac{1}{d}\Big),
\end{aligned}$$

where (a) is due to the fact that $\mathbf{logit}^{(t)}_{5,\tau(g_2(y_0))} = \Omega(1)$ and $\mathbf{logit}^{(t)}_{5,\tau(g_2(y))} = O(\frac{1}{d})$ for other $y \neq y_1, y_0$. Putting everything together, we complete the proof.

□

Lemma F.17 (Negative gradient)**.** *If Induction F.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, we have*

$$\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}^{(t)}_{4,4}}\mathsf{Loss}^{2,\ell}_5\Big]_{s,s} \geq -O\Big(\frac{1}{\log d}\Big)\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}^{(t)}_{4,3}}\mathsf{Loss}^{2,\ell}_5\Big]_{s,s}.$$

Proof. By gradient decompositions, we have

$$\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}^{(t)}_{4,4}}\mathsf{Loss}^{2,\ell}_5\Big]_{s,s} = \sum_{\ell\in[2]}\sum_{\kappa\in\{i,ii,iii\}}\mathcal{N}^{(t)}_{s,4,\ell,\kappa}.$$

Similarly as Lemma F.16, $|\mathcal{N}^{(t)}_{s,4,1,iii}|, |\mathcal{N}^{(t)}_{s,4,2,iii}| = O(\frac{1}{\mathsf{poly}d})$, and thus we can focus on other terms. By Lemma F.11, Lemma F.12, and Lemma F.14, we have

$$\begin{aligned}
\mathcal{N}^{(t)}_{s,4,1,i} + \mathcal{N}^{(t)}_{s,4,1,ii} =\ & \mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0} \cdot (1 - \mathbf{logit}^{(t)}_{5,j_1}) \cdot \\
& \Big(\Big(\psi_{j_1,r_{g_1\cdot y_0}}(y_0) - \Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1}\Bigg] \qquad (58) \\
& + \mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0} \cdot (1 - \mathbf{logit}^{(t)}_{5,j_1}) \cdot \\
& \Big(\Big(\psi_{j_1,r_{g_1\cdot y_0}}(y_0) - \Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1^c}\Bigg] \qquad (59) \\
& - \mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0} \cdot \mathbf{logit}^{(t)}_{5,j_1'} \cdot \\
& \Big(\Big(\psi_{j_1',r_{g_2\cdot y_0}}(y_0) - \Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1}\Bigg] \qquad (60)
\end{aligned}$$

$$
\begin{aligned}
&- \mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0}\cdot \sum_{g\neq g_1\in\mathcal{G}} \mathbf{logit}^{(t)}_{5,\tau(g(y_0))}\cdot \\
&\qquad \Big(\Big(\psi_{\tau(g(y_0)),r_{g\cdot y_0}}(y_0) - \Lambda^{(t)}_{5,\tau(g(y_0)),r_{g\cdot y_0}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1^c}\Bigg]
\end{aligned} \tag{61}
$$

Firstly consider the event $\{\tau(x_0)=s\}\cap\mathcal{E}_1$, we have

$$
\begin{aligned}
&(58)+(60)\\
&= \mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0}\cdot\Big((1-\mathbf{logit}^{(t)}_{5,j_1})\cdot\Big(\psi_{j_1,r_{g_1\cdot y_1}}(y_0) - \Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} \pm \widetilde{O}(\sigma_0)\pm\widetilde{O}(\delta^q)\Big)\\
&\qquad\qquad - \mathbf{logit}^{(t)}_{5,j_1'}\cdot\Big(\psi_{j_1',r_{g_2\cdot y_0}}(y_0) - \Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}} \pm \widetilde{O}(\sigma_0)\pm\widetilde{O}(\delta^q)\Big)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1}\Bigg]\\
&\overset{(a)}{=} \mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0}\cdot\Big((1-\mathbf{logit}^{(t)}_{5,j_1})\cdot\Big(\psi_{j_1,r_{g_1\cdot y_1}}(y_0) - \Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} \pm \widetilde{O}(\sigma_0)\pm\widetilde{O}(\delta^q)\Big)\\
&\quad - (1-\mathbf{logit}^{(t)}_{5,j_1} - \frac{1}{\mathsf{poly}d})\cdot\Big(\psi_{j_1',r_{g_2\cdot y_0}}(y_0) - \Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}} \pm \widetilde{O}(\sigma_0)\pm\widetilde{O}(\delta^q)\Big)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1}\Bigg]\\
&\geq -\mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0}\cdot\Big((1-\mathbf{logit}^{(t)}_{5,j_1})\cdot\\
&\qquad\Big(\psi_{j_1,r_{g_1\cdot y_1}}(y_0) - \psi_{j_1',r_{g_2\cdot y_0}}(y_0) + \Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}} - \Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} \pm \widetilde{O}(\sigma_0)\pm\widetilde{O}(\delta^q)\Big)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1}\Bigg]\\
&\overset{(b)}{\geq} -\mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0}\cdot\Big((1-\mathbf{logit}^{(t)}_{5,j_1})\cdot\Big(O(\delta)+O(1)\pm\widetilde{O}(\sigma_0)\pm\widetilde{O}(\delta^q)\Big)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1}\Bigg]\\
&\geq -O\Big(\frac{\mathcal{N}^{(t)}_{s,3,1,i}}{\log d}\Big),
\end{aligned}
$$

where (a) follows from Lemma F.13; (b) follows from Lemma F.12 and Lemma F.3.

Notice that (59) ≥ 0 since $\psi_{j_1,r_{g_1\cdot y_0}}(y_0) - \Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} \geq \Omega(B)$, thus we just need to consider the possible negative gradient from (61). By Lemma F.15, we have $\sum_{g\neq g_1\in\mathcal{G}}\mathbf{logit}^{(t)}_{5,\tau(g(y_0))} \leq O(\frac{1}{\mathsf{poly}d})$ on $\{\tau(x_0)=s\}\cap\mathcal{E}_1^c$, and hence (61) $\ll \mathcal{N}^{(t)}_{s,3,1,i}$.

Moving to the gradient from $\ell=2$, it is straightforward to see that $\mathcal{N}^{(t)}_{s,4,2,i}$ is non-negative, and thus we can focus on the the possible negative gradient from $\mathcal{N}^{(t)}_{s,4,2,ii}$.

$$
\begin{aligned}
\mathcal{N}^{(t)}_{s,4,2,ii} = &-\mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}^{(t)}_{5,j}\cdot\\
&\qquad\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda^{(t)}_{5,j,r})\cdot\Big(\psi_{j,r}(y_1) - \Lambda^{(t)}_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\mathcal{E}_2}\Bigg]\\
&-\mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}^{(t)}_{5,j}\cdot\\
&\qquad\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda^{(t)}_{5,j,r})\cdot\Big(\psi_{j,r}(y_1) - \Lambda^{(t)}_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\mathcal{E}_2^c}\Bigg]
\end{aligned}
$$

$$\geq -\Theta(\frac{1}{d})\cdot O(\frac{1}{d})\cdot\Theta(B)-\Theta(\frac{1}{d})\cdot O(1)\cdot\Theta(B)\cdot O(\frac{1}{\log d})\geq -O\Big(\frac{\mathcal{N}_{s,3,1,i}^{(t)}}{\log d}\Big).$$

Putting everything together, and combining with the fact that $\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}_{4,3}^{(t)}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s}=\Theta(\mathcal{N}_{s,3,1,i}^{(t)})$ from Lemma F.16, we complete the proof. □

Lemma F.18 (Growth of gap). *If Induction F.1 holds for all iterations $< t$, given $s\in\tau(\mathcal{X})$, we have*

$$\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}_{4,3}^{(t)}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s}+\sum_{\ell=1}^{2}\Big[\nabla_{\mathbf{Q}_{4,4}^{(t)}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s}\geq\Omega\Big(\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}_{4,3}^{(t)}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s}\Big).$$

Proof. By gradient decompositions in (44)-(55), we have

$$\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}_{4,3}^{(t)}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s}+\sum_{\ell=1}^{2}\Big[\nabla_{\mathbf{Q}_{4,4}^{(t)}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s}=\sum_{\ell\in[2]}\sum_{\kappa\in\{i,ii,iii\}}\mathcal{N}_{s,3,\ell,\kappa}^{(t)}-\mathcal{N}_{s,4,\ell,\kappa}^{(t)}.$$

Due to Lemma F.13 and Lemma F.15, $|\mathcal{N}_{s,p,\ell,iii}^{(t)}|=O(\frac{1}{\mathsf{poly}d})$ 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 F.10, we have

$$\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{ans},0}^{(t)}\leq\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}^{(t)},\quad\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}\leq\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}.$$

Hence, for $\ell\in[2]$, we have

$$\mathcal{N}_{s,3,\ell,i}^{(t)}-\mathcal{N}_{s,4,\ell,i}^{(t)}\geq -O(\delta)\cdot O(\frac{1}{d})\cdot\Theta(1)\geq -O(\delta/d).$$

$$\begin{aligned}
&\mathcal{N}_{s,3,1,ii}^{(t)}-\mathcal{N}_{s,4,1,ii}^{(t)}\\
&=-\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}^{(t)}\cdot\mathbf{logit}_{5,j_1'}^{(t)}\cdot\\
&\qquad\Big(\Big(\psi_{j_1',r_{g_2\cdot y_0}}(g_1)-\Lambda_{5,j_1',r_{g_2\cdot y_0}}^{(t)}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1}\Bigg]\\
&+\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{ans},0}^{(t)}\cdot\mathbf{logit}_{5,j_1'}^{(t)}\cdot\\
&\qquad\Big(\Big(\psi_{j_1',r_{g_2\cdot y_0}}(y_0)-\Lambda_{5,j_1',r_{g_2\cdot y_0}}^{(t)}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1}\Bigg]\\
&-\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}^{(t)}\cdot\sum_{y\neq y_0\in\mathcal{Y}}\mathbf{logit}_{5,\tau(g_1(y))}^{(t)}\cdot\\
&\qquad\Big(\Big(\psi_{\tau(g_1(y)),r_{g_1\cdot y}}(g_1)-\Lambda_{5,\tau(g_1(y)),r_{g_1\cdot y}}^{(t)}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1^c}\Bigg]\\
&+\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{ans},0}^{(t)}\cdot\sum_{y\neq y_0\in\mathcal{Y}}\mathbf{logit}_{5,\tau(g_1(y))}^{(t)}\cdot\\
&\qquad\Big(\Big(\psi_{\tau(g_1(y)),r_{g_1\cdot y}}(y_0)-\Lambda_{5,\tau(g_1(y)),r_{g_1\cdot y}}^{(t)}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1^c}\Bigg]\\
&\overset{(a)}{\geq}\Theta(\frac{1}{d})\cdot\Omega(1)\cdot\Theta(B)-\Theta(\frac{1}{d})\cdot O(\frac{1}{\mathsf{poly}d})\cdot O(B)\cdot O(\frac{1}{\log d})\geq\Omega\Big(\frac{\log d}{d}\Big).
\end{aligned}$$

where (a) follows from Lemma F.3 that $\psi_{j_1',r_{g_2\cdot y_0}}(g_1) - \Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}} \le -\Omega(B)$, and $\psi_{j_1',r_{g_2\cdot y_0}}(y_0) - \Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}} \ge \Omega(B)$.

$$
\begin{aligned}
&\mathcal{N}^{(t)}_{s,3,2,ii} - \mathcal{N}^{(t)}_{s,4,2,ii} \\
&= -\mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}^{(t)}_{5,j}\cdot \\
&\qquad \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda^{(t)}_{5,j,r})\cdot\Big(\psi_{j,r}(g_2) - \Lambda^{(t)}_{5,j,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\mathcal{E}_2}\Bigg] \\
&+ \mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}^{(t)}_{5,j}\cdot \\
&\qquad \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda^{(t)}_{5,j,r})\cdot\Big(\psi_{j,r}(y_1) - \Lambda^{(t)}_{5,j,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\mathcal{E}_2}\Bigg] \\
&- \mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}^{(t)}_{5,j}\cdot \\
&\qquad \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda^{(t)}_{5,j,r})\cdot\Big(\psi_{j,r}(g_2) - \Lambda^{(t)}_{5,j,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\mathcal{E}_2^c}\Bigg] \\
&+ \mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}^{(t)}_{5,j}\cdot \\
&\qquad \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda^{(t)}_{5,j,r})\cdot\Big(\psi_{j,r}(y_1) - \Lambda^{(t)}_{5,j,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\mathcal{E}_2^c}\Bigg] \\
&\overset{(a)}{\ge} -\Theta\Big(\frac{1}{d}\Big)\cdot O\Big(\frac{1}{d}\Big)\cdot\Theta(B) - \Theta\Big(\frac{1}{d}\Big)\cdot O(1)\cdot\Theta(B)\cdot O\Big(\frac{1}{\log d}\Big) \ge -O\Big(\frac{1}{d}\Big).
\end{aligned}
$$

where (a) follows from Lemma F.11 and Lemma F.13, which together imply that on the event $\mathcal{E}_2$, only a constant number of neurons are activated and $\mathbf{logit}^{(t)}_{5,j} \le O\left(\frac{1}{d}\right)$ for all $j \neq j_2$; and from Lemma F.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 $\mathbf{logit}^{(t)}_{5,j} \ge \Omega(1)$ while $\mathbf{logit}^{(t)}_{5,j} \le O\Big(\frac{1}{\mathsf{poly}d}\Big)$ for other $j \in \tau(\mathcal{Y})$.

Putting it all together, we finish the proof. □

Lemma F.19. *If Induction F.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}^{(t)}_{4,3}]_{s,s} \ge \Omega(\frac{\varrho}{\log d})$, we have*

$$
\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}^{(t)}_{4,4}}\mathsf{Loss}^{2,\ell}_5\Big]_{s,s} \ge \Omega\Bigg(\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}^{(t)}_{4,3}}\mathsf{Loss}^{2,\ell}_5\Big]_{s,s}\Bigg).
$$

Proof. Notice that by Lemma F.17, when $[\mathbf{Q}^{(t)}_{4,3}]_{s,s} \ge \Omega(\frac{\varrho}{\log d})$, we have $[\mathbf{Q}^{(t)}_{4,4}]_{s,s} \ge -O(\frac{\varrho}{\log^2 d})$. Hence

$$
\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} \ge \Omega\Big(\frac{\varrho}{\log d}\Big),
$$

which implies $\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}}(\mathbf{Z}^{2,1}) \ge \Omega(\frac{\varrho}{\log d})\cdot B \pm O(\delta) \ge \varrho$ already lies in the linear regime for $\mathbf{Z}^{2,1} \in \mathcal{E}_2$. Then we have

$$
\mathcal{N}^{(t)}_{s,4,2,i} \ge \mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot\Big((1-\mathbf{logit}^{(t)}_{5,j_2})\cdot
$$

$$\Big(\psi_{j_2,r_{g_2\cdot y_1}} - \Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\mathcal{E}_2}\Big]$$

$$\geq \Omega(1)\cdot\Theta(B)\cdot\Theta\Big(\frac{1}{d}\Big) \geq \Omega\Big(\frac{\log d}{d}\Big).$$

Moreover, from Lemma F.17, the magnitude of negative gradient from other $\mathcal{N}$ terms can be upper bounded by $O\Big(\frac{\mathcal{N}^{(t)}_{s,3,1,i}}{\log d}\Big)$. Therefore, combining with the fact that $\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}^{(t)}_{4,3}}\mathsf{Loss}^{2,\ell}_5\Big]_{s,s} = \Theta(\mathcal{N}^{(t)}_{s,3,1,i})$ from Lemma F.16, we complete the proof. □

Lemma F.20. *If Induction F.1 holds for all iterations $< t$, given $s \neq s' \in \tau(\mathcal{X})$, for $p \in \{3,4\}$, we have*

$$\Big|\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}^{(t)}_{4,p}}\mathsf{Loss}^{2,\ell}_5\Big]_{s,s'}\Big| \leq O\Big(\frac{\log d}{d^2}\Big) = O(\frac{1}{d})\cdot\Big|\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}^{(t)}_{4,p}}\mathsf{Loss}^{2,\ell}_5\Big]_{s,s}\Big|.$$

Proof. The proof follows directly by combining the expressions from Lemma F.8, Lemma F.9, and Lemma F.16, along with the fact that the event $\{\tau(x_0)=s, \tau(x_1)=s'\}$ occurs with probability $O\big(\frac{1}{d^2}\big)$. □

F.3.3 At the End of Stage 2.1

Putting gradient lemmas together, we can directly prove that Induction F.1 holds for all iterations t until the end of stage 2.1, where we can conclude the following:

Lemma F.21 (End of stage 2.1)**.** *Given $s \in \tau(\mathcal{X})$, Induction F.1 holds for all iterations $t < T_{2,1,s} = O\Big(\frac{d}{\eta\log^2 d}\Big)$, then at the end of stage 2.1, we have*

(a) $[\mathbf{Q}^{(t)}_{4,p}]_{s,s} = \Omega(\frac{1}{\log d})$ *for* $p \in \{3,4\}$*;*

(b) $[\mathbf{Q}^{(t)}_{4,3}]_{s,s} - [\mathbf{Q}^{(t)}_{4,4}]_{s,s} \geq \Omega\Big(\frac{1}{\log d}\Big)$*;*

(c) $\Big|[\mathbf{Q}^{(t)}_{4,p}]_{s,s'}\Big| \leq O\Big(\frac{[\mathbf{Q}^{(t)}_{4,p}]_{s,s}}{d}\Big)$ *for* $s' \in \tau(\mathcal{X}) \neq s$ *for* $p \in \{3,4\}$*; otherwise,* $[\mathbf{Q}^{(t)}_{4,p}]_{s,s'} = 0$.

F.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 F.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) $[\mathbf{Q}^{(t)}_{4,3}]_{s,s}, [\mathbf{Q}^{(t)}_{4,4}]_{s,s} \leq O(1)$ *monotonically increases;*

(b) $[\mathbf{Q}^{(t)}_{4,3}]_{s,s} - [\mathbf{Q}^{(t)}_{4,4}]_{s,s} \in \Big[\Omega(\frac{1}{\log d}), O(1)\Big]$*;*

(c) for $(p,q) \in \{(4,3),(4,4)\}$, $\Big|[\mathbf{Q}^{(t)}_{p,q}]_{s,s'}\Big| \leq O\Big(\frac{[\mathbf{Q}^{(t)}_{p,q}]_{s,s}}{d}\Big)$ *for* $s' \in \tau(\mathcal{X}) \neq s$*; other* $[\mathbf{Q}^{(t)}_{p,q}]_{s,s'} = 0$.

Throughout the following analysis, instead of $\mathcal{E}_2$ defined in (57), we consider a renewed event $\widetilde{\mathcal{E}}_2$ for $\ell = 2$:

$$\widetilde{\mathcal{E}}_2 \triangleq \Big\{g_1 \neq g_2 \wedge y_0 \neq y_1\Big\}. \tag{62}$$

F.4.1 Attention and Logit Preliminaries

The proof in this part proceeds analogously to the arguments in Appendix F.3.1, with the induction hypothesis from Induction F.2 incorporated. Hence, we omit the details here.

Lemma F.22. *If Induction F.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) $\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1} \in \Big[\frac{1}{3} + \Omega\big(\frac{1}{\log d}\big), \frac{1}{3} + c_1\Big]$, *where $c_1 > 0$ is a small constant;*
   (b) $\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},2} \in \Big[\frac{1}{3} - c_2, \frac{1}{3} - \Omega\big(\frac{1}{\log d}\big)\Big]$, *where $c_2 > 0$ is a small constant;*
   (c) $\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},2} + \Omega\big(\frac{1}{\log d}\big) \le \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0}$;
   (d) $\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0} \in \Big[\Omega\big(\frac{1}{\log d}\big), c_3\Big]$.
2. *for $\ell = 2$,*
   (a) $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \in \Big[\frac{1}{4} + \Omega\big(\frac{1}{\log d}\big), \frac{1}{4} + c_4\Big]$, *where $c_4 > 0$ is a small constant;*
   (b) $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}, \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} \in \Big[\frac{1}{4} - c_5, \frac{1}{4} - \Omega\big(\frac{1}{\log d}\big)\Big]$, *moreover,* $\big|\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\big| \le O\Big(\frac{1}{d}\Big)$, *where $c_5 > 0$ is a small constant;*
   (c) $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}, \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} \le \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \Omega\Big(\frac{1}{\log d}\Big)$;
   (d) $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} \in \Big[\Omega\big(\frac{1}{\log d}\big), c_6\Big]$, *where $c_6 > 0$ is a small constant.*

Notice that the constant $c_1 - c_6$ depends on the threshold 0.0001 in Induction F.2. We choose the threshold 0.0001 small enough to ensure $2C_B(c_1 + c_2) < 1 - c_6 C_B$ and $1 - 4c_5 C_B > 0$.

Lemma F.23. *If Induction F.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) *for $j = j_1$, $\Lambda^{(t)}_{5,j_1,r} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_{j_1} \setminus \{r_{g_1\cdot y_0}\}$;*
   (b) *for $j = j_1' \triangleq \tau\big(g_2(y_0)\big)$, $\Lambda^{(t)}_{5,j_1',r} \ll -\varrho$ 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 activted for all $r \in \widehat{\mathfrak{A}}_j$, i.e., $\Lambda^{(t)}_{5,j,r} \ll -\varrho$.*
2. *$\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \in \widetilde{\mathcal{E}}_2$, then*
   (a) *for $j = j_2$, $\Lambda^{(t)}_{5,j_2,r} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2\cdot y_1}\}$;*
   (b) *for $j = j_2' \triangleq \tau\big(g_2(y_0)\big)$, $\Lambda^{(t)}_{5,j_2',r} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_{j_2'} \setminus \{r_{g_2\cdot y_0}\}$;*
   (c) *for other $j \in \tau(\mathcal{Y})$, r is not activted for all $r \in \widehat{\mathfrak{A}}_j$, i.e., $\Lambda^{(t)}_{5,j,r} \ll -\varrho$.*

Lemma F.24. *If Induction F.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) $\Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} \ge \Big(\frac{1}{3} + \Omega\big(\frac{1}{\log d}\big)\Big)B$;
   (b) $\Omega(1) \le \Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} - \Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}} \le 2(c_1 + c_2)B$.
2. *$\ell = 2$, for $\mathbf{Z}^{2,\ell-1} \in \widetilde{\mathcal{E}}_2$,*
   (a) $\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} \in \Big[\Omega(1), 4c_5 B\Big]$;
   (b) $\Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}} \in \big[\Omega(1), c_6 B\big]$ *and* $\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} - \Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}} \ge \Omega(1)$.

Lemma F.25. *If Induction F.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 $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_1$, $\mathbf{logit}_{5,j_1'}^{(t)} \geq \Omega\Big(\frac{1}{d^{2C_B(c_1+c_2)}}\Big)$;*
2. *for $\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \in \widetilde{\mathcal{E}}_2$, $1 - \mathbf{logit}_{5,j_2}^{(t)} = \Omega(1)$, $\mathbf{logit}_{5,j_2'}^{(t)} = O\big(\frac{1}{d^{1-c_6C_B}}\big)$.*

Proof. • For $\ell = 1$, we have

$$\begin{aligned}\mathbf{logit}_{5,j_1'}^{(t)} &= \frac{1}{1 + e^{\Lambda_{5,j_1,r_{g_1\cdot y_0}}^{(t)} - \Lambda_{5,j_1',r_{g_2\cdot y_0}}^{(t)}} + O(d)\cdot e^{-\Lambda_{5,j_1',r_{g_2\cdot y_0}}^{(t)}}} \\ &\overset{(a)}{\geq} \frac{1}{1 + e^{2(c_1+c_2)B} + O(d)\cdot e^{-\left(\frac{1}{3}-2c_1-2c_2\right)B}} \geq \Omega\Big(\frac{1}{d^{2C_B(c_1+c_2)}}\Big),\end{aligned}$$

where the inequality (a) follows from Lemma F.24 and the last inequality is due to the fact that $(c_1 + c_2)$ is some sufficiently small constant s.t., $e^{-\left(\frac{1}{3}-2c_1-2c_2\right)B} = 1/\mathsf{poly}d$.

• For $\ell = 2$, we have

$$\begin{aligned}\mathbf{logit}_{j_2}^{(t)} &= \frac{1}{1 + e^{-\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)} + \Lambda_{5,j_2',r_{g_2\cdot y_0}}^{(t)}} + O(d)\cdot e^{-\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)}}} \\ &\overset{(a)}{\leq} \frac{1}{1 + O(d)\cdot e^{-4c_5B}} = O\bigg(\frac{1}{d^{1-4c_5C_B}}\bigg),\end{aligned}$$

where the inequality (a) follows from Lemma F.24. Similarly, we have

$$\begin{aligned}\mathbf{logit}_{j_2'}^{(t)} &= \frac{1}{1 + e^{\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)} - \Lambda_{5,j_2',r_{g_2\cdot y_0}}^{(t)}} + O(d)\cdot e^{-\Lambda_{5,j_2',r_{g_2\cdot y_0}}^{(t)}}} \\ &\leq \frac{1}{1 + e^{\Omega(1)} + O(d)\cdot e^{-c_6B}} = O\bigg(\frac{1}{d^{1-c_6C_B}}\bigg).\end{aligned}$$

□

Lemma F.26. *If Induction F.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) *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 $\Lambda_{5,j,r_{g_1\cdot y}}^{(t)} \in \Big[\Big(\frac{1}{3} - c_2\Big)B, \Big(\frac{1}{3} + c_1\Big)B\Big]$ 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 $\mathbf{Z}^{2,\ell-1} \notin \widetilde{\mathcal{E}}_2$, then*
   (a) *if $g_1 = g_2 \wedge y_0 \neq y_1$,*
      i. *for $j = j_2$, $\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)} \in \Big[\frac{1}{2}B + \Omega(1), \big(\frac{1}{2} + 2c_5\big)B\Big]$ 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 $j = \tau(g_2(y))$, $\Lambda_{5,j,r_{g_2\cdot y}}^{(t)} \in [\Omega(1), c_6B]$ 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* $j = j_2$, $\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} \in \Big[\frac{1}{2}B + \Omega(1), \big(\frac{1}{2} + 2c_5\big)B\Big]$ *and* $\Lambda^{(t)}_{5,j_2,r} \ll -\varrho$ *for* $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2\cdot y_1}\}$;

ii. *for* $j = \tau(g_1(y_1))$, $\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} - \Lambda^{(t)}_{5,j_2,r_{g_1\cdot y_1}} \geq \Omega(1)$ *and* $\Lambda^{(t)}_{5,j,r} \ll -\varrho$ *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 activted for all* $r \in \widehat{\mathfrak{A}}_j$, *i.e.,* $\Lambda^{(t)}_{5,j,r} \ll -\varrho$.

(c) *if* $g_1 = g_2 \wedge y_0 = y_1$,

i. *for* $j = j_2$, $\widehat{\mathfrak{A}}_{j_2} = \{r_{g_2\cdot y_1}\}$, $\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = B \pm O(\delta)$;

ii. *for other* $j \in \tau(\mathcal{Y})$, *assuming* $j = \tau(g_2(y))$, $\Lambda^{(t)}_{5,j,r_{g_2\cdot y}} \in [\Omega(1), c_6 B]$ *and* $\Lambda^{(t)}_{5,j,r} \ll -\varrho$ *for* $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2\cdot y}\}$.

Lemma F.27. *If Induction F.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* $1 - \mathbf{logit}^{(t)}_{5,j_1} = O\Big(\frac{1}{\mathsf{poly}d}\Big)$.

2. $\ell = 2$, *if* $\mathbf{Z}^{2,\ell-1} \notin \widetilde{\mathcal{E}}_2$, *then*

   (a) *if* $g_1 = g_2 \wedge y_0 \neq y_1$, $\mathbf{logit}^{(t)}_{5,j_2} = \Omega(1)$, $1 - \mathbf{logit}^{(t)}_{5,j_2} - \mathbf{logit}^{(t)}_{5,\tau(g_2(y_0))} = \frac{1}{\mathsf{poly}d}$.

   (b) *if* $g_1 \neq g_2 \wedge y_0 = y_1$, $\mathbf{logit}^{(t)}_{5,j_2} = \Omega(1)$, $1 - \mathbf{logit}^{(t)}_{5,j_2} - \mathbf{logit}^{(t)}_{5,\tau(g_1(y_1))} = \frac{1}{\mathsf{poly}d}$.

   (c) *if* $g_1 = g_2 \wedge y_0 = y_1$, $1 - \mathbf{logit}^{(t)}_{5,j_2} = O\Big(\frac{1}{\mathsf{poly}d}\Big)$.

F.4.2 Gradient Lemma

Lemma F.28. *If Induction F.2 holds for all iterations* $\in [T_{2,1,s}, t)$, *given* $s \in \tau(\mathcal{X})$, *for* $[\mathbf{Q}^{(t)}_{4,3}]_{s,s}$, *we have*

$$\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}^{(t)}_{4,3}}\mathsf{Loss}^{2,\ell}_5\Big]_{s,s} = \Theta\Big(\frac{\log d}{d}\Big).$$

Proof. The proof is similar to Lemma F.16, but we need to shift our focus to $\Big[-\nabla_{\mathbf{Q}^{(t)}_{4,3}}\mathsf{Loss}^{2,2}_5\Big]_{s,s}$. By Lemma F.23 and Lemma F.26, we have

$$\begin{aligned}
\mathcal{N}^{(t)}_{s,3,2,i} = \mathbb{E}\Bigg[&\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot(1-\mathbf{logit}^{(t)}_{5,j_2})\cdot\\
&\Big(\Big(\psi_{j_2,r_{g_2\cdot y_1}}(g_2) - \Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2}\Bigg]\\
+\mathbb{E}\Bigg[&\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot(1-\mathbf{logit}^{(t)}_{5,j_2})\cdot\\
&\Big(\Big(\psi_{j_2,r_{g_2\cdot y_1}}(g_2) - \Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2^c}\Bigg]\\
\overset{(a)}{=}\ &\Theta(\frac{1}{d})\cdot\Omega(1)\cdot\Theta(B) + \Theta(\frac{1}{d})\cdot O(1)\cdot\Theta(B)\cdot O(\frac{1}{\log d})\\
=\ &\Theta\Big(\frac{\log d}{d}\Big),
\end{aligned}$$

where the inequality (a) follows from Lemma F.25 and Lemma F.26.

Furthermore, for $\mathcal{N}^{(t)}_{s,3,2,ii}$, we have

$$\Big|\mathcal{N}^{(t)}_{s,3,2,ii}\Big|$$

$$
\begin{aligned}
&\leq \Bigg| \mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \cdot \mathbf{logit}^{(t)}_{5,j_2'} \cdot \\
&\qquad \Big(\Big(\psi_{j_2',r_{g_2\cdot y_0}}(g_2) - \Lambda^{(t)}_{5,j,r} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2} \Bigg] \Bigg| \\
&\Bigg| \mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \cdot \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}^{(t)}_{5,j} \cdot \\
&\qquad \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda^{(t)}_{5,j,r}) \cdot \Big(\psi_{j,r}(g_2) - \Lambda^{(t)}_{5,j,r} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2^c} \Bigg] \Bigg| \\
&\leq \Theta\big(\frac{1}{d}\big) \cdot O\Big(\frac{1}{d^{1-c_6 C_B}}\Big) \cdot \Theta(B) + \Theta\big(\frac{1}{d}\big) \cdot O(1) \cdot \Theta(B) \cdot O\Big(\frac{1}{\log d}\Big) \leq O(\frac{1}{d}).
\end{aligned}
$$

$|\mathcal{N}^{(t)}_{s,3,1,i}|$ and $|\mathcal{N}^{(t)}_{s,3,2,i}|$ can be upper bounded by $O\big(\frac{\log d}{d}\big)$ as Lemma F.16. Thus we complete the proof. □

Lemma F.29. *If Induction F.2 holds for all iterations $\in [T_{2,1,s}, t)$, given $s \in \tau(\mathcal{X})$, we have*

$$
\sum_{\ell=1}^{2} \Big[-\nabla_{\mathbf{Q}^{(t)}_{4,4}} \mathsf{Loss}_5^{2,\ell} \Big]_{s,s} = \Theta\Big(\frac{\log d}{d}\Big).
$$

Proof. The proof follows the similar analysis as Lemma F.19, and we thus omit the details here.

□

Lemma F.30. *If Induction F.2 holds for all iterations $\in [T_{2,1,s}, t)$, given $s \in \tau(\mathcal{X})$, we have*

$$
\begin{aligned}
&\sum_{t'=T_{2,1,s}}^{t} \Bigg(\sum_{\ell=1}^{2} \Big[-\nabla_{\mathbf{Q}^{(t')}4,3} \mathsf{Loss}_5^{2,\ell} \Big]_{s,s} - \sum_{\ell=1}^{2} \Big[-\nabla_{\mathbf{Q}^{(t')}_{4,4}} \mathsf{Loss}_5^{2,\ell} \Big]_{s,s} \Bigg) \\
&\qquad \geq -O(\frac{1}{\log d}) \cdot \Bigg(\sum_{t'=T_{2,1,s}}^{t} \sum_{\ell=1}^{2} \Big[-\nabla_{\mathbf{Q}^{(t')}_{4,4}} \mathsf{Loss}_5^{2,\ell} \Big]_{s,s} \Bigg).
\end{aligned}
$$

Proof. Following the analogous analysis as Lemma F.18, we have

$$
\sum_{\ell=1}^{2} \Big[-\nabla_{\mathbf{Q}^{(t)}_{4,3}} \mathsf{Loss}_5^{2,\ell} \Big]_{s,s} + \sum_{\ell=1}^{2} \Big[\nabla_{\mathbf{Q}^{(t)}_{4,4}} \mathsf{Loss}_5^{2,\ell} \Big]_{s,s} = \sum_{\ell\in[2]} \sum_{\kappa\in\{i,ii,iii\}} \mathcal{N}^{(t)}_{s,3,\ell,\kappa} - \mathcal{N}^{(t)}_{s,4,\ell,\kappa}.
$$

Meanwhile $|\mathcal{N}^{(t)}_{s,p,\ell,iii}| = O(\frac{1}{\mathsf{poly}d})$ for $p \in \{3,4\}$ and $\ell \in [2]$, we can focus on the gradient difference between $\mathbf{Q}^{(t)}_{4,3}$ and $\mathbf{Q}^{(t)}_{4,4}$ contributed by other terms.

For $\ell = 1$, since $\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1} \geq \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0}$, and thus it is straightforward to see that $\mathcal{N}^{(t)}_{s,3,1,i} - \mathcal{N}^{(t)}_{s,4,1,i} \geq 0$. Futhermore, we have

$$
\begin{aligned}
&\mathcal{N}^{(t)}_{s,3,1,ii} - \mathcal{N}^{(t)}_{s,4,1,ii} \\
&= -\mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1} \cdot \mathbf{logit}^{(t)}_{5,j_1'} \cdot \\
&\qquad \Big(\Big(\psi_{j_1',r_{g_2\cdot y_0}}(g_1) - \Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1} \Bigg] \\
&+ \mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0} \cdot \mathbf{logit}^{(t)}_{5,j_1'} \cdot
\end{aligned}
$$

$$\Big(\Big(\psi_{j_1',r_{g_2\cdot y_0}}(y_0)-\Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1}\Big]$$

$$-\mathbb{E}\Big[\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1}\cdot\sum_{y\neq y_0\in\mathcal{Y}}\mathbf{logit}^{(t)}_{5,\tau(g_1(y))}\cdot$$

$$\Big(\Big(\psi_{\tau(g_1(y)),r_{g_1\cdot y}}(g_1)-\Lambda^{(t)}_{5,\tau(g_1(y)),r_{g_1\cdot y}}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1^c}\Big]$$

$$+\mathbb{E}\Big[\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0}\cdot\sum_{y\neq y_0\in\mathcal{Y}}\mathbf{logit}^{(t)}_{5,\tau(g_1(y))}\cdot$$

$$\Big(\Big(\psi_{\tau(g_1(y)),r_{g_1\cdot y}}(y_0)-\Lambda^{(t)}_{5,\tau(g_1(y)),r_{g_1\cdot y}}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1^c}\Big]$$

$$\overset{(a)}{\geq}\Theta(\frac{1}{d})\cdot\Omega\Big(\frac{1}{d^{2C_B(c_1+c_2)}}\Big)\cdot\Theta(B)-\Theta(\frac{1}{d})\cdot O(\frac{1}{\mathsf{poly}d})\cdot O(B)\cdot O(\frac{1}{\log d})\geq\Omega\Big(\frac{\log d}{d^{1+2C_B(c_1+c_2)}}\Big).$$

where the inequality (a) is due to Lemma F.25 and Lemma F.27.

For $\ell=2$, since $\mathbf{Attn}^{(t)}_{\mathsf{ans},\mathsf{pred},2\to\mathsf{pred},1}\geq\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}$, and thus it is straightforward to see that $\mathcal{N}^{(t)}_{s,3,2,i}-\mathcal{N}^{(t)}_{s,4,2,i}\geq 0$. Moreover, we have

$$\begin{aligned}
&\mathcal{N}^{(t)}_{s,3,2,ii}-\mathcal{N}^{(t)}_{s,4,2,ii}\\
&=-\mathbb{E}\Big[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\mathbf{logit}^{(t)}_{5,j_2'}\cdot\\
&\qquad\Big(\Big(\psi_{j_2',r_{g_2\cdot y_0}}(g_2)-\Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2}\Big] &(63)\\
&+\mathbb{E}\Big[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot\mathbf{logit}^{(t)}_{5,j_2'}\cdot\\
&\qquad\Big(\Big(\psi_{j_2',r_{g_2\cdot y_0}}(y_1)-\Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2}\Big] &(64)\\
&-\mathbb{E}\Big[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}^{(t)}_{5,j}\cdot\\
&\qquad\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda^{(t)}_{5,j,r})\cdot\Big(\psi_{j,r}(g_2)-\Lambda^{(t)}_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2^c}\Big] &(65)\\
&+\mathbb{E}\Big[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}^{(t)}_{5,j}\cdot\\
&\qquad\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda^{(t)}_{5,j,r})\cdot\Big(\psi_{j,r}(y_1)-\Lambda^{(t)}_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2^c}\Big]. &(66)
\end{aligned}$$

By Lemma F.25, we obtain that

$$(63)+(64)\geq-O(\frac{1}{d})\cdot O\Big(\frac{1}{d^{1-c_6C_B}}\Big)\cdot\Theta(B)\geq-O\Big(\frac{\log d}{d^{2-c_6C_B}}\Big).$$

By Lemma F.26, for $\widetilde{\mathcal{E}}_2^c$, we only need to consider the case that $g_1=g_2\wedge y_0\neq y_1$, and we have

$(65)+(66)$

$$
\begin{aligned}
&\geq -\mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\sum_{y\neq y_1\in\mathcal{Y}}\mathbf{logit}^{(t)}_{5,\tau(g_2(y))}\cdot \\
&\Big(\Big(\psi_{\tau(g_2(y)),r_{g_2\cdot y}}(g_2)-\Lambda^{(t)}_{5,\tau(g_2(y)),r_{g_2\cdot y}}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\{g_1=g_2\wedge y_0\neq y_1\}}\Bigg] \\
&+\mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot\sum_{y\neq y_1\in\mathcal{Y}}\mathbf{logit}^{(t)}_{5,\tau(g_2(y))}\cdot \\
&\Big(\Big(\psi_{\tau(g_2(y)),r_{g_2\cdot y}}(y_1)-\Lambda^{(t)}_{5,\tau(g_2(y)),r_{g_2\cdot y}}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\{g_1=g_2\wedge y_0\neq y_1\}}\Bigg] \\
&\geq -\Theta\Big(\frac{1}{d}\Big)\cdot O(1)\cdot\Theta(B)\cdot O\Big(\frac{1}{\log d}\Big)\geq -O\Big(\frac{1}{d}\Big).
\end{aligned}
$$

Putting it all together, combining with the fact that c_6 and (c_1+c_2) are sufficiently small, we finish the proof. □

Lemma F.31 (Lower bound of gap)**.** *If Induction F.2 holds for all iterations $\in[T_{2,1,s},t)$, given $s\in\tau(\mathcal{X})$, we have $[\mathbf{Q}^{(t)}_{4,3}]_{s,s}-[\mathbf{Q}^{(t)}_{4,4}]_{s,s}\geq\Omega(\frac{1}{\log d})$.*

Proof. Letting $\widetilde{T}$ denote the first time that $[\mathbf{Q}^{(t)}_{4,3}]_{s,s}-[\mathbf{Q}^{(t)}_{4,4}]_{s,s}\leq\frac{1}{2}\Big([\mathbf{Q}^{(T_{2,1,s})}_{4,3}]_{s,s}-[\mathbf{Q}^{(T_{2,1,s})}_{4,4}]_{s,s}\Big)$, which implies that

$$
\sum_{t'=T_{2,1,s}}^{\widetilde{T}}\Bigg(\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}^{(t')}4,3}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s}-\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}^{(t')}_{4,4}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s}\Bigg).
$$

Hence, by Lemma F.30, we have $[\mathbf{Q}^{(\widetilde{T})}_{4,3}]_{s,s}$ and $[\mathbf{Q}^{(\widetilde{T})}_{4,3}]_{s,s}$ reaches $\Omega(1)$. Thus, we can have a refined lower bound for (65) + (66) in Lemma F.30, and obtain:

$$
\begin{aligned}
&(65)+(66)\geq -O(\frac{1}{\mathsf{poly}d}) \\
&\quad-\mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\mathbf{logit}^{(t)}_{5,\tau(g_2(y_0))}\cdot \\
&\quad\Big(\Big(\psi_{\tau(g_2(y_0)),r_{g_2\cdot y_0}}(g_2)-\Lambda^{(t)}_{5,\tau(g_2(y_0)),r_{g_2\cdot y_0}}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\{g_1=g_2\wedge y_0\neq y_1\}}\Bigg] \\
&\quad+\mathbb{E}\Bigg[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot\mathbf{logit}^{(t)}_{5,\tau(g_2(y_0))}\cdot \\
&\quad\Big(\Big(\psi_{\tau(g_2(y_0)),r_{g_2\cdot y_0}}(y_1)-\Lambda^{(t)}_{5,\tau(g_2(y_0)),r_{g_2\cdot y_0}}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\{g_1=g_2\wedge y_0\neq y_1\}}\Bigg] \\
&\quad\overset{(a)}{\geq}-O\Big(\frac{1}{\mathsf{poly}d}\Big)-\Theta\Big(\frac{1}{d}\Big)\cdot O\Big(\frac{1}{d^{\Omega(1)}}\Big)\cdot\Theta(B)\cdot O\Big(\frac{1}{\log d}\Big)\geq -O\Big(\frac{1}{d^{1+\Omega(1)}}\Big).
\end{aligned}
$$

where (a) follows from the fact that on the event $\{g_1=g_2\wedge y_0\neq y_1\}$, we have $\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}}-\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_0}}\geq\Omega(B)$ once $[\mathbf{Q}^{(t)}_{4,3}]_{s,s}$ and $[\mathbf{Q}^{(t)}_{4,4}]_{s,s}$ reach constant magnitude, and consequently, the logit satisfies $\mathbf{logit}^{(t)}_{5,\tau(g_2(y_0))}\leq O\left(\frac{1}{d^{\Omega(1)}}\right)$.

Therefore,

$$
\sum_{t'=\widetilde{T}+1}^{t}\Bigg(\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}^{(t')}4,3}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s}-\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}^{(t')}_{4,4}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s}\Bigg)\geq -O\Big(\frac{1}{\log d\cdot d^{\Omega(1)}}\Big)\cdot O(1),
$$

which means that $[\mathbf{Q}^{(t)}_{4,3}]_{s,s}-[\mathbf{Q}^{(t)}_{4,4}]_{s,s}\geq\Omega\Big(\frac{1}{\log d}\Big)-O\Big(\frac{1}{\log d\cdot d^{\Omega(1)}}\Big)\geq\Omega\Big(\frac{1}{\log d}\Big)$. □

Lemma F.32. *If Induction F.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*

$$\Big|\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}_{4,p}^{(t)}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s'}\Big| \le O\Big(\frac{\log d}{d^2}\Big) = O\Big(\frac{1}{d}\Big)\cdot\Big|\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}_{4,p}^{(t)}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s}\Big|.$$

F.4.3 At the End of Stage 2.2

Putting gradient lemmas together, we can directly prove that Induction F.2 holds for all iterations t until the end of stage 2.2, where we can conclude the following:

Lemma F.33 (End of Stage 2.2)**.** *Given $s \in \tau(\mathcal{X})$, Induction F.2 holds for all iterations $T_{2,1,s} \le t < T_{2,2,s} = O\Big(\frac{d}{\eta \log d}\Big)$, 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) $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} - [\mathbf{Q}_{4,4}^{(t)}]_{s,s} \in \Big[\Omega\big(\frac{1}{\log d}\big), O(1)\Big]$;

(c) $|[\mathbf{Q}_{4,p}^{(t)}]_{s,s'}| \le O\Big(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,s}}{d}\Big)$ *for* $s' \in \tau(\mathcal{X}) \neq s$, *and other* $[\mathbf{Q}_{4,p}]_{s,s'} = 0$.

F.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

$$\epsilon_{\mathsf{attn}}^{L,\ell}\big(\mathbf{Z}^{L,\ell-1}\big) = 1 - \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{pred},\ell}\big(\mathbf{Z}^{L,\ell-1}\big) - \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{ans},\ell-1}\big(\mathbf{Z}^{L,\ell-1}\big),$$
$$\Delta^{L,\ell}\big(\mathbf{Z}^{L,\ell-1}\big) = \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{pred},\ell}\big(\mathbf{Z}^{L,\ell-1}\big) - \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{ans},\ell-1}\big(\mathbf{Z}^{L,\ell-1}\big).$$

Throught stage 2.3, we will focus on the attention gap $\Delta^{L,\ell}\big(\mathbf{Z}^{L,\ell-1}\big)$ instead of the gap of the attention matrices. We abbreviate. $\epsilon_{\mathsf{attn}}^{L,\ell}\big(\mathbf{Z}^{L,\ell-1}\big)$ and $\Delta^{L,\ell}\big(\mathbf{Z}^{L,\ell-1}\big)$ as $\epsilon_{\mathsf{attn}}^{L,\ell}$ and $\Delta^{L,\ell}$ for simplicity.

Induction F.3. *Given $\epsilon \ge \widetilde{\Omega}(\sigma_0)$, for $s \in \tau(\mathcal{X})$, let $T_{2,3,s}$ denote the first time that $\mathbb{E}\big[\epsilon_{\mathsf{attn}}^{2,2} \mid \tau(x_1) = s\big] \le \epsilon$. For all iterations $T_{2,2,s} \le t < T_{2,3,s}$, we have the following holds:*

(a) $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$ *and* $[\mathbf{Q}_{4,4}^{(t)}]_{s,s}$ *monotonically increases* $\le \widetilde{O}(1)$;

(b) $\Delta^{2,\ell} \ge 0$ *for any* $\mathbf{Z}^{2,\ell}$ *with* $\ell \in \{1,2\}$;

(c) $\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} \le 0.5 + \widetilde{c}_1$ *for some small constant* $\widetilde{c}_1 > 0$;

(d) for $(p,q) \in \{(4,3),(4,4)\}$, $\Big|[\mathbf{Q}_{p,q}^{(t)}]_{s,s'}\Big| \le O\Big(\frac{[\mathbf{Q}_{p,q}^{(t)}]_{s,s}}{d}\Big)$ *for* $s' \in \tau(\mathcal{X}) \neq s$; *other* $[\mathbf{Q}_{p,q}^{(t)}]_{s,s'} = 0$.

F.5.1 Attention and Logit Preliminaries

Lemma F.34. *If Induction F.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)* $\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{ans},0}^{(t)} \ge \Omega(1)$, *and* $\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{ans},0}^{(t)} > \mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},2}^{(t)}$;
2. *for* $\ell = 2$,
   - *(a)* $\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} \ge \Omega(1)$, $\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}, \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)} \le \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}$;
   - *(b)* $\Big|\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\Big| \le \widetilde{O}\big(\frac{1}{d}\big)$.

Moreover, given $\mathbf{Z}^{2,1}$ *and corresponding* $\mathbf{Z}^{2,0}$, $\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1}(\mathbf{Z}^{2,0}) \geq \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}(\mathbf{Z}^{2,1})$.

Lemma F.35. *If Induction F.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) *for* $j = j_1$, $\Lambda^{(t)}_{5,j_1,r} \ll -\varrho$ *for* $r \in \widehat{\mathfrak{A}}_{j_1} \setminus \{r_{g_1\cdot y_0}\}$;
   (b) *for* $j = j_1' \triangleq \tau\big(g_2(y_0)\big) = \tau(g_1(\widetilde{y}))$, $\Lambda^{(t)}_{5,j_1',r} \ll -\varrho$ *for* $r \in \widehat{\mathfrak{A}}_{j_1'} \setminus \{r_{g_2\cdot y_0}, r_{g_1\cdot\widetilde{y}}\}$;
   (c) *for other* $j \in \tau(\mathcal{Y})$*, assuming* $j = \tau(g_1(y))$ *for some* $y \neq y_0$*, then* $\Lambda^{(t)}_{5,j,r} \ll -\varrho$ *for* $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_1\cdot y}\}$.
2. $\ell = 2$*, if* $\mathbf{Z}^{2,\ell-1} \in \widetilde{\mathcal{E}}_2$*, then*
   (a) *for* $j = j_2$, $\Lambda^{(t)}_{5,j_2,r} \ll -\varrho$ *for* $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2\cdot y_1}\}$;
   (b) *for* $j = j_2' \triangleq \tau\big(g_2(y_0)\big)$, $\Lambda^{(t)}_{5,j_2',r} \ll -\varrho$ *for* $r \in \widehat{\mathfrak{A}}_{j_2'} \setminus \{r_{g_2\cdot y_0}\}$;
   (c) *for other* $j \in \tau(\mathcal{Y})$*, if* $j = \tau(g_2(y))$ *for some* $y \in \mathcal{Y}$*, then* $\Lambda^{(t)}_{5,j,r} \ll -\varrho$ *for* $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2\cdot y}\}$.

Lemma F.36. *If Induction F.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) $\Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} = \big(1 - 2\epsilon^{2,1}_{\mathsf{attn}}\big)B \pm O(\delta) \geq \Big(\frac{1}{3} + c_1\Big)B$;
   (b) $\Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} - \Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}} = 2\big(\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},2}\big)B \pm O(\delta) \geq 2(c_1 + c_2)B$;
   (c) *for* $y \neq y_0$, $\Lambda^{(t)}_{5,\tau(g_1(y)),r_{g_1\cdot y}} = \Big(2\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1} - 1\Big)B \pm O(\delta)$*, which is only activated if* $\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1} > \frac{1}{2}$
2. $\ell = 2$*, for* $\mathbf{Z}^{2,\ell-1} \in \widetilde{\mathcal{E}}_2$,
   (a) $\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = \big(1 - 2\epsilon^{2,2}_{\mathsf{attn}}\big)B \pm O(\delta) \geq 4c_5 B$;
   (b) $\Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}} = \Delta^{2,2}\cdot B \pm O(\delta)$*, and*
   $$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} - \Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}} = 2\big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\big)B \pm O(\delta);$$
   (c) *for* $y \neq y_0, y_1$, $\Lambda^{(t)}_{5,\tau(g_2(y)),r_{g_2\cdot y}} = \Big(2\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - 1\Big)B \pm O(\delta)$*, which is only activated if* $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} > \frac{1}{2}$.

Lemma F.37. *If Induction F.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* $\mathbf{Z}^{2,\ell-1} \in \mathcal{E}_1$,
   $$1 - \mathbf{logit}^{(t)}_{5,j_1} = \Theta(1)\cdot\mathbf{logit}^{(t)}_{5,j_1'} = \Theta\bigg(\frac{1}{d^{2\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{pred},2}\right)C_B}}\bigg).$$
2. *for* $\ell = 2$*, if* $\mathbf{Z}^{2,\ell-1} \in \widetilde{\mathcal{E}}_2$,
   $$\mathbf{logit}^{(t)}_{5,j_2'} = \Theta\bigg(\frac{1}{d^{2\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right)C_B} + d^{1-\Delta^{2,2}C_B}}\bigg),$$
   moreover,
   $$1 - \mathbf{logit}^{(t)}_{5,j_2} \geq \min\bigg\{\Omega(1), \Omega\Big(\frac{1}{d^{C_B\cdot(1-2\epsilon^{2,2}_{\mathsf{attn}})-1}}\Big)\bigg\}.$$

Lemma F.38. *If Induction F.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 $j = j_1$, $\widehat{\mathfrak{A}}_{j_1} = \{r_{g_1 \cdot y_0}\}$, $\Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} = B \pm O(\delta)$;*
   (b) *for $j \neq j_1 \in \tau(\mathcal{Y})$, assuming $j = \tau(g_1(y))$ for $y \neq y_0$, then $\Lambda^{(t)}_{5,j_1,r_{g_1\cdot y_0}} - \Lambda^{(t)}_{5,j,r_{g_1\cdot y}} \geq 2\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0} \cdot B \geq \Omega(B)$ and $\Lambda^{(t)}_{5,j,r} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_1\cdot y}\}$.*
2. *$\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \notin \widetilde{\mathcal{E}}_2$, then*
   (a) *if $g_1 = g_2 \wedge y_0 \neq y_1$,*
      i. *for $j = j_2$, $\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = (1 - \epsilon^{2,2}_{\mathsf{attn}}) \cdot B \pm O(\delta)$ and $\Lambda^{(t)}_{5,j_2,r} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2\cdot y_1}\}$;*
      ii. *for $j = \tau(g_2(y_0))$, $\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} - \Lambda^{(t)}_{5,j,r_{g_2\cdot y_0}} = 2(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0})B \pm O(\delta)$ and $\Lambda^{(t)}_{5,j,r} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2\cdot y_0}\}$;*
      iii. *for other $j \in \tau(\mathcal{Y})$, assuming $j = \tau(g_2(y))$ for some $y \neq y_0, y_1$, $\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} - \Lambda^{(t)}_{5,j,r_{g_2\cdot y}} = 2\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} B \pm O(\delta)$ and $\Lambda^{(t)}_{5,j,r} \ll -\varrho$ 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 $j = j_2$, $\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = (1 - \epsilon^{2,2}_{\mathsf{attn}}) \cdot B \pm O(\delta)$ and $\Lambda^{(t)}_{5,j_2,r} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_{j_2} \setminus \{r_{g_2\cdot y_1}\}$;*
      ii. *for $j = \tau(g_1(y_1)) = \tau(g_2(\widetilde{y}))$,*
      $$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} - \Lambda^{(t)}_{5,j,r_{g_1\cdot y_1}} = 2(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1})B \pm O(\delta)$$
      $$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} - \Lambda^{(t)}_{5,j,r_{g_2\cdot \widetilde{y}}} = 2(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1})B \pm O(\delta),$$
      where $r_{g_2\cdot\widetilde{y}}$ is only activated if $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} > \frac{1}{2}$. $\Lambda^{(t)}_{5,j,r} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_1\cdot y_1}, r_{g_2\cdot\widetilde{y}}\}$;
      iii. *for other $j \in \tau(\mathcal{Y})$, assuming $j = \tau(g_2(y))$ for $y \neq y_1, \widetilde{y}$, then $r_{g_2\cdot y}$ is only activated if $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} > \frac{1}{2}$ and*
      $$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} - \Lambda^{(t)}_{5,j_2,r_{g_2\cdot y}} = 2(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1})B \pm O(\delta).$$
      $\Lambda^{(t)}_{5,j,r} \ll -\varrho$ *for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2\cdot y}\}$.*
   (c) *if $g_1 = g_2 \wedge y_0 = y_1$,*
      i. *for $j = j_2$, $\widehat{\mathfrak{A}}_{j_2} = \{r_{g_2\cdot y_1}\}$, $\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = B \pm O(\delta)$;*
      ii. *for other $j \in \tau(\mathcal{Y})$, assuming $j = \tau(g_2(y))$,*
      $$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} - \Lambda^{(t)}_{5,j_2,r_{g_2\cdot y}} = 2(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1})B \pm O(\delta).$$
      and $\Lambda^{(t)}_{5,j,r} \ll -\varrho$ for $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2\cdot y}\}$.

Lemma F.39. *If Induction F.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 $1 - \mathbf{logit}^{(t)}_{5,j_1} = O\Big(\frac{1}{d^{2\mathbf{Attn}^{(t)}_{\mathsf{ans},0\to\mathsf{ans},0} C_B}}\Big) = O\Big(\frac{1}{\mathsf{poly}d}\Big)$.*
2. *$\ell = 2$, if $\mathbf{Z}^{2,\ell-1} \notin \widetilde{\mathcal{E}}_2$, then*
   (a) *if $g_1 = g_2 \wedge y_0 \neq y_1$, $\mathbf{logit}^{(t)}_{5,\tau(g_2(y_0))} = O\Big(\frac{1}{d^{2\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right) C_B}}\Big)$.*
   (b) *if $g_1 \neq g_2 \wedge y_0 = y_1$, $\mathbf{logit}^{(t)}_{5,\tau(g_1(y_1))} = O\Big(\frac{1}{d^{2\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\right) C_B}}\Big)$.*
   (c) *if $g_1 = g_2 \wedge y_0 = y_1$, $1 - \mathbf{logit}^{(t)}_{5,j_2} = O\Big(\frac{1}{d^{2\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\right) C_B}}\Big) = \frac{1}{\mathsf{poly}d}$.*

F.5.2 Gradient Lemma

Lemma F.40. *If Induction F.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*

$$\sum_{\ell=1}^{2} \Big[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \mathsf{Loss}_5^{2,\ell} \Big]_{s,s} \geq \Omega\Big(\frac{\epsilon \log d}{d^{(1-2\epsilon)C_B}} \Big).$$

Proof. By gradient decomposition, we have

$$\sum_{\ell=1}^{2} \Big[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \mathsf{Loss}_5^{2,\ell} \Big]_{s,s} = \sum_{\ell\in[2]} \sum_{\kappa\in\{i,ii,iii\}} \mathcal{N}_{s,4,\ell,\kappa}^{(t)}.$$

Firstly, for $\mathcal{N}_{s,4,2,i}^{(t)}$, by Lemma F.35 and Lemma F.38, we have

$$\begin{aligned}
\mathcal{N}_{s,4,2,i}^{(t)} &= \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \\
&\qquad \Big(\Big(\psi_{j_2, r_{g_2\cdot y_1}}(y_1) - \Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2} \Bigg] \\
&\quad + \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \\
&\qquad \Big(\Big(\psi_{j_2, r_{g_2\cdot y_1}}(y_1) - \Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2^c} \Bigg] \\
&\overset{(a)}{\geq} \Theta\Big(\frac{1}{d}\Big) \cdot \mathbb{E}\Big[\min\Big\{ \Omega(1), \Omega\Big(\frac{1}{d^{C_B\cdot(1-2\epsilon_{\mathsf{attn}}^{2,2})-1}} \Big) \Big\} \cdot 2\epsilon_{\mathsf{attn}}^{2,2} \cdot B \mid \tau(x_1) = s \Big] \\
&\geq \Omega\Big(\frac{\epsilon \log d}{d^{(1-2\epsilon)C_B}} \Big),
\end{aligned} \tag{67}$$

where (a) follows from Lemma F.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 $\sum_{\ell=1}^{2} \Big[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \mathsf{Loss}_5^{2,\ell} \Big]_{s,s}$, we only need to focus on the negative gradient from $\mathcal{N}_{s,4,1,ii}^{(t)}$ and $\mathcal{N}_{s,4,2,ii}^{(t)}$.

$$\begin{aligned}
&\mathcal{N}_{s,4,2,ii}^{(t)} = \\
&- \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} \cdot \mathbf{logit}_{5,j_2'}^{(t)} \cdot \\
&\qquad \Big(\Big(\psi_{j_2', r_{g_2\cdot y_0}}(y_1) - \Lambda_{5,j_2',r_{g_2\cdot y_0}}^{(t)} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2} \Bigg] \\
&- \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} \cdot \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)} \cdot \\
&\qquad \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \cdot \Big(\psi_{j,r}(y_1) - \Lambda_{5,j,r}^{(t)} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2^c} \Bigg] \\
&\geq -\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} \cdot \mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} \cdot \Big(\mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}}^{(t)}) \cdot \\
&\qquad \Big(\psi_{\tau(g_1(y_1)),r_{g_1\cdot y_1}}(y_1) - \Lambda_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}}^{(t)} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \mathbb{1}_{\{\tau(x_1)=s\}\cap\{g_1\neq g_2, y_0=y_1\}} \Bigg].
\end{aligned} \tag{68}$$

When $\mathbb{E}\Big[C_B \cdot (1-2\epsilon_{\mathsf{attn}}^{2,2}) \mid \tau(x_1)=s\Big] < 1$, by Lemma F.37 and Lemma F.39, we have

$$\begin{aligned}&\mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)}\mathbb{1}_{\{\tau(x_1)=s\}\cap\{g_1\neq g_2, y_0=y_1\}}\\ &\leq O\bigg(\frac{1}{d^{2\big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\big)C_B}}\bigg)\big(1-\mathbf{logit}_{5,j_2}^{(t)}\big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2}.\end{aligned}$$

During this time, $\epsilon_{\mathsf{attn}}^{2,2} \geq \Omega(1)$, thus we can lower bound $\mathcal{N}_{s,4,2,ii}^{(t)}$ by $\mathcal{N}_{s,4,2,ii}^{(t)} \geq -O\Big(\frac{1}{d^{\Omega(1)}\cdot\log d}\Big)\cdot \mathcal{N}_{s,4,2,i}^{(t)}$, 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 $\mathbb{E}\Big[C_B \cdot (1-2\epsilon_{\mathsf{attn}}^{2,2}) \mid \tau(x_1)=s\Big] \geq 1$, since $2\big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\big)C_B \geq (1-2\eta^{2,2})C_B$, we obtain

$$\mathbf{logit}_{\tau(g_1(y_1))}^{(t)}\mathbb{1}_{\{\tau(x_1)=s\}\cap\{g_1\neq g_2, y_0=y_1\}} \leq O(\frac{1}{d})\cdot(1-\mathbf{logit}_{j_2}^{(t)})\mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2}. \tag{69}$$

For the event $\{g_1\neq g_2, y_0=y_1\}$, if $\Lambda_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}}^{(t)}$ is still in the linear regime, we have

$$\Lambda_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}}^{(t)} = \big(1-2\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\cdot B \pm O(\delta) \geq \varrho,$$

which implies

$$\epsilon_{\mathsf{attn}}^{2,2} \geq 1-2\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2} \geq \Omega\Big(\frac{\varrho}{B}\Big). \tag{70}$$

Hence, putting (70) back to (68), and putting (69) back to (68), we can lower bound $\mathcal{N}_{s,4,2,ii}^{(t)}$ as follows $\mathcal{N}_{s,4,2,ii}^{(t)} \geq -O\big(\frac{1}{d}\big)\cdot\mathcal{N}_{s,4,2,i}^{(t)}$. If $\Lambda_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}}^{(t)}$ falls into the smoothed regime, we can upper bound $\mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}}^{(t)})$ by $O(\frac{\epsilon_{\mathsf{attn}}^{2,2}B}{\varrho})$, and then similarly, we can obatian $\mathcal{N}_{s,4,2,ii}^{(t)} \geq -O\big(\frac{1}{d}\big)\cdot\mathcal{N}_{s,4,2,i}^{(t)}$.

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.

□

Lemma F.41. *If Induction F.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*

$$\Big|\Big[-\nabla_{\mathbf{Q}_{4,p}^{(t)}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s'}\Big| \leq O\Big(\frac{1}{d}\Big)\Big|\Big[-\nabla_{\mathbf{Q}_{4,p}^{(t)}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s}\Big|.$$

F.5.3 Non-negative Gap

Lemma F.42. *If Induction F.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 $\widetilde{T}$ denote the first time that $\mathbb{E}[\Delta^{2,2}\mid\tau(x_1)=s] < \alpha$, where $\alpha = \frac{\epsilon^{\frac{1}{\alpha-2}}}{\mathsf{poly}d}$.

Following the analogous analysis as Lemma F.30, we have

$$\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}_{4,3}^{(\widetilde{T})}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s} + \sum_{\ell=1}^{2}\Big[\nabla_{\mathbf{Q}_{4,4}^{(\widetilde{T})}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s} = \sum_{\ell\in[2]}\sum_{\kappa\in\{i,ii,iii\}}\mathcal{N}_{s,3,\ell,\kappa}^{(\widetilde{T})} - \mathcal{N}_{s,4,\ell,\kappa}^{(\widetilde{T})}.$$

We can ignore the negligible difference introduced by $\mathcal{N}_{s,p,\ell,iii}^{(\widetilde{T})}$ for $p\in\{3,4\}$ and $\ell\in[2]$.

For $\ell = 1$, since $\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},0\to\mathsf{pred},1} \geq \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},0\to\mathsf{ans},0}$, and thus it is straightforward to see that $\mathcal{N}^{(\widetilde{T})}_{s,3,1,i} - \mathcal{N}^{(\widetilde{T})}_{s,4,1,i} \geq 0$. Futhermore, we have

$$
\begin{aligned}
&\mathcal{N}^{(\widetilde{T})}_{s,3,1,ii} - \mathcal{N}^{(\widetilde{T})}_{s,4,1,ii} \\
&= -\mathbb{E}\Bigg[\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},0\to\mathsf{pred},1}\cdot\mathbf{logit}^{(\widetilde{T})}_{5,j_1'}\cdot\mathbf{sReLU}'\big(\Lambda^{(t)}_{5,j_1',r_{g_2\cdot y_0}}\big) \\
&\qquad\Big(\Big(\psi_{j_1',r_{g_2\cdot y_0}}(g_1) - \Lambda^{(\widetilde{T})}_{5,j_1',r_{g_2\cdot y_0}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1}\Bigg] \qquad (71)\\
&+\mathbb{E}\Bigg[\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},0\to\mathsf{ans},0}\cdot\mathbf{logit}^{(\widetilde{T})}_{5,j_1'}\cdot\mathbf{sReLU}'\big(\Lambda^{(\widetilde{T})}_{5,j_1',r_{g_2\cdot y_0}}\big) \\
&\qquad\Big(\Big(\psi_{j_1',r_{g_2\cdot y_0}}(y_0) - \Lambda^{(\widetilde{T})}_{5,j_1',r_{g_2\cdot y_0}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1}\Bigg] \qquad (72)\\
&-\mathbb{E}\Bigg[\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},0\to\mathsf{pred},1}\cdot\sum_{y\neq y_0\in\mathcal{Y}}\mathbf{logit}^{(\widetilde{T})}_{5,\tau(g_1(y))}\cdot\mathbf{sReLU}'\big(\Lambda^{(\widetilde{T})}_{5,\tau(g_1(y)),r_{g_1\cdot y}}\big) \\
&\qquad\Big(\Big(\psi_{\tau(g_1(y)),r_{g_1\cdot y}}(g_1) - \Lambda^{(\widetilde{T})}_{5,\tau(g_1(y)),r_{g_1\cdot y}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1^c}\Bigg] \\
&+\mathbb{E}\Bigg[\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},0\to\mathsf{ans},0}\cdot\sum_{y\neq y_0\in\mathcal{Y}}\mathbf{logit}^{(\widetilde{T})}_{5,\tau(g_1(y))}\cdot\mathbf{sReLU}'\big(\Lambda^{(\widetilde{T})}_{5,\tau(g_1(y)),r_{g_1\cdot y}}\big) \\
&\qquad\Big(\Big(\psi_{\tau(g_1(y)),r_{g_1\cdot y}}(y_0) - \Lambda^{(\widetilde{T})}_{5,\tau(g_1(y)),r_{g_1\cdot y}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\mathcal{E}_1^c}\Bigg].
\end{aligned}
$$

By Lemma F.39, we obtain that $\mathcal{N}^{(\widetilde{T})}_{s,3,1,ii} - \mathcal{N}^{(\widetilde{T})}_{s,4,1,ii}$ is dominated by (71) + (72). Moreover, by Lemma F.37, we have

$$
\begin{aligned}
\mathbf{logit}^{(\widetilde{T})}_{5,\tau(g_1(y))}(\mathbf{Z}^{2,0}) &\geq \Omega\bigg(\frac{1}{d^{2\left(\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},0\to\mathsf{pred},1} - \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},0\to\mathsf{pred},2}\right)C_B}}\bigg) \\
&\geq \Omega\bigg(\frac{1}{d^{2\left(\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},1}\right)C_B}}\bigg). \qquad (73)
\end{aligned}
$$

For $\ell = 2$, we have

$$
\begin{aligned}
&\mathcal{N}^{(\widetilde{T})}_{s,3,2,i} - \mathcal{N}^{(\widetilde{T})}_{s,4,2,i} \\
&\geq \mathbb{E}\Bigg[\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2})\cdot \\
&\qquad\Big(\Big(\psi_{j_2,r_{g_2\cdot y_1}}(g_2) - \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2}\Bigg] \\
&-\mathbb{E}\Bigg[\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2})\cdot \\
&\qquad\Big(\Big(\psi_{j_2,r_{g_2\cdot y_1}}(y_1) - \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2}\Bigg] \\
&\geq \Omega\big(\alpha\epsilon\log d\big)\mathbb{E}\Big[(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2})\cdot\mathbb{1}_{\tau(x_1)=s}\Big], \qquad (74)
\end{aligned}
$$

where the last inequality follows from the definition of $\widetilde{T}$.

$$
\begin{aligned}
&\mathcal{N}_{s,3,2,ii}^{(\widetilde{T})} - \mathcal{N}_{s,4,2,ii}^{(\widetilde{T})} \\
&= -\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})} \cdot \mathbf{logit}_{5,j_2'}^{(\widetilde{T})} \cdot \mathbf{sReLU}'\big(\Lambda_{5,j_2',r_{g_2\cdot y_0}}^{(\widetilde{T})}\big) \\
&\qquad \Big(\Big(\psi_{j_2',r_{g_2\cdot y_0}}(g_2) - \Lambda_{5,j_2',r_{g_2\cdot y_0}}^{(\widetilde{T})} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2}\Bigg] \qquad (75) \\
&+ \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(\widetilde{T})} \cdot \mathbf{logit}_{5,j_2'}^{(\widetilde{T})} \cdot \mathbf{sReLU}'\big(\Lambda_{5,j_2',r_{g_2\cdot y_0}}^{(\widetilde{T})}\big) \\
&\qquad \Big(\Big(\psi_{j_2',r_{g_2\cdot y_0}}(y_1) - \Lambda_{5,j_2',r_{g_2\cdot y_0}}^{(\widetilde{T})} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2}\Bigg] \qquad (76) \\
&- \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})} \cdot \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(\widetilde{T})} \cdot \\
&\qquad \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(\widetilde{T})}) \cdot \Big(\psi_{j,r}(g_2) - \Lambda_{5,j,r}^{(\widetilde{T})} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2^c}\Bigg] \qquad (77) \\
&+ \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(\widetilde{T})} \cdot \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(\widetilde{T})} \cdot \\
&\qquad \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(\widetilde{T})}) \cdot \Big(\psi_{j,r}(y_1) - \Lambda_{5,j,r}^{(\widetilde{T})} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2^c}\Bigg]. \qquad (78)
\end{aligned}
$$

Notice that by Lemma F.36 and the definition of $\widetilde{T}$, we have

$$
|(75)+(76)| \le O\big((\alpha\log d)^{q-1}\big)\cdot\mathbb{E}\Big[\Big(1-\mathbf{logit}_{5,j_2}^{(\widetilde{T})}\Big)\cdot\mathbb{1}_{\tau(x_1)=s}\Big],
$$

which is dominated by (74) due to the choice of α.

Furthermore, by Lemma F.38, for $\widetilde{\mathcal{E}}_2^c$, 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

$$
\begin{aligned}
&(77)+(78) \\
&\ge -\mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})} \cdot \mathbf{logit}_{5,\tau(g_2(y_0))}^{(\widetilde{T})} \cdot \mathbf{sReLU}'\big(\Lambda_{5,\tau(g_2(y_0)),r_{g_2\cdot y_0}}^{(\widetilde{T})}\big) \\
&\Big(\Big(\psi_{\tau(g_2(y_0)),r_{g_2\cdot y_0}}(g_2) - \Lambda_{5,\tau(g_2(y_0)),r_{g_2\cdot y_0}}^{(\widetilde{T})} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\{g_1=g_2\wedge y_0\neq y_1\}}\Bigg] \\
&+ \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(\widetilde{T})} \cdot \mathbf{logit}_{5,\tau(g_2(y_0))}^{(\widetilde{T})} \cdot \mathbf{sReLU}'\big(\Lambda_{5,\tau(g_2(y_0)),r_{g_2\cdot y_0}}^{(\widetilde{T})}\big) \\
&\Big(\Big(\psi_{\tau(g_2(y_0)),r_{g_2\cdot y}}(y_1) - \Lambda_{5,\tau(g_2(y_0)),r_{g_2\cdot y_0}}^{(\widetilde{T})} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\{g_1=g_2\wedge y_0\neq y_1\}}\Bigg].
\end{aligned}
$$

By Lemma F.39, we have

$$
\mathbf{logit}_{5,\tau(g_2(y_0))}^{(\widetilde{T})}(\mathbf{Z}^{2,1}) = O\bigg(\frac{1}{d^{2\left(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(\widetilde{T})} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(\widetilde{T})}\right)C_B}}\bigg).
$$

Since $\mathbb{E}[\Delta^{2,2} \mid \tau(x_1)=s] < \alpha$, where α is sufficiently small, then combining with (73), we have

$$
\mathbf{logit}_{5,\tau(g_2(y_0))}^{(\widetilde{T})}(\mathbf{Z}^{2,1}) = O(1)\cdot\mathbf{logit}_{5,\tau(g_2(y_0))}^{(\widetilde{T})}(\mathbf{Z}^{2,0}).
$$

Therefore, since $\widetilde{\mathcal{E}}_2^c$ happens with probability $O(\frac{1}{\log d})$, we obtain that

$$(77)+(78) \geq -O(\frac{1}{\log d})\Big(\mathcal{N}_{s,3,1,ii}^{(t)} - \mathcal{N}_{s,4,1,ii}^{(t)}\Big).$$

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.,

$$\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}_{4,3}^{(\widetilde{T})}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s} + \sum_{\ell=1}^{2}\Big[\nabla_{\mathbf{Q}_{4,4}^{(\widetilde{T})}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s} > 0,$$

which completes the proof. □

F.5.4 Upper Bound for Q

Lemma F.43. *If Induction F.3 holds for all iterations $\in [T_{2,2,s}, t)$, given $s \in \tau(\mathcal{X})$, then at time t, we have $[\mathbf{Q}_{4,p}^{(t)}]_{s,s} \leq \widetilde{O}(1)$ 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 $\widetilde{T}$. Then by direct calculations, we have

$$\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},2}^{(t)} \leq O\bigg(\frac{1}{e^{\Omega(\log^{1+c} d)}}\bigg),$$
$$\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)} + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)} \leq O\bigg(\frac{1}{e^{\Omega(\log^{1+c} d)}}\bigg).$$

Moreover, by Lemma F.37, we can simply bound the logits as follows:

$$1-\mathbf{logit}_{j_1}^{(t)} \leq O\bigg(\frac{1}{d^{C_B\left(1-2\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},2}^{(t)}\right)-1}}\bigg) \leq O\Big(\frac{1}{d^{C_B/2-1}}\Big) \text{ for } \mathbf{Z}^{2,0} \in \mathcal{E}_1$$
$$1-\mathbf{logit}_{j_2}^{(t)} \leq O\bigg(\frac{1}{d^{C_B\left(1-2\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}-2\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\right)-1}}\bigg)$$
$$+ O\bigg(\frac{1}{d^{2\left(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\right)C_B}}\bigg) \leq O\Big(\frac{1}{d^{C_B/2-1}}\Big) \text{ for } \mathbf{Z}^{2,1} \in \widetilde{\mathcal{E}}_2$$

Thus, by focusing on $\mathcal{N}_{s,3,1,i}$ and $\mathcal{N}_{s,4,1,i}$, we have

$$\bigg|\sum_{\ell=1}^{2}\Big[-\nabla_{\mathbf{Q}_{4,3}^{(t)}}\mathsf{Loss}_5^{2,\ell}\Big]_{s,s}\bigg| \leq \frac{1}{d}\cdot O\Big(\frac{1}{d^{C_B/2-1}}\Big)\cdot O\Big(\frac{1}{e^{\Omega(\log^{1+c} d)}}\Big)\cdot B \leq O\bigg(\frac{\log d}{e^{\Omega(\log^{1+c} d)} d^{C_B/2}}\bigg),$$

which implies

$$\mathbf{Q}_{4,3}^{(T)} \leq \mathbf{Q}_{4,3}^{(\widetilde{T})} + O\bigg(\frac{T\log d}{e^{\Omega(\log^{1+c} d)} d^{C_B/2}}\bigg) \leq \mathbf{Q}_{4,3}^{(\widetilde{T})} + O\bigg(\frac{\mathsf{poly} d\cdot \log d}{e^{\Omega(\log^{1+c} d)} d^{C_B/2}}\bigg) \leq O(\log^{1+c} d).$$

□

F.5.5 Attention Upper Bound

Lemma F.44. *If Induction F.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 $\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} \leq 0.5 + \widetilde{c}_1$, where $\widetilde{c}_1 > 0$ is some small constant.*

Proof. Let $\widetilde{T}$ denote the first time that $\mathbb{E}[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} \mid \tau(x_1)=s]$ reaches $0.5+\widetilde{c}$. where $\widetilde{c} > 0$ is some small constant s.t., $2\widetilde{c}\cdot C_B > 1$. At $\widetilde{T}$, we have $\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{pred},1}^{(\widetilde{T})}\mathbb{1}_{\tau(x_0)=s} \geq \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})}\mathbb{1}_{\tau(x_1)=s}$; moreover, $\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(\widetilde{T})}$ and $\mathbf{Attn}_{\mathsf{ans},0\to\mathsf{ans},0}^{(\widetilde{T})}$ is still at the constant level.

- For $\ell = 1$, by Lemma F.35 and Lemma F.38, for $y \in \mathcal{Y} \setminus \{y_0\}$, we have only $r_{g_1 \cdot y}$ has been activted to the linear regime for the prediction $\tau(g_1(y))$. Furthermore, we obtain

$$\sum_{y \in \mathcal{Y} \setminus \{y_0\}} \mathbf{logit}^{(\widetilde{T})}_{5,\tau(g_1(y))} \geq \frac{\log d \cdot e^{2\widetilde{c}C_B \log d}}{\log d \cdot e^{2\widetilde{c}C_B \log d} + O(d)} \Big(1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_1}\Big)$$
$$= (1 - o(1)) \cdot \Big(1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_1}\Big).$$

Thus,

$$\Big[- \nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,3}} \mathsf{Loss}^{2,1}_5 \Big]_{s,s}$$
$$= \mathbb{E}\Bigg[\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},0\to\mathsf{pred},1} \cdot \Bigg(\Big(1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_1}\Big) \cdot \Big(\Big(\psi_{j_1, r_{g_1 \cdot y_0}}(g_1) - \Lambda^{(\widetilde{T})}_{5,j_1,r_{g_1\cdot y_0}} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big)$$
$$- \sum_{y \in \mathcal{Y}\setminus\{y_0\}} \mathbf{logit}^{(\widetilde{T})}_{5,\tau(g_1(y))} \cdot \Big(\Big(\psi_{\tau(g_1(y)), r_{g_1\cdot y}}(g_1) - \Lambda^{(\widetilde{T})}_{5,\tau(g_1(y)),r_{g_1\cdot y}} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big)$$
$$\pm \frac{1}{\mathsf{poly} d} \Big(1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_1}\Big) \cdot \widetilde{O}(\sigma_0^q) \Bigg) \mathbb{1}_{\tau(x_0)=s} \Bigg]$$
$$\leq \mathbb{E}\Bigg[\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},0\to\mathsf{pred},1} \cdot \Bigg(\Big(1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_1}\Big) \cdot \Big(\Big(2\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},0\to\mathsf{pred},2} \cdot B \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big)$$
$$- \sum_{y \in \mathcal{Y}\setminus\{y_0\}} \mathbf{logit}^{(\widetilde{T})}_{5,\tau(g_1(y))} \cdot \Big(\Big(2\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},0\to\mathsf{ans},0} \cdot B \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big)$$
$$\pm \frac{1}{\mathsf{poly} d} \Big(1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_1}\Big) \cdot \widetilde{O}(\sigma_0^q) \Bigg) \mathbb{1}_{\tau(x_0)=s} \Bigg] < 0$$

- For $\ell = 2$, for $\mathbf{Z}^{2,1} \in \widetilde{\mathcal{E}}_2 \cup \{g_1 = g_2, y_0 \neq y_1 x\}$, we have

$$\sum_{y \in \mathcal{Y}\setminus\{y_0,y_1\}} \mathbf{logit}^{(\widetilde{T})}_{5,\tau(g_2(y))} \geq \frac{\log d \cdot e^{2\widetilde{c}C_B \log d}}{\log d \cdot e^{2\widetilde{c}C_B \log d} + O(d)} \Big(1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_2} - \mathbf{logit}^{(\widetilde{T})}_{5,j_2'}\Big)$$
$$= (1 - o(1)) \cdot \Big(1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_2} - \mathbf{logit}^{(\widetilde{T})}_{5,j_2'}\Big)$$

Else, we have

$$\sum_{y \in \mathcal{Y}\setminus\{y_1\}} \mathbf{logit}^{(\widetilde{T})}_{5,\tau(g_2(y))} = (1 - o(1)) \cdot \Big(1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_2} - \mathbf{logit}^{(\widetilde{T})}_{5,j_2'}\Big)$$

Therefore, similar to $\ell = 1$, we obtain $\Big[- \nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,3}} \mathsf{Loss}^{2,2}_5 \Big]_{s,s} < 0$.

Combing the above two cases, we have $\sum_{\ell=1}^{2} \Big[- \nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,3}} \mathsf{Loss}^{2,\ell}_5 \Big]_{s,s} < 0$. It is also clear from Lemma F.40 that $\sum_{\ell=1}^{2} \Big[- \nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,4}} \mathsf{Loss}^{2,\ell}_5 \Big]_{s,s} \geq 0$. Hence $\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2}$ cannot further grow once it reaches $0.5 + \widetilde{c}$. □

F.5.6 Decreasing Gap at the End of Convergence

Let $\widetilde{T}$ denote the first time that $\mathbb{E}[\epsilon^{2,2}_{\mathsf{attn}} \mid \tau(x_1) = s] \leq 3\epsilon$, if $\mathbb{E}[\Delta^{2,2} \mid \tau(x_1) = s] \leq O\big(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d}\big)$, then we can let $T^\star = \widetilde{T}$ and stop the training. Otherwise, we have $\mathbb{E}[\Delta^{2,2} \mid \tau(x_1) = s] \geq \Omega\big(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d}\big)$. Following the similar argument as in Lemma F.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

$$
\begin{aligned}
&\Big[-\nabla_{\mathbf{Q}_{4,3}^{(\widetilde{T})}} \mathsf{Loss}_5^{2,2} \Big]_{s,s} \\
&= \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})} \cdot \bigg(\Big(1 - \mathbf{logit}_{5,j_2}^{(\widetilde{T})}\Big) \cdot \Big(\Big(\psi_{j_2, r_{g_2\cdot y_1}}(g_2) - \Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(\widetilde{T})} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \\
&\qquad \pm \frac{1}{\mathsf{poly}d} \Big(1 - \mathbf{logit}_{5,j_2}^{(\widetilde{T})}\Big) \cdot \widetilde{O}(\sigma_0^q) \bigg) \mathbb{1}_{\{\tau(x_0)=s\}} \Bigg] \\
&- \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})} \cdot \bigg(\sum_{y\in\mathcal{Y}\setminus\{y_1\}} \mathbf{logit}_{5,\tau(g_2(y))}^{(\widetilde{T})} \cdot \mathbf{sReLU}'\big(\Lambda_{5,\tau(g_2(y)),r_{g_2\cdot y}}^{(\widetilde{T})}\big) \\
&\qquad \cdot \Big(\Big(\psi_{\tau(g_2(y)),r_{g_2\cdot y}}(g_2) - \Lambda_{5,\tau(g_2(y)),r_{g_2\cdot y}}^{(\widetilde{T})} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \bigg) \mathbb{1}_{\{\tau(x_0)=s\}\cap\widetilde{\mathcal{E}}_2} \Bigg] \\
&- \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} \cdot \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)} \cdot \\
&\qquad \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \cdot \Big(\psi_{j,r}(g_2) - \Lambda_{5,j,r}^{(t)} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2^c} \Bigg] \\
&\le \mathbb{E}\bigg[\Big(1 - \mathbf{logit}_{5,j_2}^{(t)}\Big) \cdot O(\epsilon \log d) \mathbb{1}_{\tau(x_1)=s} \bigg].
\end{aligned}
$$

Turn to $\mathbf{Q}_{4,4}$, we have

$$
\begin{aligned}
&\Big[-\nabla_{\mathbf{Q}_{4,4}^{(\widetilde{T})}} \mathsf{Loss}_5^{2,2} \Big]_{s,s} \\
&= \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(\widetilde{T})} \cdot \bigg(\Big(1 - \mathbf{logit}_{5,j_2}^{(\widetilde{T})}\Big) \cdot \Big(\Big(\psi_{j_2, r_{g_2\cdot y_1}}(y_1) - \Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(\widetilde{T})} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \\
&\qquad \pm \frac{1}{\mathsf{poly}d} \Big(1 - \mathbf{logit}_{5,j_2}^{(\widetilde{T})}\Big) \cdot \widetilde{O}(\sigma_0^q) \bigg) \mathbb{1}_{\{\tau(x_0)=s\}} \Bigg] \\
&- \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(\widetilde{T})} \cdot \bigg(\sum_{y\in\mathcal{Y}\setminus\{y_1\}} \mathbf{logit}_{5,\tau(g_2(y))}^{(\widetilde{T})} \cdot \mathbf{sReLU}'\big(\Lambda_{5,\tau(g_2(y)),r_{g_2\cdot y}}^{(\widetilde{T})}\big) \\
&\qquad \cdot \Big(\Big(\psi_{\tau(g_2(y)),r_{g_2\cdot y}}(y_1) - \Lambda_{5,\tau(g_2(y)),r_{g_2\cdot y}}^{(\widetilde{T})} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \bigg) \mathbb{1}_{\{\tau(x_0)=s\}\cap\widetilde{\mathcal{E}}_2} \Bigg] \qquad (79) \\
&- \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} \cdot \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}_{5,j}^{(t)} \cdot \\
&\qquad \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \cdot \Big(\psi_{j,r}(y_1) - \Lambda_{5,j,r}^{(t)} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \mathbb{1}_{\{\tau(x_1)=s\}\cap\widetilde{\mathcal{E}}_2^c} \Bigg] \qquad (80)
\end{aligned}
$$

For (79), we have

$$
\begin{aligned}
(79) &\ge \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(\widetilde{T})} \cdot \bigg(-\mathbf{logit}_{5,j_2'}^{(\widetilde{T})} \cdot \\
&\qquad \min\Big\{ \Omega(\log d), \Omega\Big(\big((\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(\widetilde{T})}) \log d \big)^{q-1} \log d \Big) \Big\} \bigg) \mathbb{1}_{\{\tau(x_0)=s\}\cap\widetilde{\mathcal{E}}_2} \Bigg] \\
&\ge \mathbb{E}\bigg[\Big(1 - \mathbf{logit}_{5,j_2}^{(\widetilde{T})}\Big) \cdot \Omega\Big(\frac{1}{d}\Big) \cdot
\end{aligned}
$$

$$\min\left\{\Omega(\log d), \Omega\Big(\big((\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},1})\log d\big)^{q-1}\log d\Big)\right\}\mathbb{1}_{\{\tau(x_0)=s\}\cap\widetilde{\mathcal{E}}_2}\Bigg].$$

Moreover, for (80), we have

$$\begin{aligned}
(80) &\geq -\mathbb{E}\Bigg[\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot\mathbf{logit}^{(\widetilde{T})}_{5,\tau(g_1(y_1))}\cdot\mathbf{sReLU}'\big(\Lambda^{(\widetilde{T})}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}}\big)\\
&\quad\Big(\Big(\psi_{\tau(g_1(y_1)),r_{g_1\cdot y_1}}(y_1) - \Lambda^{(\widetilde{T})}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\{\tau(x_1)=s\}\cap\{g_1\neq g_2\wedge y_0=y_1\}}\Bigg]\\
&\geq -\mathbb{E}\Bigg[\Big(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2}\Big)\cdot O\Big(\frac{1}{d}\Big)\cdot O\Big(\frac{1}{\log d}\Big)\cdot O\Big((\epsilon\log d)^{q-1}\Big)\mathbb{1}_{\{\tau(x_0)=s\}\cap\widetilde{\mathcal{E}}_2}\Bigg],
\end{aligned}$$

where the last inequality holds since $\Lambda^{(\widetilde{T})}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}} \leq \Big(\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)B \leq O(\epsilon\log d)$. Since $\mathbb{E}\Big[\Delta^{2,2}\mid\tau(x_1)=s\Big] \geq \Omega\big(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d}\big)$ we have

$$(79) \geq d^{0.01}\cdot\Big|\Big[-\nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,3}}\mathsf{Loss}^{2,2}_5\Big]_{s,s}\Big|,$$

and thus (79) $\gg$ (80). Thus, for $\epsilon \ll \frac{1}{d}$, if the attention gap does not decrease to the level of $O\Big(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d}\Big)$, $[\mathbf{Q}_{4,4}]_{s,s}$ 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

$$\begin{aligned}
\epsilon^{2,2}_{\mathsf{attn}} = 1 - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} &= \frac{O(1)}{O(1)+e^{[\mathbf{Q}^{(t)}_{4,4}]_{s,s}}+e^{[\mathbf{Q}^{(t)}_{4,3}]_{s,s}}}\\
&\geq \frac{O(1)}{O(1)+2e^{[\mathbf{Q}^{(\widetilde{T})}_{4,3}]_{s,s}}} \geq \frac{1}{2}\cdot 3\epsilon > \epsilon.
\end{aligned}$$

This implies, we can find some time between $\widetilde{T}$ and $T_{2,3,s}$, s.t., the gap will decrease to $O\Big(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d}\Big)$. We denote this time as $T^\star$ and stop the training.

F.5.7 At the End of the Training

Putting everything together, we have that at the end of the training, we have

Lemma F.45. *Given $s\in\tau(\mathcal{X})$, at $T^\star = \widetilde{O}\Big(\frac{d^{(1-2\epsilon)C_B}}{\eta\epsilon}\Big)$, if $\epsilon = o(\frac{1}{d^{1.01}})$, we have*

- *(a) Attention convergence: $\epsilon^{2,\ell}_{\mathsf{attn}} \leq O(\epsilon)$ for $\ell\in[2]$, and $\mathbb{E}\big[\Delta^{2,2}\mid\tau(x_1)=s\big] \leq O(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d})$;*
- *(b) $\Big[\mathbf{Q}^{(T_{2,3,s})}_{4,p}\Big]_{s,s'} \geq \Omega(\log d)$ for $p\in\{3,4\}$ if $s=s'\in\tau(\mathcal{X})$, otherwise, $\Big[\mathbf{Q}^{(T_{2,3,s})}_{4,p}\Big]_{s,s'} \leq \widetilde{O}(\frac{1}{d})$;*
- *(c) Loss convergence: $\sum_{\ell=1}^2 \mathsf{Loss}^{2,\ell}_5 \leq \frac{1}{\mathsf{poly}d}$.*

F.6 Proof of Main Theorem

Theorem F.1 (Restatement of Theorem 4.1)**.** *Under Assumptions 3.1, 3.2, A.2, and 4.1, for every $L \leq \frac{1}{\sigma_0 d^{0.01}}$, the transformer model $F^{(T_1+T_2)}$ obtained by Algorithm 1 with learning rate $\eta = \frac{1}{\mathsf{poly}(d)}$, and stage 1 and 2 iteration $T_1 = \widetilde{O}\Big(\frac{1}{\eta(\sigma_0)^{q-2}}\Big)$, $T_2 = \widetilde{O}\Big(\frac{\mathsf{poly}(d)}{\eta\sigma_0}\Big)$ satisfies*

$$\mathrm{Acc}_L\big(F^{(T_1+T_2)}\big) \geq 1 - \frac{1}{\mathsf{poly}(d)},$$

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

Proof. By Lemma F.45, at the end of Stage 2 training, we have $\left[\mathbf{Q}_{4,p}^{(T_{2,3,s})}\right]_{s,s} \ge \Omega(\log d)$ for all $p \in \{3,4\}$ and $s \in \tau(\mathcal{X})$.

Therefore, for task $\mathcal{T}^L$ with $L \le \mathsf{poly}(d)$, we obtain

$$\epsilon_{\mathsf{attn}}^{L,\ell} \le \frac{O(1)\cdot L}{O(1)\cdot L + e^{\left[\mathbf{Q}_{4,3}^{(T_{2,3,s})}\right]_{s,s}} + e^{\left[\mathbf{Q}_{4,4}^{(T_{2,3,s})}\right]_{s,s}}} = o(1).$$

Moreover, we have

$$\Delta^{L,\ell} \le \Delta^{2,1} = \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(T_{2,3,s})} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(T_{2,3,s})} \le o(1).$$

These together guarantee that

$$1 - \mathbf{logit}_{5,\tau(g_{\ell+1}(y_\ell))}(F^{(T^\star)}, \mathbf{Z}^{(L,\ell)}) \le \frac{O(1)\cdot d + e^{o(1)}}{O(1)\cdot d + e^{o(1)} + e^{\Omega(\log d)}} \le \frac{1}{\mathsf{poly}(d)},$$

which implies

$$\mathrm{Acc}_L\big(F^{(T^\star)}\big) \ge 1 - \frac{1}{\mathsf{poly}(d)}.$$

□

G 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 4.2, we assume that $\mathcal{Y} = \{0, 1, \ldots, n_y - 1\}$, and let $\mathcal{G}$ be the symmetry group for $\mathcal{Y}$, with order $|\mathcal{G}| = n_y! = \Theta(\mathrm{polylog} d) \gg \frac{1}{\varrho}$, 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 $\mathsf{Loss}_5^{2,2}$. Given $s \in \tau(\mathcal{X})$, let $\mathsf{Loss}_{5,s}^{2,2} = -\mathbb{E}\Big[\log p_F(\mathbf{Z}_{\mathsf{ans},2,5} \mid \mathbf{Z}^{2,1}) \,\Big|\, \tau(x_1) = s\Big]$. 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 $\mathsf{Loss}_{5,s}^{2,2}$ in what follows.

G.1 Gradient Computations

We start with the gradient computations for the attention layer.

Notations for gradient expressions. We firsr introduce some notations for the gradients of the attention layer. For $1 \le \ell \le L$, given $\mathbf{Z}^{L,\ell-1}$ and $\mathbf{k} \in \mathcal{I}^{L,\ell-1}$, define

$$\Xi_{\ell,i,\mathbf{k}}^{L}(\mathbf{Z}^{L,\ell-1}) \triangleq \sum_{j\in[d]} \mathcal{E}_{i,j}(\mathbf{Z}^{L,\ell-1}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{i,j,r}(\mathbf{Z}^{L,\ell-1})\big)\langle \mathbf{W}_{i,j,r}, \mathbf{Z}_{\mathbf{k}}\rangle, \quad i \in [5]. \tag{81}$$

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 G.1 (Gradients of $\mathbf{Q}$)**.** For any $p, q \in [5]$, we have

$$-\nabla_{\mathbf{Q}_{p,q}} \mathsf{Loss}^{L} = \sum_{\ell=1}^{L}\sum_{i\in[5]} -\nabla_{\mathbf{Q}_{p,q}} \mathsf{Loss}_i^{L,\ell}, \quad \text{where}$$

$$-\nabla_{\mathbf{Q}_{p,q}} \mathsf{Loss}_i^{L,\ell} =$$

$$\mathbb{E}\left[\sum_{\mathbf{k}\in\mathcal{I}^{L,\ell-1}} \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathbf{k}} \cdot \left(\Xi_{\ell,i,\mathbf{k}}^{L} - \sum_{\mathbf{k}'\in\mathcal{I}^{L,\ell-1}} \mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathbf{k}'}\Xi_{\ell,i,\mathbf{k}'}^{L}\right)\mathbf{Z}_{\mathsf{ans},\ell-1,p}\mathbf{Z}_{\mathbf{k},q}^{\top}\right].$$

Lemma G.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*

$$\Big[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}_5^{2,2}\Big]_{s,s} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot \Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},2}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Bigg)\mathbb{1}_{s=\tau(x_1)}\Bigg].$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\Big[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}_5^{2,2}\Big]_{s,s'} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot \Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},1}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Bigg)\mathbb{1}_{s=\tau(x_1),s'=\tau(x_0)}\Bigg].$$

Lemma G.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*

$$\Big[-\nabla_{\mathbf{Q}_{4,4}} \mathsf{Loss}_5^{2,2}\Big]_{s,s} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot$$

$$\Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{ans},1}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Big)\mathbb{1}_{s=\tau(x_1)}\Big].$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,4}]_{s,s'}$ *with* $s \neq s'$*, we have*

$$\Big[-\nabla_{\mathbf{Q}_{4,4}} \mathsf{Loss}_5^{2,2}\Big]_{s,s'} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\Bigg(\sum_{j\in[d]}\mathcal{E}_{5,j}(\mathbf{Z}^{2,1})\sum_{r\in[m]}\mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot$$
$$\Big(\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{ans},0}\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Big)\mathbb{1}_{s=\tau(x_1),s'=\tau(x_0)}\Bigg].$$

G.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 $\mathrm{Fiber}_{j,y}$, and the set of feature combinations for predicting $y = \tau^{-1}(j)$ are defined as

$$\mathrm{Fiber}_{j,y} = \{g \in \mathcal{G} \mid \tau\big(g(y)\big) = j\}, \quad \mathfrak{F}_j = \Big\{(\mathrm{Fiber}_{j,y}, y) \in 2^{\mathcal{G}} \times \mathcal{Y}\Big\}.$$

As established in Lemma D.10 and Theorem D.1, at the end of stage 1.1, we have

- **Sparse activations**: For $j \in \tau(\mathcal{Y})$, let $\phi = (\mathrm{Fiber}_{j,y}, y) \in \mathfrak{F}_j$, then there exists exactly one *activated* neuron $r \in [m]$ such that when $g_1 = g \in \mathrm{Fiber}_{j,y}, y_0 = y$ happens:
$$\Lambda_{5,j,r}^{(T_{1.1})} \geq B - O(d^{0.01}\sigma_0), \qquad \Big|n_y\psi_{j,r}^{(t)}(y) - \psi_{j,r}^{(t)}(g)\Big| \leq O(\delta)$$
$$\Lambda_{5,j,r'}^{(T_{1.1})} \leq O(d^{-\Omega(1)}) \quad \forall r' \neq r,$$
for some $\delta = O\Big(\Big(\big(d^{0.01}\sigma_0\big)^{q-2} d/\boldsymbol{\lambda}\Big)^{1/(q-1)}\Big)$.
- **Cancellation of incorrect features**: For $j \in \tau(\mathcal{Y})$, let $\phi = (\mathrm{Fiber}_{j,y}, y) \in \mathfrak{F}_j$, and let the $r \in [m]$ be the activated neuron, then for any $g' \notin \mathrm{Fiber}_{j,y}$, and any $y' \neq y \in \mathcal{Y}$, we have
$$\Big|\psi_{j,r}^{(T_{1.1})}(g) + \psi_{j,r}^{(T_{1.1})}(y')\Big| \leq O(\delta) \quad \text{and} \quad \Big|\psi_{j,r}^{(T_{1.1})}(g') + \psi_{j,r}^{(T_{1.1})}(y)\Big| \leq O(\delta).$$

Notations for activated neurons. We denote by $r_{j,y}$ the unique activated neuron corresponding to $\phi = (\mathrm{Fiber}_{j,y}, y) \in \mathfrak{F}_j$. For any $g \in \mathrm{Fiber}_{j,y}$, we also write $r_{g\cdot y}$ for the same neuron $r_{j,y}$. Note that $r_{g_1\cdot y} = r_{g_2\cdot y}$ for distinct $g_1, g_2 \in \mathrm{Fiber}_{j,y}$. Moreover, define

$$\mathfrak{A} \triangleq \cup_{j\in\tau(\mathcal{Y})}\mathfrak{A}_j, \text{ where } \mathfrak{A}_j \triangleq \{r_{j,y} \mid y \in \mathcal{Y}\}.$$

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 cluases. Similarly $\widehat{\mathcal{Y}} = \cup\{y_{\ell'}\}_{\ell'=0}^{\ell-1}$. Then define $\widehat{\mathfrak{A}}_j(\mathbf{Z}^{L,\ell-1}) = \Big\{r_{g\cdot y} \mid g \in \mathrm{Fiber}_{j,y} \wedge \Big(g \in \widehat{\mathcal{G}}(\mathbf{Z}^{L,\ell-1}) \vee y \in \widehat{\mathcal{Y}}\Big)\Big\}$. 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 G.3 (Properties of target feature magnitude)**.** *Given* $j \in \tau(\mathcal{Y})$ *and* $y \in \mathcal{Y}$ *then for* $g \in \mathrm{Fiber}_{j,y}$ *, the following properties hold.*

$$\psi_{j,r_{g\cdot y}}(g) + \psi_{j,r_{g\cdot y}}(y) \geq 2B - O(d^{0.01}\sigma_0), \quad \Big|n_y\psi_{j,r}^{(t)}(y) - \psi_{j,r}^{(t)}(g)\Big| \leq O(\delta), \tag{82}$$
$$\big|\psi_{j,r_{g\cdot y}}(g) + \psi_{j,r_{g\cdot y}}(y')\big| \leq O(\delta), \psi_{j,r_{g\cdot y}}(y') < 0 \quad \textit{for all } y' \neq y, \tag{83}$$
$$\big|\psi_{j,r_{g\cdot y}}(g') + \psi_{j,r_{g\cdot y}}(y)\big| \leq O(\delta), \psi_{j,r_{g\cdot y}}(g') < 0 \quad \textit{for all } g' \notin \mathrm{Fiber}_{j,y}. \tag{84}$$
$$|\psi_{j,r}(g)|, |\psi_{j,r}(y)| \leq O(\delta) \quad \textit{for all } r \notin \mathfrak{A}_j. \tag{85}$$

Lemma G.4 (Properties of irrelavant 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*

$$\left|\langle \mathbf{W}_{5,j,r,p}, e_v \rangle\right| \leq \widetilde{O}(\sigma_0). \tag{86}$$

The above lemmas give us some direct computations of the inner products between the weight matrices and input embedding vectors.

Lemma G.5. *Let $j \in \tau(\mathcal{Y})$ and $\ell \in [2]$. Then for any $r \in [m]$, the following holds:*

$$\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},\ell} \rangle = \Psi_{j,r}(g_\ell) \pm \widetilde{O}(\sigma_0), \tag{87}$$

$$\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{ans},\ell-1} \rangle = \Psi_{j,r}(y_{\ell-1}) \pm \widetilde{O}(\sigma_0). \tag{88}$$

Moreover, for $j \notin \tau(\mathcal{Y})$ and any $\mathbf{k} \in \mathcal{I}^{2,1}$ and $r \in [m]$, we have

$$\left|\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathbf{k}} \rangle\right| = \widetilde{O}(\sigma_0). \tag{89}$$

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 G.6 (Characterizations of Lambda)**.** *Given $\mathbf{Z}^{2,1}$ with $\{\mathbf{Attn}_{\mathsf{ans},1\to\mathbf{k}}\}_{\mathbf{k}\in\mathcal{I}^{2,1}}$, then*

(a) for $j \in \tau(\mathcal{Y})$, for activated neuron $r \in \mathfrak{A}_j$, we have

$$\Lambda_{5,j,r} = \sum_{\ell'=1}^{2} \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},\ell'} \psi_{j,r}(g_{\ell'}) + \sum_{\ell'=1}^{2} \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},\ell'-1} \psi_{j,r}(y_{\ell'-1}) \pm \widetilde{O}(\sigma_0).$$

(b) for $j \in \tau(\mathcal{Y})$, for any non-activated neuron $r \notin \mathfrak{A}_j$ we have

$$\left|\Lambda_{5,j,r}\right| \leq O(\delta).$$

(c) for $j \notin \tau(\mathcal{Y})$, for any $r \in [m]$, we have

$$\left|\Lambda_{5,j,r}\right| \leq \widetilde{O}(\sigma_0).$$

A direct consequence of the above lemma is the following finer characterization of the activated neurons.

Lemma G.7. *Given $j \in \tau(\mathcal{Y})$, for $r \in \mathfrak{A}_j \setminus \widehat{\mathfrak{A}}_j$, we have $\mathbf{sReLU}'\big(\Lambda_{5,j,r}\big) = 0$.*

Now we are ready to further derive the gradients of the attention layer starting from Lemmas G.1 and G.2 and the properties established above.

Lemma G.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*

$$\begin{aligned}
\Big[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}_5^{2,2}\Big]_{s,s} = \mathbb{E}\Bigg[& \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2} \cdot \\
& \Bigg((1 - \mathbf{logit}_{5,j_2}) \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \Big(\psi_{j_2,r}(g_2) - \Lambda_{5,j_2,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q) \Big) \\
& - \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \Big(\psi_{j,r}(g_2) - \Lambda_{5,j,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q) \Big) \\
& \pm \sum_{j\notin\tau(\mathcal{Y})} \mathbf{logit}_{5,j} \widetilde{O}(\sigma_0^q) \Bigg) \mathbb{1}_{\tau(x_1)=s} \Bigg]
\end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\Big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}_5^{2,2}\Big]_{s,s'} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot \Big((1-\mathbf{logit}_{5,j_2})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot\Big(\psi_{j_2,r}(g_1)-\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big) - \sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(g_1)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big) \pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_1)=s,\tau(x_0)=s'}\Bigg]$$

Lemma G.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*

$$\Big[-\nabla_{\mathbf{Q}_{4,4}}\mathsf{Loss}_5^{2,2}\Big]_{s,s} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot \Big((1-\mathbf{logit}_{5,j_2})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot\Big(\psi_{j_2,r}(y_1)-\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big) - \sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(y_1)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big) \pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_1)=s}\Bigg].$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\Big[-\nabla_{\mathbf{Q}_{4,4}}\mathsf{Loss}_5^{2,2}\Big]_{s,s'} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot \Big((1-\mathbf{logit}_{5,j_2})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot\Big(\psi_{j_2,r}(y_0)-\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big) - \sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(y_0)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big) \pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_1)=s,\tau(x_0)=s'}\Bigg].$$

Following the above calculations, we can further obtain the gradient summation of $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ as follows:

Lemma G.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*

$$\begin{aligned}&\Big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}_5^{2,2}\Big]_{s,s} + \Big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}_5^{2,2}\Big]_{s,s}\\ &= \mathbb{E}\Bigg[\Big((1-\mathbf{logit}_{5,j_2})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot \Big(-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j_2,r}(y_0)-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j_2,r}(g_1)\\ &\qquad +\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\\
&\Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r}(y_0)+\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j,r}(g_1)\\
&\quad-\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\\
&\pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}+\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_1)=s}\Big].
\end{aligned}
$$

Notations for gradient decompositions. We shall define some useful notations to further simplify the expressions of gradient.

Lemma G.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 $\big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}_5^{2,2}\big]_{s,s}=\mathbb{E}\big[\mathcal{N}_{s,3,2,i}+\mathcal{N}_{s,3,2,ii}+\mathcal{N}_{s,3,2,iii}\big]$, where*

$$
\begin{aligned}
\mathcal{N}_{s,3,2,i}=\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot(1-\mathbf{logit}_{5,j_2})\cdot\\
\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot\Big(\psi_{j_2,r}(g_2)-\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s},
\end{aligned}
\tag{90}
$$

$$
\begin{aligned}
\mathcal{N}_{s,3,2,ii}=-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\\
\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(g_2)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s},
\end{aligned}
\tag{91}
$$

$$
\mathcal{N}_{s,3,2,iii}=\pm\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\mathbb{1}_{\tau(x_1)=s}.
\tag{92}
$$

2. *for $[\mathbf{Q}_{4,4}]_{s,s}$, we have $\big[-\nabla_{\mathbf{Q}_{4,4}}\mathsf{Loss}_5^{2,2}\big]_{s,s}=\mathbb{E}\big[\mathcal{N}_{s,4,2,i}+\mathcal{N}_{s,4,2,ii}+\mathcal{N}_{s,4,2,iii}\big]$, where*

$$
\begin{aligned}
\mathcal{N}_{s,4,2,i}=\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot(1-\mathbf{logit}_{5,j_2})\cdot\\
\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot\Big(\psi_{j_2,r}(y_1)-\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s},
\end{aligned}
\tag{93}
$$

$$
\begin{aligned}
\mathcal{N}_{s,4,2,ii}=-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\\
\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(y_1)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s},
\end{aligned}
\tag{94}
$$

$$
\mathcal{N}_{s,4,2,iii}=\pm\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\mathbb{1}_{\tau(x_1)=s}.
\tag{95}
$$

3. *for the summation of $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$, we have $\big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}_5^{2,2}\big]_{s,s}+\big[-\nabla_{\mathbf{Q}_{4,4}}\mathsf{Loss}_5^{2,2}\big]_{s,s}=\mathbb{E}\big[\mathcal{N}_{s,2,i}+\mathcal{N}_{s,2,ii}+\mathcal{N}_{s,2,iii}\big]$, where*

$$
\begin{aligned}
\mathcal{N}_{s,2,i}=(1-\mathbf{logit}_{5,j_2})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot\\
\Big(-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j_2,r}(y_0)-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j_2,r}(g_1)\\
+\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}
\end{aligned}
$$

$$\mathcal{N}_{s,2,ii} = \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r})\cdot \Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r}(y_0) + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j,r}(g_1) - \big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}$$

$$\mathcal{N}_{s,2,iii} = \pm\sum_{j\notin\tau(\mathcal{Y})} \mathbf{logit}_{5,j}\cdot\big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}+\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\widetilde{O}(\sigma_0^q)\mathbb{1}_{\tau(x_1)=s}.$$

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 $\mathbf{Q}_{4,4}$.

$$\mathcal{E}_1 = \Big\{y_0\neq y_1, g_1(y_{\ell-1})\neq g_2(y_{\ell-1}),\ \text{for all } \ell\in[2]\Big\} \tag{96}$$

$$\mathcal{E}_2 = \Big\{g_1,g_2\in\mathrm{Fiber}_{\tau(g_2(y_0)),y_0}, g_1(y_1)\neq g_2(y_1)\Big\}, \tag{97}$$

$$\mathcal{E}_3 = \Big\{y_0=y_1, g_1(y_1)\neq g_2(y_1)\Big\}. \tag{98}$$

$$\mathcal{E}_4 = \Big\{y_0\neq y_1, g_1(y_{\ell-1})= g_2(y_{\ell-1}),\ \text{for all } \ell\in[2]\Big\} \tag{99}$$

$$\tag{100}$$

We first observe that event $\mathcal{E}_1$ occurs with high probability $1-O\big(\frac{1}{n_c}\big)$, 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.

$$\mathcal{E}_5 = \Big\{y_0\neq y_1, g_1(y_1)=g_2(y_1), g_1(y_0)\neq g_2(y_0)\Big\} \tag{101}$$

$$\mathcal{E}_6 = \Big\{y_0=y_1, g_1(y_1)=g_2(y_1)\Big\}. \tag{102}$$

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 $\cup_{i\in[6]}\mathcal{E}_i$ forms a partition of the entire sample space.

G.3 Stage 1.2.1: Inital Growth of Q

We define the following notations:

$$\epsilon_{\mathsf{attn}}^{L,\ell}\big(\mathbf{Z}^{L,\ell-1}\big) = 1-\mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{pred},\ell}\big(\mathbf{Z}^{L,\ell-1}\big)-\mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{ans},\ell-1}\big(\mathbf{Z}^{L,\ell-1}\big),$$
$$\Delta^{L,\ell}\big(\mathbf{Z}^{L,\ell-1}\big) = \Big|\mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{pred},\ell}\big(\mathbf{Z}^{L,\ell-1}\big)-\mathbf{Attn}_{\mathsf{ans},\ell-1\to\mathsf{ans},\ell-1}\big(\mathbf{Z}^{L,\ell-1}\big)\Big|.$$

We abbreviate $\epsilon_{\mathsf{attn}}^{L,\ell}\big(\mathbf{Z}^{L,\ell-1}\big)$ and $\Delta^{L,\ell}\big(\mathbf{Z}^{L,\ell-1}\big)$ as $\epsilon_{\mathsf{attn}}^{L,\ell}$ 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 G.1. *Given $s\in\tau(\mathcal{X})$, let $T_{1,2,1,s}$ denote the first time that $\mathbb{E}[\epsilon_{\mathsf{attn}}\mid\tau(x_1)=s]\le 0.4$ For all iterations $t\le T_{1,2,1,s}$, we have the following holds*

(a) $\big[\mathbf{Q}_{4,3}^{(t)}\big]_{s,s}+\big[\mathbf{Q}_{4,4}^{(t)}\big]_{s,s}\le O(1)$ *monotonically increases;*

(b) for $p\in\{3,4\}$, for $s'\in\tau(\mathcal{X})\neq s$, $\Big|\big[\mathbf{Q}_{4,p}^{(t)}\big]_{s,s'}\Big|\le O\Big(\frac{\big[\mathbf{Q}_{4,p}^{(t)}\big]_{s,s'}}{d}\Big)$ *; otherwise* $\big[\mathbf{Q}_{4,p}^{(t)}\big]_{s,s'}=0$*;*

(c) *for any sample* $\mathbf{Z}^{2,1}$*, we have* $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} \geq -O\big(\frac{\log\log d}{\log d}\big)$*;*

(d) *for any sample* $\mathbf{Z}^{2,1}$*, we have* $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} \leq c_1$ *for some small constant* $c_1 > 0$*;*

G.3.1 Attention and Lambda Preliminaries

Lemma G.12. *If Induction G.1 holds for all iterations* $< t$*,then we have*

1. $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} \in [0.4 \pm \widetilde{O}\big(\frac{1}{d}\big), 0.5]$*;*
2. $\big|\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\big| \leq \widetilde{O}\big(\frac{1}{d}\big)$*;*

Lemma G.13. *If Induction G.1 holds for all iterations* $< t$*, then given* $\mathbf{Z}^{2,1} \in \mathcal{E}_1$*,*

1. *for the prediction* j_2*, we have*

$$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\Big)\cdot 2B + \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\Big)\cdot\frac{2B}{n_y} + O(\delta).$$

2. *for the prediction* $j_2' = \tau\big(g_2(y_0)\big)$*, we have*

$$\Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta).$$

3. *for the prediction* $\tau(g_1(y_0))$*, we have*

$$\Lambda^{(t)}_{5,\tau(g_1(y_0)),r_{g_1\cdot y_0}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\Big)\cdot\frac{2B}{n_y} + O(\delta).$$

   furthermore, we have $\Lambda^{(t)}_{5,\tau(g_1(y_0)),r_{g_1\cdot y_0}} \leq \Lambda^{(t)}_{5,\tau(g_2(y_1)),r_{g_2\cdot y_1}}$*.*

4. *for the prediction* $\tau(g_1(y_1))$*, we have*

$$\Lambda^{(t)}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\Big)\cdot\frac{2B}{n_y} + \widetilde{O}\Big(\frac{B}{d}\Big) + O(\delta).$$

5. *for other* $j \in \tau(\mathcal{Y})$*, if there exsit* j *and* y*, s.t.,* $g_1, g_2 \in \mathrm{Fiber}_{j,y}$ *(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*

$$\Lambda^{(t)}_{5,j,r_{g_2\cdot y}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta);$$

   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* $r \in \cup_{\ell,\ell'\in[2]}\Big(\widehat{\mathfrak{A}}_{\tau(g_\ell\cdot y_{\ell'-1})} \setminus \{r_{g_\ell\cdot y_{\ell'-1}}\}\Big)$ *cannot be activated.*

Lemma G.14. *If Induction G.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*

$$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\Big)\cdot 2B + \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\Big)\cdot\frac{2B}{n_y} + O(\delta).$$

2. *for the prediction $j_2' = \tau\big(g_2(y_0)\big)$, $r \in \widehat{\mathfrak{A}}_{j_2'} \setminus \{r_{g_2\cdot y_0}\}$ (notice that in this case $r_{g_2\cdot y_0} = r_{g_1\cdot y_0}$) cannot be activated, moreover, we have*

$$\Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + O(\delta).$$

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*

$$\Lambda^{(t)}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\Big)\cdot \frac{2B}{n_y} + O(\delta).$$

4. *for other $j \in \tau(\mathcal{Y})$, if there exsit j and y, s.t., $g_1, g_2 \in \mathrm{Fiber}_{j,y}$ (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*

$$\Lambda^{(t)}_{5,j,r_{g_2\cdot y}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta);$$

else, none of $r \in \widehat{\mathfrak{A}}_j$ can be activated.

Lemma G.15. *If Induction G.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*

$$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot 2B \pm O(\delta).$$

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*

$$\Lambda^{(t)}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}} = \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot 2B \pm O(\delta).$$

3. *for other $j = g(y_1)$, where $g_1, g_2 \notin \mathrm{Fiber}_{j,y_1}$, we have*

$$\Lambda^{(t)}_{5,j,r_{g\cdot y_1}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\Big)\cdot \frac{2B}{n_y} + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta);$$

moreover, if there exsit y, s.t., $g_1, g_2 \in \mathrm{Fiber}_{j,y}$ (notice that $y \neq y_1$ for $\mathcal{E}_3$), we have

$$\Lambda^{(t)}_{5,j,r_{g_2\cdot y}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta);$$

besides, other $r \in \widehat{\mathfrak{A}}_j$ cannot be activated.

Lemma G.16. *If Induction G.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*

$$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot 2B + O(\delta).$$

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*

$$\Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + O(\delta).$$

3. *for other $j \in \tau(\mathcal{Y})$, if there exsit j and y, s.t., $g_1, g_2 \in \mathrm{Fiber}_{j,y}$, then for such a j, $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2\cdot y}\}$ cannot be activated, moreover, we have*

$$\Lambda^{(t)}_{5,j,r_{g_2\cdot y}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta);$$

else, none of $r \in \widehat{\mathfrak{A}}_j$ can be activated.

Lemma G.17. *If Induction G.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*
$$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \cdot 2B \pm O(\delta).$$
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*
$$\Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta).$$
3. *for the prediction $\tau(g_1(y_0))$, $r \in \widehat{\mathfrak{A}}_{\tau(g_2(y_0))} \setminus \{r_{g_1\cdot y_0}\}$ cannot be activated, moreover, we have*
$$\begin{aligned}\Lambda^{(t)}_{5,\tau(g_1(y_0)),r_{g_1\cdot y_0}} &= \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B \\ &+ \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\Big)\cdot\frac{2B}{n_y} + O(\delta).\end{aligned}$$
4. *for other $j \in \tau(\mathcal{Y})$, if there exsit j and y, s.t., $g_1, g_2 \in \mathrm{Fiber}_{j,y}$, then for such a j, $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2\cdot y}\}$ cannot be activated, moreover, we have*
$$\Lambda^{(t)}_{5,j,r_{g_2\cdot y}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta);$$
else, none of $r \in \widehat{\mathfrak{A}}_j$ can be activated.

Lemma G.18. *If Induction G.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*
$$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\Big)\cdot 2B \pm O(\delta).$$
2. *for other $j = g(y_1)$, where $g_2 \notin \mathrm{Fiber}_{j,y_1}$, we have*
$$\Lambda^{(t)}_{5,j,r_{g\cdot y_1}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\Big)\cdot\frac{2B}{n_y} + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta);$$
moreover, if there exsit $y \neq y_1$, s.t., $g_1, g_2 \in \mathrm{Fiber}_{j,y}$, we have
$$\Lambda^{(t)}_{5,j,r_{g_2\cdot y}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta);$$
besides, other $r \in \widehat{\mathfrak{A}}_j$ cannot be activated.

G.3.2 Gradient Lemma

Lemma G.19. *If Induction G.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, we have*
$$\begin{aligned}&\Big[-\nabla_{\mathbf{Q}^{(t)}_{4,3}}\mathsf{Loss}^{2,2}_5\Big]_{s,s} + \Big[-\nabla_{\mathbf{Q}^{(t)}_{4,4}}\mathsf{Loss}^{2,2}_5\Big]_{s,s} \\ &\quad\geq \min\Big\{\Omega\Big(\frac{1}{d\cdot n_y}\Big), \Omega(\frac{B}{d})\cdot\mathbb{E}\Big[(1-\mathbf{logit}^{(t)}_{5,j_2})\Big|\tau(x_1)=s,\mathcal{E}_1\Big]\Big\} > 0.\end{aligned}$$

Proof. By Lemma G.10 and Lemma G.11, we have $\big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}^{2,2}_5\big]_{s,s} + \big[-\nabla_{\mathbf{Q}_{4,4}}\mathsf{Loss}^{2,2}_5\big]_{s,s} = \mathbb{E}\big[\mathcal{N}_{s,2,i} + \mathcal{N}_{s,2,ii} + \mathcal{N}_{s,2,iii}\big]$. Based onLemma G.13-Lemma G.18, we can first directly bound the term $\mathcal{N}^{(t)}_{s,2,iii}$ as follows:
$$\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,iii}\Big] \leq \widetilde{O}(\sigma_0^q) = \frac{1}{\mathsf{poly}d}.$$

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) $\mathbb{E}\big[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)} \mid \tau(x_1) = s\big] \le \frac{\varrho}{B}$, and (ii) $\mathbb{E}\big[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)} \mid \tau(x_1) = s\big] \ge \frac{\varrho}{B}$. 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 G.1 and Lemma G.13, we have an immediate logit upper bound for $j \in \tau(\mathcal{Y})$: $\mathbf{logit}_{5,j}^{(t)} \le O\big(\frac{1}{d}\big)$. Then by Lemma G.10, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}_{s,2,i}^{(t)}\mathbb{1}_{\mathcal{E}_1}\Big] \\
&= \mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r}^{(t)})\cdot \\
&\quad\Big(-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j_2,r}(y_0) - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{j_2,r}(g_1) \\
&\quad + \big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j_2,r}^{(t)} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Big]
\end{aligned}
$$

   By Lemma G.13, we can obtain

$$
\begin{aligned}
&\mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\cdot\Big(\mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)})\cdot \\
&\quad\Big(-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(y_0) - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(g_1) \\
&\quad + \big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Big] \\
&\ge \Omega\Big(\frac{B}{d}\Big)\cdot\mathbb{E}\Big[\mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)})\big|\tau(x_1)=s,\mathcal{E}_1\Big] \qquad (103)
\end{aligned}
$$

   On the other hand,

$$
\begin{aligned}
&-\mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\cdot\mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_1\cdot y_0}}^{(t)})\cdot \\
&\qquad\Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j_2,r_{g_1\cdot y_0}}(y_0) + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{j_2,r_{g_1\cdot y_0}}(g_1)\Big) \\
&\qquad\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\mathbb{1}_{g_1(y_0)=g_2(y_1)}\Big] \\
&\ge -O\big(\frac{B}{d\cdot n_y}\big)\cdot\mathbb{E}\Big[\mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)})\big|\tau(x_1)=s,\mathcal{E}_1\Big],
\end{aligned}
$$

   which implies that $\mathbb{E}\Big[\mathcal{N}_{s,2,i}^{(t)}\mathbb{1}_{\mathcal{E}_1}\Big] \ge \Omega\Big(\frac{B}{d}\Big)\cdot\mathbb{E}\Big[\mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)})\big|\tau(x_1) = s,\mathcal{E}_1\Big] \ge 0$. Moving to $\mathcal{N}_{s,2,ii}^{(t)}$, by Lemma G.10 and Lemma G.13, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}_{s,2,ii}^{(t)}\mathbb{1}_{\mathcal{E}_1}\Big] \\
&= \mathbb{E}\Big[\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}^{(t)}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot \\
&\quad\Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j,r}(y_0) + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{j,r}(g_1) \\
&\quad - \big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j,r}^{(t)} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Big]
\end{aligned}
$$

$$\geq -O\Big(\frac{n_y \cdot B}{d^2}\Big).$$

(b) For $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, by Induction G.1 and Lemma G.14, we can also obtain a crude bound of logit: $\mathbf{logit}_{5,j}^{(t)} \leq O\big(\frac{1}{d}\big)$ for $j \neq j_2'$. Then $\mathcal{N}_{s,2,i}^{(t)}$ can be bounded in the same way as (103), and we have

$$\mathbb{E}\Big[\mathcal{N}_{s,2,i}^{(t)} \mathbb{1}_{\mathcal{E}_2}\Big] \geq \Omega\Big(\frac{B}{d \cdot n_y}\Big) \cdot \mathbb{E}\Big[\mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \big| \tau(x_1) = s, \mathcal{E}_2\Big] \geq 0.$$

Moving to $\mathcal{N}_{s,2,ii}^{(t)}$,

- for $j = \tau(g_1(y_1))$, we have

$$\begin{aligned}
&= \mathbb{E}\Big[\mathbf{logit}_{5,j}^{(t)}\Big(\mathbf{sReLU}'(\Lambda_{5,j,r_{g_1 \cdot y_1}}) \cdot \\
&\Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)} \cdot \psi_{j,r_{g_1 \cdot y_1}}(y_0) + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)} \cdot \psi_{j,r_{g_1 \cdot y_1}}(g_1) \\
&- \big(1 - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j,r}^{(t)} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_2}\Big] \\
&\geq -O\Big(\frac{B}{n_y^2 \cdot d^2}\Big).
\end{aligned}$$

  where the last inequality is due to the cancellation of the term $\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)} \cdot \psi_{j,r_{g_1 \cdot y_1}}(y_0) + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)} \cdot \psi_{j,r_{g_1 \cdot y_1}}(g_1)$ and the fact that

$$\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} = \Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\Big) \cdot \frac{2B}{n_y} + O(\delta) \leq O\big(\frac{B}{n_y}\big).$$

- for $j = \tau(g_2(y))$ if there exists $y \neq y_0, y_1$ s.t., $g_1(y) = g_2(y)$
  - when $\mathbb{E}\big[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} \mid \tau(x_1) = s\big] \leq 0.01$, by Induction G.1 and Lemma G.14, we have $\mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \ll O\big(\frac{1}{d}\big) \cdot \mathbf{logit}_{5,j_2'}^{(t)}$ and $\mathbf{logit}_{5,j_2'}^{(t)} = \Omega(1)$. Then

$$\begin{aligned}
&\Big|\mathbb{E}\Big[\mathbf{logit}_{5,j}^{(t)} \cdot \Big(\mathbf{sReLU}'(\Lambda_{5,j,r_{g_1 \cdot y}}^{(t)}) \cdot \\
&\Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)} \cdot \psi_{j,r_{g_1 \cdot y}}(y_0) + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)} \cdot \psi_{j,r_{g_1 \cdot y}}(g_1) \\
&- \big(1 - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j,r_{g_1 \cdot y}}^{(t)} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_2}\Big]\Big| \\
&\leq O\Big(\frac{B}{n_y \cdot d^2}\Big).
\end{aligned}$$

  On the other hand,

$$\begin{aligned}
&\mathbb{E}\Big[\mathbf{logit}_{5,j_2'}^{(t)} \cdot \Big(\mathbf{sReLU}'(\Lambda_{5,j_2',r_{g_1 \cdot y_0}}^{(t)}) \cdot \\
&\Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)} \cdot \psi_{j_2',r_{g_1 \cdot y_0}}(y_0) + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)} \cdot \psi_{j_2',r_{g_1 \cdot y_0}}(g_1) \\
&- \big(1 - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j_2',r_{g_1 \cdot y_0}}^{(t)} \pm \widetilde{O}(\sigma_0)\Big) \\
&\quad \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Big] \\
&\geq \Omega\Big(\frac{B}{n_y \cdot d}\Big),
\end{aligned}$$

  where the last inequality follows from the fact that

$$\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)} \cdot \psi_{j_2',r_{g_1 \cdot y_0}}(y_0) + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)} \cdot \psi_{j_2',r_{g_1 \cdot y_0}}(g_1)$$

$$
\begin{aligned}
&- \big(1 - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j_2',r_{g_1\cdot y_0}} \\
&= \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\Big(2B - 2\Lambda^{(t)}_{5,j_2',r_{g_1\cdot y_0}}\Big) \pm O(\delta) \geq \Omega(B).
\end{aligned}
$$

– when $\mathbb{E}\big[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \mid \tau(x_1) = s\big] \geq 0.01$, by Lemma G.14, we have $\Lambda^{(t)}_{5,j,r_{g_1\cdot y}}$ cannot be activated. Furthermore,

$$
\begin{aligned}
&\mathbb{E}\Big[\mathbf{logit}^{(t)}_{5,j_2'}\cdot\Big(\mathbf{sReLU}'(\Lambda^{(t)}_{5,j_2',r_{g_1\cdot y_0}})\cdot \\
&\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j_2',r_{g_1\cdot y_0}}(y_0) + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j_2',r_{g_1\cdot y_0}}(g_1) \\
&- \big(1 - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j_2',r_{g_1\cdot y_0}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Big] \\
&\geq 0.
\end{aligned}
$$

Putting the above discussion together, we have

$$
\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_2}\Big] \geq -O\Big(\frac{B}{n_y^2\cdot d^2}\Big).
$$

(c) For $\mathbf{Z}^{2,1}\in\mathcal{E}_3$, by Induction G.1 and Lemma G.15, we can first have some facts of logit:

$$
1 - \mathbf{logit}^{(t)}_{5,j_2} = \Omega(1), \quad \mathbf{logit}^{(t)}_{5,\tau(g_1(y_1))} = \Omega(1)
$$
$$
\mathbf{logit}^{(t)}_{5,j} \leq \frac{1}{\mathsf{poly}d} \text{ for other } j.
$$

Therefore, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,i}\mathbb{1}_{\mathcal{E}_3}\Big] \\
&= \mathbb{E}\Big[(1 - \mathbf{logit}^{(t)}_{5,j_2})\cdot\Big(-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(y_1) - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(g_1) \\
&\quad + \big(1 - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} \pm \widetilde{O}(\sigma_0) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_3}\Big] \\
&\geq \Omega\Big(\frac{B}{n_y\cdot d}\Big).
\end{aligned}
$$

Moving to $\mathcal{N}^{(t)}_{s,2,ii}$, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_3}\Big] \\
&\geq \mathbb{E}\Big[\mathbf{logit}^{(t)}_{5,\tau(g_1(y_1))}\cdot \\
&\quad\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{\tau(g_1(y_1)),r_{g_1\cdot y_1}}(y_1) + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{\tau(g_1(y_1)),r}(g_1) \\
&- \big(1 - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}} \pm \widetilde{O}(\sigma_0) \\
&\pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_3}\Big] - O\Big(\frac{n_y\cdot B}{d\cdot\mathsf{poly}d}\Big) \\
&\geq \Omega\Big(\frac{B}{n_y\cdot d}\Big) - O\Big(\frac{B}{d\cdot\mathsf{poly}d}\Big) = \Omega\Big(\frac{B}{n_y\cdot d}\Big).
\end{aligned}
$$

(d) For $\mathbf{Z}^{2,1}\in\mathcal{E}_4$, by comparing Lemma G.16 with Lemma G.14, we can directly bound $\mathcal{N}^{(t)}_{s,2,ii}$ 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 $\mathbb{E}\Big[\mathcal{N}_{s,2,ii}^{(t)}\mathbb{1}_{\mathcal{E}_4}\Big] \geq 0$. Moreover, it is clear that

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}_{s,2,i}^{(t)}\mathbb{1}_{\mathcal{E}_4}\Big] \\
&= \mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\cdot\Big(\mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)})\cdot \\
&\quad\Big(-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(y_0)-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(g_1) \\
&\quad+\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_4}\Big] \\
&\geq 0.
\end{aligned}$$

where the last inequality is due to the cancellation of $\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(y_0)+\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(g_1)$, and the fact that

$$\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)} = \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\cdot 2B + O(\delta) \geq \Omega(B).$$

(e) For $\mathbf{Z}^{2,1}\in\mathcal{E}_5\cup\mathcal{E}_6$, by Induction G.1, Lemma G.17 and Lemma G.18, we can derive the following logit condition: $1-\mathbf{logit}_{5,j_2}^{(t)}, \mathbf{logit}_{5,j}^{(t)} \leq \frac{1}{\mathsf{poly}d}$ 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:

$$\Big|\mathbb{E}\Big[\mathcal{N}_{s,2,i}^{(t)}\mathbb{1}_{\mathcal{E}_m}\Big]\Big|, \Big|\mathbb{E}\Big[\mathcal{N}_{s,2,ii}^{(t)}\mathbb{1}_{\mathcal{E}_m}\Big]\Big| \leq O\Big(\frac{B}{dn_y}\cdot\frac{1}{\mathsf{poly}d}\Big) \text{ for } m\in\{5,6\}.$$

Putting it all together, we have for the regmie (1),

$$\Big[-\nabla_{\mathbf{Q}_{4,3}^{(t)}}\mathsf{Loss}_5^{2,2}\Big]_{s,s}+\Big[-\nabla_{\mathbf{Q}_{4,4}^{(t)}}\mathsf{Loss}_5^{2,2}\Big]_{s,s} = \sum_{\kappa\in\{i,ii,iii\}}\sum_{i\in[6]}\mathbb{E}\Big[\mathcal{N}_{s,2,\kappa}^{(t)}\mathbb{1}_{\mathcal{E}_i}\Big] \geq \Omega(\frac{B}{d\cdot n_y}).$$

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 $\Lambda_{5,j_2,rg_2\cdot y_1}^{(t)}$ 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 G.1 and Lemma G.13, we have

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}_{s,2,i}^{(t)}\mathbb{1}_{\mathcal{E}_1}\Big] \\
&= \mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r}^{(t)})\cdot \\
&\quad\Big(-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j_2,r}(y_0)-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{j_2,r}(g_1) \\
&\quad+\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j_2,r}^{(t)}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Big]
\end{aligned}$$

By Lemma G.13, we can obtain

$$\begin{aligned}
&\mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\cdot\Big(\Big(-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(y_0)-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(g_1) \\
&\quad+\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Big] \qquad (104) \\
&\geq \Omega\Big(\frac{B}{d}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\big|\tau(x_1)=s,\mathcal{E}_1\Big].
\end{aligned}$$

On the other hand,

$$
\begin{aligned}
&-\mathbb{E}\Big[(1-\mathbf{logit}^{(t)}_{5,j_2})\cdot \mathbf{sReLU}'(\Lambda^{(t)}_{5,j_2,r_{g_1\cdot y_0}})\cdot \\
&\qquad \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot \psi_{j_2,r_{g_1\cdot y_0}}(y_0)+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot \psi_{j_2,r_{g_1\cdot y_0}}(g_1)\Big) \\
&\qquad \mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\mathbb{1}_{g_1(y_0)=g_2(y_1)}\Big] \\
&\geq -O\big(\frac{B}{d\cdot n_y}\big)\cdot \mathbb{E}\Big[(1-\mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1)=s,\mathcal{E}_1\Big],
\end{aligned}
$$

which implies that $\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,i}\mathbb{1}_{\mathcal{E}_1}\Big]\geq \Omega\Big(\frac{B}{d}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1)=s,\mathcal{E}_1\Big]\geq 0$.

Moving to $\mathcal{N}^{(t)}_{s,2,ii}$, by Lemma G.13, we have

- for $j=\tau(g_1(y_0)), r=r_{g_1\cdot y_0}$

$$
\begin{aligned}
&\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r}(y_0)+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j,r}(g_1) \\
&\qquad -\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\geq 0.
\end{aligned}
$$

- for $j=\tau(g_1(y_1))$, $r=r_{g_1\cdot y_1}$, due to the cancellation of $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r}(y_0)+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j,r}(g_1)$, we have

$$
\begin{aligned}
&\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r}(y_0)+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j,r}(g_1) \\
&\qquad -\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big) \\
&\geq -\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j,r}.
\end{aligned}
$$

- for $j=\tau(g_2(y_0)), r=r_{g_2\cdot y_0}$

$$
\begin{aligned}
&\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r}(y_0)+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j,r}(g_1) \\
&\qquad -\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big) \\
&\geq \big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j,r}.
\end{aligned}
$$

- for $j=g_1(y)$, where $\exists y\neq y_0,y_1$, s.t., $g_2(y)=g_1(y)$, we have

$$
\begin{aligned}
&\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r}(y_0)+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j,r}(g_1) \\
&\qquad -\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big) \\
&\geq -\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j,r}.
\end{aligned}
$$

Putting them together, and upper bound $\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}^{(t)}_{5,j}$ by $1-\mathbf{logit}^{(t)}_{5,j_2}$, we have

$$
\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_1}\Big] \\
&=\mathbb{E}\Big[\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}^{(t)}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot \\
&\qquad \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r}(y_0)+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j,r}(g_1) \\
&\qquad -\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Big]
\end{aligned}
$$

$$\geq -\mathbb{E}\Big[\Big(1-\mathbf{logit}^{(t)}_{5,j_2}\Big)\cdot\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)$$
$$\cdot\max_{y\neq y_0,y_1}\Big\{\Lambda^{(t)}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}},\Lambda^{(t)}_{5,\tau(g_2(y_0)),r_{g_2\cdot y_0}}\Big\}\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Big],$$

where the last inequality is due to the fact that $\Lambda^{(t)}_{5,\tau(g_2(y_0)),r_{g_2\cdot y_0}}=\Lambda^{(t)}_{5,\tau(g_2(y)),r_{g_2\cdot y}}\pm\widetilde{O}(\frac{B}{d\cdot n_y})$ for $y\neq y_0,y_1$ s.t., $g_1(y)=g_2(y)$. Notice that

$$\max_{y\neq y_0,y_1}\Big\{\Lambda^{(t)}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}},\Lambda^{(t)}_{5,\tau(g_2(y_0)),r_{g_2\cdot y_0}}\Big\}\leq$$
$$\max\Big\{\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1},\Theta(\frac{1}{n_y})\Big\}2B,$$

while

$$(104)\geq\mathbb{E}\Big[\Big(1-\mathbf{logit}^{(t)}_{5,j_2}\Big)\cdot\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot 2B\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Big]$$
$$=\frac{1}{2}\mathbb{E}\Big[\Big(1-\mathbf{logit}^{(t)}_{5,j_2}\Big)\cdot\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)2B\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Big].$$

Since by Induction G.1, $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\leq c_1\ll\frac{1}{2}$, thus we have

$$\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,i}\mathbb{1}_{\mathcal{E}_1}\Big]+\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_1}\Big]\geq\Omega\Big(\frac{B}{d}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1)=s,\mathcal{E}_1\Big].$$

(b) For $\mathbf{Z}^{2,1}\in\mathcal{E}_2$, $\mathcal{N}^{(t)}_{s,2,i}$ can be bounded in the same way as (104), and we have

$$\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,i}\mathbb{1}_{\mathcal{E}_2}\Big]\geq\Omega\Big(\frac{B}{d\cdot n_y}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1)=s,\mathcal{E}_2\Big]\geq 0.$$

Moving to $\mathcal{N}^{(t)}_{s,2,ii}$,

- for $j=\tau(g_1(y_1))$, we have

$$=\mathbb{E}\Big[\mathbf{logit}^{(t)}_{5,j}\Big(\mathbf{sReLU}'(\Lambda_{5,j,r_{g_1\cdot y_1}})\cdot$$
$$\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r_{g_1\cdot y_1}}(y_0)+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j,r_{g_1\cdot y_1}}(g_1)\qquad(105)$$
$$-\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_2}\Big]$$
$$\geq -O\Big(\frac{B}{n_y^2\cdot d}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1)=s,\mathcal{E}_2\Big].$$

  where the last inequality is due to the cancellation of the term $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r_{g_1\cdot y_1}}(y_0)+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j,r_{g_1\cdot y_1}}(g_1)$ and the fact that

$$\Lambda^{(t)}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}}=\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\Big)\cdot\frac{2B}{n_y}+O(\delta)\leq O(\frac{B}{n_y}).$$

- for $j_2'=\tau(g_2(y_0))=\tau(g_1(y_0))$, clearly, we have

$$\mathbb{E}\Big[\mathbf{logit}^{(t)}_{5,j_2'}\cdot\Big(\mathbf{sReLU}'(\Lambda^{(t)}_{5,j_2',r_{g_1\cdot y_0}})\cdot$$
$$\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j_2',r_{g_1\cdot y_0}}(y_0)+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j_2',r_{g_1\cdot y_0}}(g_1)$$
$$-\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j_2',r_{g_1\cdot y_0}}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)$$
$$\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_2}\Big]\geq 0,$$

where the last inequality is due to the fact that

$$\begin{aligned}
&\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j_2',r_{g_1\cdot y_0}}(g_1)-\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\\
&\quad-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j_2',r_{g_1\cdot y_0}}\\
&\geq\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\\
&\quad\cdot\bigg(1-2\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\bigg)2B\\
&\geq\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\bigg(1-2\Big(c_1+0.25\Big)\bigg)2B\geq 0.
\end{aligned}$$

- for $j=\tau(g_2(y))$ if there exists $y\neq y_0,y_1$ s.t., $g_1(y)=g_2(y)$
  - when $\mathbb{E}\big[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\mid\tau(x_1)=s\big]\leq 0.01$, by Induction G.1 and Lemma G.14, we have $\mathbf{logit}^{(t)}_{5,\tau(g_2(y))}\ll O\big(\frac{1}{d}\big)\cdot\mathbf{logit}^{(t)}_{5,j_2'}$. Then

$$\begin{aligned}
&\bigg|\mathbb{E}\bigg[\mathbf{logit}^{(t)}_{5,j}\cdot\Big(\mathbf{sReLU}'(\Lambda^{(t)}_{5,j,r_{g_1\cdot y}})\cdot\\
&\quad\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r_{g_1\cdot y}}(y_0)+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j,r_{g_1\cdot y}}(g_1)\\
&\quad-\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j,r_{g_1\cdot y}}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_2}\bigg]\bigg|\\
&\leq O\Big(\frac{1}{d}\Big)\cdot(105).
\end{aligned}$$

  - when $\mathbb{E}\big[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\mid\tau(x_1)=s\big]\geq 0.01$, by Lemma G.14, we have $\Lambda^{(t)}_{5,j,r_{g_1\cdot y}}$ cannot be activated.

Putting the above discussion together, we have

$$\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_1}\Big]+\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_2}\Big]\geq\Omega\Big(\frac{B}{d\cdot n_y}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1)=s,\mathcal{E}_2\Big].$$

(c) For $\mathbf{Z}^{2,1}\in\mathcal{E}_3$, by Induction G.1 and Lemma G.15, we can first have the following logit bound:

$$\mathbf{logit}^{(t)}_{5,j}\leq\frac{1}{\mathsf{poly}d}\cdot\Big(1-\mathbf{logit}^{(t)}_{5,j_2}\Big)\text{ for }j\neq j_2,\tau\big(g_1(y_1)\big).$$

Therefore, we have

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,i}\mathbb{1}_{\mathcal{E}_3}\Big]\\
&=\mathbb{E}\bigg[(1-\mathbf{logit}^{(t)}_{5,j_2})\cdot\Big(-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(y_1)-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(g_1)\\
&\quad+\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}}\pm\widetilde{O}(\sigma_0)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_3}\bigg]\\
&\geq\Omega\Big(\frac{B}{n_y\cdot d}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1)=s,\mathcal{E}_3\Big].
\end{aligned}$$

Moving to $\mathcal{N}^{(t)}_{s,2,ii}$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_3}\Big]\\
&\geq\mathbb{E}\bigg[\mathbf{logit}^{(t)}_{5,\tau(g_1(y_1))}\cdot\\
&\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{\tau(g_1(y_1)),r_{g_1\cdot y_1}}(y_1)+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{\tau(g_1(y_1)),r}(g_1)
\end{aligned}$$

$$- \big(1 - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}} \pm \widetilde{O}(\sigma_0) \pm \widetilde{O}(\delta^q)\Big) \mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_3}\Big]$$
$$- O\Big(\frac{n_y \cdot B}{n_y \cdot d \cdot \mathsf{poly}d}\Big) \cdot \mathbb{E}\Big[(1 - \mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1) = s, \mathcal{E}_3\Big].$$

Thus,

$$\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_3}\Big] + \mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_3}\Big] \geq \Omega\Big(\frac{B}{n_y \cdot d}\Big) \cdot \mathbb{E}\Big[(1 - \mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1) = s, \mathcal{E}_3\Big].$$

(d) For $\mathbf{Z}^{2,1} \in \mathcal{E}_4$, we can bound the gradient in a manner similar to regime (i), and obtain

$$\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_4}\Big] + \mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_4}\Big] \geq 0.$$

(e) For $\mathbf{Z}^{2,1} \in \mathcal{E}_5 \cup \mathcal{E}_6$, by Induction G.1, Lemma G.17 and Lemma G.18, we can derive the following logit condition: $1 - \mathbf{logit}^{(t)}_{5,j_2}, \mathbf{logit}^{(t)}_{5,j} \leq \frac{1}{\mathsf{poly}d} \cdot \mathbb{E}\Big[(1 - \mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1) = s, \mathcal{E}_1\Big]$ for $j \neq j_2$. Hence, we can simply bound $\mathcal{N}^{(t)}_{s,2,i}$ and $\mathcal{N}^{(t)}_{s,2,ii}$ as follows:

$$\Big|\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,i}\mathbb{1}_{\mathcal{E}_m}\Big]\Big|, \Big|\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_m}\Big]\Big|$$
$$\leq O\Big(\frac{B}{dn_y} \cdot \frac{1}{\mathsf{poly}d}\Big) \cdot \mathbb{E}\Big[(1 - \mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1) = s, \mathcal{E}_1\Big], \text{ for } m \in \{5, 6\}.$$

Putting everything together, we have for the regmie (ii),

$$\Big[-\nabla_{\mathbf{Q}^{(t)}_{4,3}}\mathsf{Loss}^{2,2}_5\Big]_{s,s} + \Big[-\nabla_{\mathbf{Q}^{(t)}_{4,4}}\mathsf{Loss}^{2,2}_5\Big]_{s,s} = \sum_{\kappa\in\{i,ii,iii\}}\sum_{i\in[6]}\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,\kappa}\mathbb{1}_{\mathcal{E}_i}\Big]$$
$$\geq \Omega(\frac{B}{d}) \cdot \mathbb{E}\Big[(1 - \mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1) = s, \mathcal{E}_1\Big].$$

□

Lemma G.20. *If Induction G.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*

$$\Big|\Big[-\nabla_{\mathbf{Q}^{(t)}_{4,p}}\mathsf{Loss}^{2,2}_5\Big]_{s,s'}\Big| \leq O\Big(\frac{1}{d}\Big)\Big|\Big[-\nabla_{\mathbf{Q}^{(t)}_{4,p}}\mathsf{Loss}^{2,2}_5\Big]_{s,s}\Big|.$$

G.3.3 Bounded Decrease of Attention to Related Context Clause

Lemma G.21. *If Induction G.1 holds for all iterations $< t$, then for any sample $\mathbf{Z}^{2,1}$, we have*

$$\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \geq -O(\frac{\log d \log d}{\log d}).$$

Proof. Denote the first time that $\mathbb{E}\Big[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big|\tau(x_1) = s\Big] \leq -\Omega\big(\frac{\log d \log d}{\log d}\big)$ as $\widetilde{T}$. Notice that $\Big|\Big[\mathbf{Q}^{(t)}_{4,p}\Big]_{s,s'}\Big| \leq O\left(\frac{\big[\mathbf{Q}^{(t)}_{4,p}\big]_{s,s}}{d}\right)$ for $p \in \{3, 4\}$, thus for any sample $\mathbf{Z}^{2,1}$ satisfying $\tau(x_1) = s$, at time $\widetilde{T}$, we have

$$\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2} \leq -\Omega\Big(\frac{\log d \log d}{\log d}\Big).$$

Based on the the gradient compositions from Lemma G.11, we have $\Big[-\nabla_{\mathbf{Q}_{4,3}} \mathrm{Loss}_5^{2,2} \Big]_{s,s} = \mathbb{E}\Big[\mathcal{N}_{s,3,2,i} + \mathcal{N}_{s,3,2,ii} + \mathcal{N}_{s,3,2,iii}\Big]$, 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 G.19, we can first directly bound the term $\mathcal{N}_{s,3,2,iii}^{(\widetilde{T})}$ as follows:

$$\mathbb{E}\Big[\mathcal{N}_{s,3,2,iii}^{(\widetilde{T})}\Big] \leq \widetilde{O}(\sigma_0^q) = \frac{1}{\mathsf{poly}d}.$$

- for $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, at time $\widetilde{T}$, since the neurons for predicting j_2 cannot be activated, and $\mathbf{logit}_{5,j}^{(\widetilde{T})} \leq O(\frac{1}{d})$ for $j \neq j_2$, thus we can naively bound the gradient on the event $\mathcal{E}_1$ as follows:

$$\Big| \sum_{\kappa \in \{i,ii\}} \mathbb{E}\big[\mathcal{N}_{s,3,2,\kappa}^{(\widetilde{T})} \mathbb{1}_{\mathcal{E}_1}\big] \Big| \leq O\Big(\frac{n_y B}{d^2}\Big) + \widetilde{O}(\delta^q).$$

- 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 (91) and Lemma G.14, we have

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}_{s,3,2,ii}^{(\widetilde{T})} \mathbb{1}_{\mathcal{E}_2}\Big] \geq -\mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})} \\
&\cdot \Big(1 + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(\widetilde{T})} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(\widetilde{T})}\Big) \mid \tau(x_1) = s\Big] \\
&\cdot \mathbb{E}\Big[\mathbf{logit}_{5,j_2^{'}}^{(\widetilde{T})} \mid \tau(x_1) = s, \mathcal{E}_2\Big] \frac{2B}{n_y \cdot d} \pm \widetilde{O}(\sigma_0^q).
\end{aligned} \tag{106}$$

- for $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, by Lemma G.15, we can mainly focus on the term $\mathcal{N}_{s,3,2,i}^{(\widetilde{T})}$, and the prediction of $\tau(g_1(y_1))$ in $\mathcal{N}_{s,3,2,ii}^{(\widetilde{T})}$ since $\mathbf{logit}_{5,\tau(g_1(y_1))}^{(\widetilde{T})} = 1 - O\big(\frac{1}{\log d}\big)$. Hence, we have

$$\begin{aligned}
&\mathbb{E}\Big[\Big(\mathcal{N}_{s,3,2,ii}^{(\widetilde{T})} + \mathcal{N}_{s,3,2,ii}^{(\widetilde{T})}\Big) \mathbb{1}_{\mathcal{E}_3}\Big] \geq \mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})} \\
&\cdot \Big(1 + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(\widetilde{T})} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})}\Big) \mid \tau(x_1) = s\Big] \cdot \Big(1 - O\big(\frac{1}{\log d}\big)\Big) \frac{2B}{n_y \cdot d} \pm \widetilde{O}(\sigma_0^q).
\end{aligned} \tag{107}$$

- for $\mathbf{Z}^{2,1} \in \mathcal{E}_4$, the negative gradient can be bounded in the same way as (106), 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 for $\mathbf{Z}^{2,1} \in \mathcal{E}_5 \cup \mathcal{E}_6$, the overall gradient is also negeligble since $\mathbf{logit}_{5,j_2}^{(t)}$ is very close to 1.

Putting it all together, we have

$$\begin{aligned}
&\Big[-\nabla_{\mathbf{Q}_{4,3}^{(\widetilde{T})}} \mathsf{Loss}_5^{2,2}\Big]_{s,s} \geq -\mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})} \\
&\cdot \Big(1 + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(\widetilde{T})} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(\widetilde{T})}\Big) \mid \tau(x_1) = s\Big] \\
&\cdot \mathbb{E}\Big[\mathbf{logit}_{5,j_2^{'}}^{(\widetilde{T})} \mid \tau(x_1) = s, \mathcal{E}_2\Big] \frac{B}{n_y \cdot d} \\
&+ \mathbb{E}\Big[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})} \\
&\cdot \Big(1 + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(\widetilde{T})} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(\widetilde{T})}\Big) \mid \tau(x_1) = s\Big] \cdot \Big(1 - O\big(\frac{1}{\log d}\big)\Big) \frac{2B}{n_y \cdot d} \pm \widetilde{O}(\sigma_0^q)
\end{aligned}$$

Notice that

$$\begin{aligned}&\Big(1+\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},1}-\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2}\Big)\\&-\Big(1+\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2}-\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},1}\Big)\\&=2\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},1}-\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},1}.\end{aligned}$$

If $2\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},1}-\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},1}<c$ for some small constant $c>0$, s.t., $\frac{cB}{\log d}<1$, then $\Lambda^{(\widetilde{T})}_{5,j_2',r_{g_2\cdot y_0}}\le cB$. Hence $\mathbf{logit}^{(\widetilde{T})}_{5,j_2'}\le\frac{1}{d^{1-\frac{cB}{\log d}}}=o(1)$, and (106) is dominated by the positive term (107). Else, clearly, (106) can be cancelled out by the positive term (107). Therefore,

$$\Big[-\nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,3}}\mathsf{Loss}^{2,2}_5\Big]_{s,s}\ge\Omega\Big(\frac{B}{n_y\cdot d}\Big).$$

As a consequence, together with the growth of $\mathbf{Q}^{(\widetilde{T})}_{4,3}+\mathbf{Q}^{(\widetilde{T})}_{4,4}$, and nearly no change of $[\mathbf{Q}^{(t)}_{4,p}]_{s,s'}$ in Lemma G.20, we have $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}$ must start to decrease and cannot be larger than $O\big(\frac{\log\log d}{\log d}\big)$. □

G.3.4 Attention gap is small

Lemma G.22. *If Induction G.1 holds for all iterations $< t$, then for any sample $\mathbf{Z}^{2,1}$, we have*

$$\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\le c_1,$$

where $c_1>0$ is a small constant.

Proof. Let $\widetilde{T}$ denote the first time $\mathbb{E}\Big[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\big|\tau(x_1)=s\Big]\ge\frac{1.0005\log d}{2B}$. Notice that $\Big|\Big[\mathbf{Q}^{(t)}_{4,p}\Big]_{s,s'}\Big|\le O\left(\frac{\big[\mathbf{Q}^{(t)}_{4,p}\big]_{s,s}}{d}\right)$ for $p\in\{3,4\}$, thus for any sample $\mathbf{Z}^{2,1}$ satisfying $\tau(x_1)=s$, at time $\widetilde{T}$, we have

$$\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2}-\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},1}\ge\frac{1.0005\log d}{2B}.$$

Based on Lemma G.11, we will discuss the gradients of $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ on different samples $\mathbf{Z}^{2,1}$. Since $\Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}},\Lambda^{(\widetilde{T})}_{5,j_2',r_{g_2\cdot y_0}}$ is guranteed 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 G.13, at time $\widetilde{T}$, we have

$$\mathbf{logit}^{(\widetilde{T})}_{5,j_2'}=\frac{e^{\frac{1.0005\log d}{2B}\cdot 2B}}{e^{\frac{1.0005\log d}{2B}\cdot 2B}+O(d)}\Big(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2}\Big)=\Big(1-O\big(1/d^{0.0005}\big)\Big)\Big(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2}\Big).$$

Then by Lemma G.11, we can obtain the gradient on the event $\mathcal{E}_1$ as follows:

$$\begin{aligned}&\mathbb{E}\Big[\Big(\mathcal{N}^{(\widetilde{T})}_{s,3,2,i}+\mathcal{N}^{(\widetilde{T})}_{s,3,2,ii}+\mathcal{N}^{(\widetilde{T})}_{s,3,2,iii}\Big)\mathbb{1}_{\mathcal{E}_1}\Big]\\&=\mathbb{E}\Bigg[\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\Big(\big(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2}\big)\cdot\Big(\psi_{j_2,r_{g_2\cdot y_1}}(g_2)-\Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\pm\widetilde{O}(\sigma_0)\Big)\\&\quad-\Big(1-O\Big(\frac{1}{d^{0.0005}}\Big)\Big)\big(1-\mathbf{logit}^{(t)}_{5,j_2}\big)\cdot\Big(\psi_{j_2',r_{g_2\cdot y_0}}(g_2)-\Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_0}}\pm\widetilde{O}(\sigma_0)\Big)\\&\quad\pm\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Bigg]\end{aligned}$$

$$= \mathbb{E}\Bigg[\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\Bigg(\Big(1-O\Big(\frac{1}{d^{0.0005}}\Big)\Big)(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2})\cdot\Big(\Lambda^{(\widetilde{T})}_{5,j_2',r_{g_2\cdot y_0}}-\Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\Big)$$
$$+O\Big(\frac{1}{d^{0.0005}}\Big)(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2})\cdot\Big(\psi_{5,j_2,r_{g_2\cdot y_1},2}(g_2)-\Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\Big)\Bigg)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Bigg],$$

on the other hand, we have

$$\mathbb{E}\Big[\Big(\mathcal{N}^{(\widetilde{T})}_{s,4,2,i}+\mathcal{N}^{(\widetilde{T})}_{s,4,2,ii}+\mathcal{N}^{(\widetilde{T})}_{s,4,2,iii}\Big)\mathbb{1}_{\mathcal{E}_1}\Big]$$
$$= \mathbb{E}\Bigg[\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\Bigg((1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2})\cdot\Big(\psi_{j_2,r_{g_2\cdot y_1}}(y_1)-\Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\pm\widetilde{O}(\sigma_0)\Big)$$
$$-\Big(1-O\Big(\frac{1}{d^{0.0005}}\Big)\Big)(1-\mathbf{logit}^{(t)}_{5,j_2})\cdot\Big(\psi_{j_2',r_{g_2\cdot y_0}}(y_1)-\Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_0}}\pm\widetilde{O}(\sigma_0)\Big)$$
$$\pm\widetilde{O}(\sigma_0^q)\Bigg)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Bigg]$$
$$= \mathbb{E}\Bigg[\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\Bigg(\Big(1-O\Big(\frac{1}{d^{0.0005}}\Big)\Big)(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2})$$
$$\cdot\Big(-\psi_{j_2',r_{g_2\cdot y_0}}(y_1)+\Lambda^{(\widetilde{T})}_{5,j_2',r_{g_2\cdot y_0}}-\Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\Big)$$
$$-O\Big(\frac{1}{d^{0.0005}}\Big)(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2})\cdot\Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\Bigg)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Bigg].$$

Since at time $\widetilde{T}$, we have $\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},1}-\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},1}\le\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2}-\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},1}\le\frac{1.005\log d}{2B}$, we have $\Lambda^{(\widetilde{T})}_{5,j_2',r_{g_2\cdot y_0}}-\Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\ll-\psi_{j_2',r_{g_2\cdot y_0}}(y_1)$. Therefore,

$$\Big[-\nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,4}}\mathsf{Loss}^{2,2}_5\Big]_{s,s}+\Big[\nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,3}}\mathsf{Loss}^{2,2}_5\Big]_{s,s}\ge\Omega\Big(\frac{B}{d}\Big)\cdot\mathbb{E}\Big[1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2}\big|\tau(x_1)=s\Big]$$

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 increasing. □

G.3.5 At the End of Stage 1.2.1

Lemma G.23. *For all iterations $t\le T_{1,2,1,s}=O\Big(\frac{n_yd}{\eta\log d}\Big)+O\Big(\frac{d^{\frac{(0.1+c_1/2)B}{\log d}}}{\eta\log d}\Big)$, we have Induction G.1 holds, and at time $T_{1,2,1,s}$, we have*

(a) $[\mathbf{Q}^{(T_{1,2,1,s})}_{4,3}]_{s,s},[\mathbf{Q}^{(T_{1,2,1,s})}_{4,4}]_{s,s}\ge\Omega(1)$;

(b) other $|[\mathbf{Q}^{(T_{2,1,s})}_{4,p}]_{s,s'}|$ *for* $p\in\{3,4\}$, $s'\in\tau(\mathcal{X})\ne s$ *are at most* $\widetilde{O}(\frac{1}{d})$.

Proof. The existence of $T_{1,2,1,s}$ can be directly obtained by using Lemmas G.19 to G.22. Furthermore, $[\mathbf{Q}^{(T_{1,2,1,s})}_{4,4}]_{s,s}\ge\Omega(1)$ can be guaranteed since Lemma G.22 implies that $\mathbf{Attn}^{(T_{1,2,1,s})}_{\mathsf{ans},1\to\mathsf{ans},1}\ge0.3-\frac{c_1}{2}\pm\widetilde{O}(1/d)>0.2$, which means that $[\mathbf{Q}^{(T_{1,2,1,s})}_{4,4}]_{s,s}$ should at least grow to a constant level compared to $|[\mathbf{Q}^{(T_{1,2,1,s})}_{4,p}]_{s,s'}|=\widetilde{O}(1/d)$ with $p\in\{3,4\}$.

We will handle $[\mathbf{Q}^{(T_{1,2,1,s})}_{4,3}]_{s,s}$ by means of a proof by contradiction. Suppose that $\mathbf{Attn}^{(T_{1,2,1,s})}_{\mathsf{ans},1\to\mathsf{ans},1}\ge0.4-\widetilde{c}$, where $\widetilde{c}=\frac{\log d}{8B}$ is a sufficiently small constant. Then denote $\widetilde{T}$ the first time that $\mathbb{E}\Big[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\big|\tau(x_1)=s\Big]\ge0.4-2\widetilde{c}$. Notice that $\Big|\Big[\mathbf{Q}^{(t)}_{4,p}\Big]_{s,s'}\Big|\le O\Big(\frac{\big[\mathbf{Q}^{(t)}_{4,p}\big]_{s,s}}{d}\Big)$ for

$p \in \{3,4\}$, thus for any sample $\mathbf{Z}^{2,1}$ satisfying $\tau(x_1) = s$, at time $\widetilde{T}$, we have

$$\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},1} \geq 0.4 - 2\widetilde{c} \pm \widetilde{O}(1/d).$$

- If $\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},0} \geq 2\varrho$, then
  - for $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, $\Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2}\cdot y_1}$ has already been activated to the linear regime. Furthermore, $\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},0} \leq 0.2 + 2\widetilde{c} - 0.2 \leq 2\widetilde{c}$, which implies $1 - \mathbf{logit}^{(1)}_{5,j_2} = 1 - o(1)$. Thus, by Lemma G.11, we have

    $$\mathbb{E}\Big[\Big(\mathcal{N}^{(\widetilde{T})}_{s,3,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,iii}\Big)\mathbb{1}_{\mathcal{E}_1}\Big] \geq \Omega\Big(\frac{B}{d}\Big),$$

    while

    $$\mathbb{E}\Big[\Big(\mathcal{N}^{(\widetilde{T})}_{s,4,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,iii}\Big)\mathbb{1}_{\mathcal{E}_1}\Big] \leq -\Omega\Big(\frac{B}{d}\Big).$$

  - for $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, $\Lambda^{(\widetilde{T})}_{5,j_2',r_{g_2}\cdot y_0} = (\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2} + \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},0})2B \leq (0.4 + \frac{4}{3}\widetilde{c} - 0.4 + 2\widetilde{c})2B = \frac{5\log d}{6}$. Thus $\mathbf{logit}^{(1)}_{5,j_2} = O(\frac{1}{d^{1/6}})$. Hence by Lemma G.11, we have

    $$\mathbb{E}\Big[\Big(\mathcal{N}^{(\widetilde{T})}_{s,3,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,iii}\Big)\mathbb{1}_{\mathcal{E}_1}\Big] \geq -O\Big(\frac{B}{d^{7/6}n_y}\Big),$$

    while

    $$\mathbb{E}\Big[\Big(\mathcal{N}^{(\widetilde{T})}_{s,4,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,iii}\Big)\mathbb{1}_{\mathcal{E}_1}\Big] \leq -\Omega\Big(\frac{B}{n_y d}\Big).$$

  - $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, we can use the following naive bounds for $p \in \{3,4\}$:

    $$|\mathbb{E}\Big[\Big(\mathcal{N}^{(\widetilde{T})}_{s,p,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,p,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,p,2,iii}\Big)\mathbb{1}_{\mathcal{E}_1}\Big]| \leq O\Big(\frac{B}{dn_y}\Big).$$

  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

  $$\Big[-\nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,3}}\mathsf{Loss}^{2,2}_5\Big]_{s,s} \geq \Omega(\frac{B}{d}), \Big[-\nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,4}}\mathsf{Loss}^{2,2}_5\Big]_{s,s} \leq -\Omega(\frac{B}{d}),$$

  which implies $\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},1}$ cannot further increase above $0.4 - 2\widetilde{c}$.

- If $\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},0} < 2\varrho$, 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

    $$\mathbb{E}\Big[\Big(\mathcal{N}^{(\widetilde{T})}_{s,3,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,iii}\Big)\mathbb{1}_{\mathcal{E}_1}\Big] \geq -\widetilde{O}(\sigma_0^q),$$

    while

    $$\mathbb{E}\Big[\Big(\mathcal{N}^{(\widetilde{T})}_{s,4,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,iii}\Big)\mathbb{1}_{\mathcal{E}_1}\Big] \leq \widetilde{O}(\sigma_0^q).$$

  - for $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, similarly as previous case, we have

    $$\mathbb{E}\Big[\Big(\mathcal{N}^{(\widetilde{T})}_{s,3,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,iii}\Big)\mathbb{1}_{\mathcal{E}_2}\Big] \geq -O\Big(\frac{B}{d^{7/6}n_y}\Big),$$

    while

    $$\mathbb{E}\Big[\Big(\mathcal{N}^{(\widetilde{T})}_{s,4,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,iii}\Big)\mathbb{1}_{\mathcal{E}_2}\Big] \leq O\Big(\frac{B}{d^{7/6}n_y}\Big).$$

– $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, $\Big|\Lambda^{(\widetilde{T})}_{5,j_2',r_{g_1\cdot y_1}} - \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\Big| \le 4\rho B = o(1)$, hence, we have

$$\mathbb{E}\Big[\Big(\mathcal{N}^{(\widetilde{T})}_{s,3,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,iii}\Big)\mathbb{1}_{\mathcal{E}_3}\Big] \ge \Omega\Big(\frac{B}{dn_y}\Big),$$

while

$$\mathbb{E}\Big[\Big(\mathcal{N}^{(\widetilde{T})}_{s,4,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,iii}\Big)\mathbb{1}_{\mathcal{E}_3}\Big] \le O\Big(\frac{B\log\log d}{dn_y\cdot\log d}\Big).$$

Putting them together, again we can conclude that

$$\Big[-\nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,3}}\mathsf{Loss}^{2,2}_5\Big]_{s,s} \ge \Omega(\frac{B}{n_y d}), \Big[-\nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,4}}\mathsf{Loss}^{2,2}_5\Big]_{s,s} \le O\Big(\frac{B\log\log d}{dn_y\cdot\log d}\Big),$$

which implies $\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},1}$ cannot further increase above $0.4 - 2\widetilde{c}$.

Consequently, this leads to a contradiction, and $\mathbf{Attn}^{(T_{1,2,1,s})}_{\mathsf{ans},1\to\mathsf{ans},1} < 0.4 - \widetilde{c}$, where $\widetilde{c} = \frac{\log d}{8B}$. Then it would follow that $\mathbf{Attn}^{(T_{1,2,1,s})}_{\mathsf{ans},1\to\mathsf{pred},2} \ge \widetilde{c}$, and thus $[\mathbf{Q}^{(T_{1,2,1,s})}_{4,3}]_{s,s} \ge \Omega(1)$. □

G.4 Stage 1.2.2: Convergence with Small Wrong Attention

Recall that $\mathsf{Loss}^{2,2}_{5,s} = -\mathbb{E}\Big[\log p_F(\mathbf{Z}_{\mathsf{ans},2,5}|\mathbf{Z}^{(2,1)})\big|\tau(x_1) = s\Big]$ for $s \in \tau(\mathcal{X})$.

Induction G.2. *Given $s \in \tau(\mathcal{X})$, let $T_{1,2,2,s}$ denote the first time that $\mathsf{Loss}^{2,2}_{5,s}$ decreases below $\Theta\Big(e^{(-\frac{1}{2}+3.01c_1)\cdot 2B}\Big)$. For all iterations $T_{1,2,1,s} \le t < T_{1,2,2,s}$, we have the following holds*

(a) $\Big[\mathbf{Q}^{(t)}_{4,3}\Big]_{s,s} + \Big[\mathbf{Q}^{(t)}_{4,4}\Big]_{s,s}$ *monotonically increases;*

(b) for $p \in \{3,4\}$, for $j \in \tau(\mathcal{X}) \ne s$, $|[\mathbf{Q}^{(t)}_{4,p}]_{s,j}| \le O(\frac{[\mathbf{Q}^{(t)}_{4,p}]_{s,j}}{d})$;

(c) for any sample $\mathbf{Z}^{2,1}$, we have $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} \le c_1$ for some sufficiently small constant $c_1 = \frac{1.005\log d}{B} > 0$;

(d) for any $\mathbf{Z}^{2,1}$, we have $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \le \min\Big\{\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - c_2, 0\Big\}$, where $c_2 = \frac{\log d}{4B} > 0$ is some sufficiently small constant.

G.4.1 Attention and Lambda Preliminaries

Lemma G.24. *If Induction G.2 holds for all iterations $[T_{1,2,1,s}, t)$,then for any sample $\mathbf{Z}^{2,1}$, we have*

1. $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} \in \Big[\frac{4}{3}c_1, 0.4 \pm \widetilde{O}\big(\frac{1}{d}\big)\Big]$;
2. $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} \ge \Omega(1)$;
3. $\big|\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\big| \le \widetilde{O}\big(\frac{1}{d}\big)$;

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

$$\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}, \quad \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} = c + \widetilde{O}\Big(\frac{1}{d}\Big).$$

- If $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \ge \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}$, then

– If $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, we have

$$\begin{aligned}
&\log p_F(\mathbf{Z}_{\mathsf{ans},2,5}|\mathbf{Z}^{(2,1)}) \\
&\le \Theta(1) \cdot \max\Bigg\{ e^{\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\right)2B - \left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right)2B}, \\
&\qquad\qquad e^{\log d - \left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right)2B} \Bigg\} \\
&\le \Theta\left(e^{2c_1 B - \left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right)2B} \right) \\
&\le \Theta\left(e^{-2\left(\frac{1}{2} - 2c - c_1\right)B} \right),
\end{aligned}$$

where the last inequality follows from the fact that $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \ge \frac{1}{2}(1-2c) = \frac{1}{2} - c$.

– If $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, we have

$$\begin{aligned}
&\log p_F(\mathbf{Z}_{\mathsf{ans},2,5}|\mathbf{Z}^{(2,1)}) \le \Theta(1) \\
&\cdot \max\Bigg\{ e^{\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\right)B - \left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right)2B}, \\
&\qquad\qquad e^{\log d - \left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right)2B} \Bigg\} \\
&\le \Theta\left(e^{2(c+c_1)B - \left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right)2B} \right) \\
&\le \Theta\left(e^{-2\left(\frac{1}{2} - 3c - c_1\right)B} \right).
\end{aligned}$$

– If $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, we have

$$\begin{aligned}
&\log p_F(\mathbf{Z}_{\mathsf{ans},2,5}|\mathbf{Z}^{(2,1)}) \\
&\le \Theta(1) \cdot \max\left\{ e^{\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\right)2B}, e^{\log d - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}2B} \right\} \\
&\le \Theta(1) \cdot \max\left\{ e^{-2(\frac{1}{2} - c_1 - c)B},\ e^{-2(\frac{1}{2} - 2c)B} \right\}.
\end{aligned}$$

• If $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \le \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}$, then

– If $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, we have

$$\begin{aligned}
\log p_F(\mathbf{Z}_{\mathsf{ans},2,5}|\mathbf{Z}^{(2,1)}) &\le \Theta(1) \cdot e^{\log d - \left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right)2B} \\
&\le \Theta\left(e^{2c_1 B - \left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right)2B} \right) \\
&\le \Theta\left(e^{-2\left(\frac{1}{2} - \frac{5}{2}c - \frac{c_2}{2} - c_1\right)B} \right),
\end{aligned}$$

where the last inequality uses the fact that $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \le c - c_2$, implying $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \ge \frac{1}{2} - \frac{3}{2}c + \frac{c_2}{2}$.

– If $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, we have

$$\begin{aligned}
&\log p_F(\mathbf{Z}_{\mathsf{ans},2,5}|\mathbf{Z}^{(2,1)}) \le \Theta(1) \cdot \\
&\max\Bigg\{ e^{\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\right)2B - \left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right)2B},
\end{aligned}$$

$$e^{\log d-\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right)2B}\Bigg\}$$
$$\le \Theta\left(e^{\max\{c,c_1\}2B-\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right)2B}\right)$$
$$\le \Theta\left(e^{-2\left(\frac{1}{2}-2c-\max\{c,c_1\}\right)B}\right).$$

– If $\mathbf{Z}^{2,1}\in\mathcal{E}_3$, we have

$$\log p_F(\mathbf{Z}_{\mathsf{ans},2,5}|\mathbf{Z}^{(2,1)})$$
$$\le \Theta(1)\cdot\max\left\{e^{\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\right)2B}, e^{\log d-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}2B}\right\}$$
$$\le \Theta(1)\cdot\max\left\{e^{-2(\frac{1}{2}-c_1-c)B},\ e^{-2(\frac{1}{2}-2c)B}\right\}.$$

Putting all cases together, if $c\le\frac{2}{3}c_1$, we must have

$$\mathsf{Loss}^{2,2}_{5,s}\ \le\ \Theta\left(e^{-2(\frac{1}{2}+3\cdot\frac{2c_1}{3}+c_1)B}\right)=\Theta\left(e^{(-\frac{1}{2}+3c_1)2B}\right),$$

which contradicts the definition of $T_{1,2,2,s}$. Therefore, it must hold that $c\ge\frac{2}{3}c_1$ for all $t\le T_{1,2,2,s}$.

□

Lemma G.25. *If Induction G.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*
$$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}}=\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right)\cdot 2B+O\Big(\frac{B}{n_y}\Big)+O(\delta).$$
2. *for the prediction $j_2'=\tau\big(g_2(y_0)\big)$, we have*
$$\Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}}=\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\right)\cdot 2B+\widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big)+O(\delta).$$
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*
$$\Lambda^{(t)}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}}=\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\right)\cdot\frac{2B}{n_y}+\widetilde{O}\Big(\frac{B}{d}\Big)+O(\delta).$$
5. *for other $j\in\tau(\mathcal{Y})$, if there exsit j and y, s.t., $g_1,g_2\in\mathrm{Fiber}_{j,y}$ (notice that $y\ne y_0,y_1$ 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*
$$\Lambda^{(t)}_{5,j,r_{g_2\cdot y}}=\left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\right)\cdot 2B+\widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big)+O(\delta);$$
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 $r\in\cup_{\ell,\ell'\in[2]}\Big(\widehat{\mathfrak{A}}_{\tau(g_\ell\cdot y_{\ell'-1})}\setminus\{r_{g_\ell\cdot y_{\ell'-1}}\}\Big)$ cannot be activated.

Lemma G.26. *If Induction G.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*

$$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{n_y}\Big) + O(\delta).$$

2. *for the prediction $j_2' = \tau\big(g_2(y_0)\big)$, $r \in \widehat{\mathfrak{A}}_{j_2'} \setminus \{r_{g_2\cdot y_0}\}$ (notice that in this case $r_{g_2\cdot y_0} = r_{g_1\cdot y_0}$) cannot be activated, moreover, we have*

$$\Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + O(\delta).$$

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*

$$\Lambda^{(t)}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\Big)\cdot \frac{2B}{n_y} + O(\delta).$$

4. *for other $j \in \tau(\mathcal{Y})$, if there exsit j and y, s.t., $g_1, g_2 \in \mathrm{Fiber}_{j,y}$ (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*

$$\Lambda^{(t)}_{5,j,r_{g_2\cdot y}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta);$$

else, none of $r \in \widehat{\mathfrak{A}}_j$ can be activated.

Lemma G.27. *If Induction G.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*

$$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot 2B \pm O(\delta).$$

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*

$$\Lambda^{(t)}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}} = \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot 2B \pm O(\delta).$$

3. *for other $j = g(y_1)$, where $g_1, g_2 \notin \mathrm{Fiber}_{j,y_1}$, we have*

$$\Lambda^{(t)}_{5,j,r_{g\cdot y_1}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\Big)\cdot \frac{2B}{n_y} + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta);$$

moreover, if there exsit y, s.t., $g_1, g_2 \in \mathrm{Fiber}_{j,y}$ (notice that $y \neq y_1$ for $\mathcal{E}_3$), we have

$$\Lambda^{(t)}_{5,j,r_{g_2\cdot y}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta);$$

besides, other $r \in \widehat{\mathfrak{A}}_j$ cannot be activated.

Lemma G.28. *If Induction G.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*

$$\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot 2B + O(\delta).$$

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*

$$\Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}} = \Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + O(\delta).$$

3. *for other $j \in \tau(\mathcal{Y})$, if there exsit j and y, s.t., $g_1, g_2 \in \mathrm{Fiber}_{j,y}$, then for such a j, $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$ cannot be activated, moreover, we have*

$$\Lambda_{5,j,r_{g_2\cdot y}}^{(t)} = \Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta);$$

else, none of $r \in \widehat{\mathfrak{A}}_j$ can be activated.

Lemma G.29. *If Induction G.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*

$$\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)} = \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\cdot 2B \pm O(\delta).$$

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*

$$\Lambda_{5,j_2',r_{g_2\cdot y_0}}^{(t)} = \Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta).$$

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 exsit j and y, s.t., $g_1, g_2 \in \mathrm{Fiber}_{j,y}$, then for such a j, $r \in \widehat{\mathfrak{A}}_j \setminus \{r_{g_2 \cdot y}\}$ cannot be activated, moreover, we have*

$$\Lambda_{5,j,r_{g_2\cdot y}}^{(t)} = \Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta);$$

else, none of $r \in \widehat{\mathfrak{A}}_j$ can be activated.

Lemma G.30. *If Induction G.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*

$$\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)} = \Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\Big)\cdot 2B \pm O(\delta).$$

2. *for other $j = g(y_1)$, where $g_2 \notin \mathrm{Fiber}_{j,y_1}$, we have $r_{g\cdot y_1}$ cannot be activated, moreover, if there exsit $y \neq y_1$, s.t., $g_1, g_2 \in \mathrm{Fiber}_{j,y}$, we have*

$$\Lambda_{5,j,r_{g_2\cdot y}}^{(t)} = \Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}\Big)\cdot 2B + \widetilde{O}\Big(\frac{B}{d\cdot n_y}\Big) + O(\delta);$$

besides, other $r \in \widehat{\mathfrak{A}}_j$ cannot be activated.

G.4.2 Gradient Lemma

Lemma G.31. *If Induction G.2 holds for all iterations $t \in [T_{1,2,1,s}, T_{1,2,2,s})$, given $s \in \tau(\mathcal{X})$, if $\mathbb{E}\big[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} \mid \tau(x_1) = s\big] \le \frac{1}{2}$ we have*

$$\Big[-\nabla_{\mathbf{Q}_{4,3}^{(t)}}\mathsf{Loss}_5^{2,2}\Big]_{s,s} + \Big[-\nabla_{\mathbf{Q}_{4,4}^{(t)}}\mathsf{Loss}_5^{2,2}\Big]_{s,s}$$

$$\ge \Omega\Big(\frac{B}{d}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\big|\tau(x_1)=s,\mathcal{E}_1\Big] + \Omega\Big(\frac{B}{n_y d}\Big)\cdot\sum_{m=2}^{3}\mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\big|\tau(x_1)=s,\mathcal{E}_m\Big].$$

Proof. The analysis follows the similar idea as Lemma G.19, while Induction G.2 ensures that $\mathbb{E}\big[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)} \mid \tau(x_1) = s\big] \ge \Omega(1) > \frac{\varrho}{B}$, which falls into the regime (ii). Thus for the main event $\mathcal{E}_1$, the term $\Lambda^{(t)}5, j_2, rg_2 \cdot y_1$ is guaranteed to remain within the linear regime. We can first directly upper bound the term $\mathcal{N}_{s,2,iii}^{(t)}$ by $\widetilde{O}(\sigma_0^q)$.

(a) For $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, by Induction G.2 and Lemma G.25, we can obtain

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}_{s,2,i}^{(t)}\mathbb{1}_{\mathcal{E}_1}\Big] \\
&= \mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\cdot\Big(\Big(-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(y_0)-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(g_1) \\
&\quad+\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Big] \qquad (108)\\
&\geq \Omega\Big(\frac{B}{d}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\big|\tau(x_1)=s,\mathcal{E}_1\Big].
\end{aligned}$$

Moving to $\mathcal{N}_{s,2,ii}^{(t)}$, by Lemma G.25, we have

- for $j=\tau(g_1(y_1))$, $r=r_{g_1\cdot y_1}$, due to the cancellation of $\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j,r}(y_0)+\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{j,r}(g_1)$, we have

$$\begin{aligned}
&\Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j,r}(y_0)+\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{j,r}(g_1) \\
&\qquad-\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j,r}^{(t)}\pm\widetilde{O}(\sigma_0)\Big) \\
&\geq-\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j,r}^{(t)}.
\end{aligned}$$

- for $j=\tau(g_2(y_0))$, $r=r_{g_2\cdot y_0}$

$$\begin{aligned}
&\Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j,r}(y_0)+\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{j,r}(g_1) \\
&\qquad-\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j,r}^{(t)}\pm\widetilde{O}(\sigma_0)\Big) \\
&\geq-\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j,r}^{(t)}.
\end{aligned}$$

- for $j=g_1(y)$, where $\exists y\neq y_0,y_1$, s.t., $g_2(y)=g_1(y)$, we have

$$\begin{aligned}
&\Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j,r}(y_0)+\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{j,r}(g_1) \\
&\qquad-\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j,r}^{(t)}\pm\widetilde{O}(\sigma_0)\Big) \\
&\geq-\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j,r}^{(t)}.
\end{aligned}$$

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

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}_{s,2,ii}^{(t)}\mathbb{1}_{\mathcal{E}_1}\Big] \\
&=\mathbb{E}\Big[\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}^{(t)}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot \\
&\quad\Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j,r}(y_0)+\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{j,r}(g_1) \\
&\quad-\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j,r}^{(t)}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Big] \\
&\geq-\mathbb{E}\Big[\Big(1-\mathbf{logit}_{5,j_2}^{(t)}\Big)\cdot\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big) \\
&\quad\cdot\max_{y\neq y_0,y_1}\Big\{\Lambda_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}}^{(t)},\Lambda_{5,\tau(g_2(y_0)),r_{g_2\cdot y_0}}^{(t)}\Big\}\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\Big],
\end{aligned}$$

where the last inequality is due to the fact that $\Lambda^{(t)}_{5,\tau(g_2(y_0)),r_{g_2\cdot y_0}} = \Lambda^{(t)}_{5,\tau(g_2(y)),r_{g_2\cdot y}} \pm \widetilde{O}(\frac{B}{d\cdot n_y})$ for $y \neq y_0, y_1$ s.t., $g_1(y) = g_2(y)$. Notice that

$$\max_{y\neq y_0,y_1}\left\{\Lambda^{(t)}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}}, \Lambda^{(t)}_{5,\tau(g_2(y_0)),r_{g_2\cdot y_0}}\right\} \leq \max\left\{\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}, \Theta(\frac{1}{n_y})\right\}2B,$$

while

$$\begin{aligned}(108) &\geq \mathbb{E}\left[\left(1-\mathbf{logit}^{(t)}_{5,j_2}\right)\cdot \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot 2B\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\right]\\ &= \mathbb{E}\left[\left(1-\mathbf{logit}^{(t)}_{5,j_2}\right)\cdot \left(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\right)B\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_1}\right]\end{aligned}$$

Since by Induction G.2, $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} \leq c_1 \ll \frac{1}{2}$, thus we have

$$\mathbb{E}\left[\mathcal{N}^{(t)}_{s,2,i}\mathbb{1}_{\mathcal{E}_1}\right] + \mathbb{E}\left[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_1}\right] \geq \Omega\Big(\frac{B}{d}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1)=s,\mathcal{E}_1\Big].$$

(b) For $\mathbf{Z}^{2,1}\in\mathcal{E}_2$, $\mathcal{N}^{(t)}_{s,2,i}$ can be bounded analogously to (108), yielding

$$\mathbb{E}\left[\mathcal{N}^{(t)}_{s,2,i}\mathbb{1}_{\mathcal{E}_2}\right] \geq \Omega\Big(\frac{B}{d\cdot n_y}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1)=s,\mathcal{E}_2\Big] \geq 0.$$

Moving to $\mathcal{N}^{(t)}_{s,2,ii}$,

- For $j=\tau(g_1(y_1))$, we obtain

$$\begin{aligned}&= \mathbb{E}\Bigg[\mathbf{logit}^{(t)}_{5,j}\Big(\mathbf{sReLU}'(\Lambda_{5,j,r_{g_1\cdot y_1}})\cdot\\ &\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r_{g_1\cdot y_1}}(y_0) + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j,r_{g_1\cdot y_1}}(g_1) \qquad (109)\\ &- \left(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\right)\Lambda^{(t)}_{5,j,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_2}\Bigg]\\ &\geq -O\left(\frac{B}{n_y^2\cdot d}\right)\cdot\mathbb{E}\Big[(1-\mathbf{logit}^{(t)}_{5,j_2})\,\big|\,\tau(x_1)=s,\mathcal{E}_2\Big],\end{aligned}$$

  where the inequality follows from the cancellation of the term

$$\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r_{g_1\cdot y_1}}(y_0) + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j,r_{g_1\cdot y_1}}(g_1),$$

  together with the fact that

$$\Lambda^{(t)}_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}} = \left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\right)\cdot\frac{2B}{n_y} + O(\delta) \;\leq\; O\left(\frac{B}{n_y}\right).$$

- For $j_2' = \tau(g_2(y_0)) = \tau(g_1(y_0))$, it is clear that

$$\begin{aligned}&\mathbb{E}\Bigg[\mathbf{logit}^{(t)}_{5,j_2'}\Big(\mathbf{sReLU}'(\Lambda^{(t)}_{5,j_2',r_{g_1\cdot y_0}})\cdot\\ &\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j_2',r_{g_1\cdot y_0}}(y_0) + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j_2',r_{g_1\cdot y_0}}(g_1)\\ &- \left(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\right)\Lambda^{(t)}_{5,j_2',r_{g_1\cdot y_0}} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_2}\Bigg]\\ &\geq 0,\end{aligned}$$

where the last inequality is due to the fact that

$$\begin{aligned}
&\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j_2',r_{g_1\cdot y_0}}(g_1)-\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j_2',r_{g_1\cdot y_0}}\\
&\geq \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\Big(1-2\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\Big)2B\\
&\geq \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\Big(1-2\Big(c_1+0.2\Big)\Big)2B\geq 0.
\end{aligned}$$

- 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
  $$\mathbb{E}\big[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\mid\tau(x_1)=s\big]\leq\frac{c_1}{4},$$
  then by Induction G.2 and Lemma G.26, we have we have $\Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}}-\Lambda^{(t)}_{5,j,r_{g_2\cdot y}}\geq\big(-\frac{c_1}{6}+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\big)2B\geq(-\frac{c_1}{4}+\frac{3}{4}c_1)2B=c_1B$, which implies
  $$\sum_{y\neq y_0,y_1}\mathbf{logit}^{(t)}_{5,\tau(g_2(y))}\mathbb{1}_{g_1(y)=g_2(y)}\ll\widetilde{O}\big(\tfrac{1}{d^{1/2}}\big)\cdot\mathbf{logit}^{(t)}_{5,j_2'}.$$
  Consequently,
  $$\begin{aligned}
  &\sum_{y\neq y_0,y_1}\Bigg|\mathbb{E}\Bigg[\mathbf{logit}^{(t)}_{5,j}\Big(\mathbf{sReLU}'(\Lambda^{(t)}_{5,j,r_{g_1\cdot y}})\cdot\\
  &\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r_{g_1\cdot y}}(y_0)+\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j,r_{g_1\cdot y}}(g_1)\\
  &-\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j,r_{g_1\cdot y}}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\\
  &\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_2}\mathbb{1}_{g_1(y)=g_2(y)}\Bigg]\Bigg|\leq O\big(\tfrac{1}{d^{1/2}}\big)\cdot(109).
  \end{aligned}$$
  - If
  $$\mathbb{E}\big[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\mid\tau(x_1)=s\big]\geq\frac{c_1}{4},$$
  then by Lemma G.26, the activation $\Lambda^{(t)}_{5,j,r_{g_1\cdot y}}$ cannot be triggered.

Putting the above discussion together, we have

$$\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_1}\Big]+\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,ii}\mathbb{1}_{\mathcal{E}_2}\Big]\geq\Omega\Big(\frac{B}{d\cdot n_y}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}^{(t)}_{5,j_2})\big|\tau(x_1)=s,\mathcal{E}_2\Big].$$

(c) For $\mathbf{Z}^{2,1}\in\mathcal{E}_3$, the gradient admits an analysis analogous to that in Lemma G.19. In particular, by Induction G.2 and Lemma G.27, we first obtain the following logit bound:

$$\mathbf{logit}^{(t)}_{5,j}\ \leq\ \frac{1}{\mathsf{poly}d}\cdot\Big(1-\mathbf{logit}^{(t)}_{5,j_2}\Big),\qquad j\neq j_2,\ \tau\big(g_1(y_1)\big).$$

Therefore,

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}^{(t)}_{s,2,i}\,\mathbb{1}_{\mathcal{E}_3}\Big]\\
&=\mathbb{E}\Bigg[(1-\mathbf{logit}^{(t)}_{5,j_2})\cdot\Big(-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(y_1)-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(g_1)\\
&\qquad+\big(1-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}}\pm\widetilde{O}(\sigma_0)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_3}\Bigg]\\
&\geq\ \Omega\Big(\frac{B}{n_y\cdot d}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}^{(t)}_{5,j_2})\ \big|\ \tau(x_1)=s,\,\mathcal{E}_3\Big].
\end{aligned}$$

Moving to $\mathcal{N}_{s,2,ii}^{(t)}$, we have

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}_{s,2,ii}^{(t)}\,\mathbb{1}_{\mathcal{E}_3}\Big] \\
&\geq \mathbb{E}\Bigg[\mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)}\cdot\Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{\tau(g_1(y_1)),r_{g_1\cdot y_1}}(y_1)+\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}\cdot\psi_{\tau(g_1(y_1)),r}(g_1) \\
&-\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,\tau(g_1(y_1)),r_{g_1\cdot y_1}}^{(t)}\pm\widetilde{O}(\sigma_0)\pm\widetilde{O}(\delta^q)\Big)\,\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_3}\Bigg] \\
&\quad-O\Big(\frac{B}{d\cdot\mathsf{poly}d}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\;\big|\;\tau(x_1)=s,\,\mathcal{E}_3\Big].
\end{aligned}$$

Combining the two bounds, we conclude

$$\mathbb{E}\Big[\mathcal{N}_{s,2,i}^{(t)}\,\mathbb{1}_{\mathcal{E}_3}\Big]+\mathbb{E}\Big[\mathcal{N}_{s,2,ii}^{(t)}\,\mathbb{1}_{\mathcal{E}_3}\Big]\;\geq\;\Omega\Big(\frac{B}{n_y\cdot d}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\;\big|\;\tau(x_1)=s,\,\mathcal{E}_3\Big].$$

(d) For $\mathbf{Z}^{2,1}\in\mathcal{E}_4$, in analogy with Lemma G.19, we obtain the following crude bound:

$$\mathbb{E}\Big[\mathcal{N}_{s,2,i}^{(t)}\,\mathbb{1}_{\mathcal{E}_4}\Big]+\mathbb{E}\Big[\mathcal{N}_{s,2,ii}^{(t)}\,\mathbb{1}_{\mathcal{E}_4}\Big]\;\geq\;0.$$

(e) For $\mathbf{Z}^{2,1}\in\mathcal{E}_5\cup\mathcal{E}_6$, by Induction G.2, Lemma G.29 and Lemma G.30, we obtain the following logit condition:

$$1-\mathbf{logit}_{5,j_2}^{(t)},\;\;\mathbf{logit}_{5,j}^{(t)}\;\leq\;\frac{1}{\mathsf{poly}d}\cdot\mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\;\big|\;\tau(x_1)=s,\,\mathcal{E}_1\Big],\qquad j\neq j_2.$$

Hence, $\mathcal{N}_{s,2,i}^{(t)}$ and $\mathcal{N}_{s,2,ii}^{(t)}$ can be bounded as

$$\begin{aligned}
&\Big|\mathbb{E}[\mathcal{N}_{s,2,i}^{(t)}\,\mathbb{1}_{\mathcal{E}_m}]\Big|,\quad\Big|\mathbb{E}[\mathcal{N}_{s,2,ii}^{(t)}\,\mathbb{1}_{\mathcal{E}_m}]\Big| \\
&\qquad\leq O\Big(\frac{B}{dn_y}\cdot\frac{1}{\mathsf{poly}d}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\;\big|\;\tau(x_1)=s,\,\mathcal{E}_1\Big],\qquad m\in\{5,6\}.
\end{aligned}$$

Moreover, for $\mathbf{Z}^{2,1}\in\mathcal{E}_6$, if

$$\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}+\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}>\tfrac{1}{2},$$

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

$$\begin{aligned}
&\Big[-\nabla_{\mathbf{Q}_{4,3}^{(t)}}\mathsf{Loss}_5^{2,2}\Big]_{s,s}+\Big[-\nabla_{\mathbf{Q}_{4,4}^{(t)}}\mathsf{Loss}_5^{2,2}\Big]_{s,s}=\sum_{\kappa\in\{i,ii,iii\}}\sum_{m\in[6]}\mathbb{E}\Big[\mathcal{N}_{s,2,\kappa}^{(t)}\mathbb{1}_{\mathcal{E}_m}\Big] \\
&\geq\Omega\Big(\frac{B}{d}\Big)\cdot\mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\big|\tau(x_1)=s,\mathcal{E}_1\Big]+\Omega\Big(\frac{B}{n_yd}\Big)\cdot\sum_{m=2}^{3}\mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\big|\tau(x_1)=s,\mathcal{E}_m\Big].
\end{aligned}$$

□

Lemma G.32. *If Induction G.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*

$$\Big|\Big[-\nabla_{\mathbf{Q}_{4,p}^{(t)}}\mathsf{Loss}_5^{2,2}\Big]_{s,s'}\Big|\leq O\Big(\frac{1}{d}\Big)\Big|\Big[-\nabla_{\mathbf{Q}_{4,p}^{(t)}}\mathsf{Loss}_5^{2,2}\Big]_{s,s}\Big|.$$

G.4.3 Attention gap is small

Lemma G.33. *If Induction G.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*

$$\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} \le c_1,$$

where $c_1 > 0$ is a small constant.

The proof follows along the same lines as Lemma G.22, and hence the details are omitted for brevity.

Lemma G.34. *If Induction G.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*

$$\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \le \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - c_2,$$

where $c_2 = \frac{\log d}{4B} > 0$ is a small constant.

Proof. Let $\widetilde{T}$ denote the first time such that

$$\mathbb{E}\Big[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} \,\Big|\, \tau(x_1) = s\Big] \le \frac{\log d}{4.02B}.$$

Since $\big|[\mathbf{Q}^{(t)}_{4,p}]_{s,s'}\big| \le O\Big(\frac{[\mathbf{Q}^{(t)}_{4,p}]_{s,s}}{d}\Big)$ for $p \in \{3,4\}$, it follows that for any sample $\mathbf{Z}^{2,1}$ with $\tau(x_1) = s$, at time $\widetilde{T}$ we have

$$\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} \ge \frac{\log d}{4.02B} \pm \widetilde{O}(1/d).$$

By Lemma G.24, we also have $\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},0} \ge \Omega(1) \gg \varrho$. We then consider the following cases:

- If $\mathbf{Z}^{2,1} \in \mathcal{E}_1$, then $\Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}$ is already activated into the linear regime. Moreover, since $\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},1} \le \frac{\log d}{4.02B} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} < 0$, it follows that $\Lambda^{(\widetilde{T})}_{5,j_2',r_{g_2\cdot y_0}}$ cannot be activated. Thus, by Lemma G.11,

  $$\mathbb{E}\Big[\big(\mathcal{N}^{(\widetilde{T})}_{s,3,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,iii}\big)\mathbb{1}_{\mathcal{E}_1}\Big] \ge \Omega\big(\tfrac{B}{d}\big)\cdot\mathbb{E}\big[1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_2} \,\big|\, \tau(x_1) = s,\, \mathcal{E}_1\big],$$

  while

  $$\mathbb{E}\Big[\big(\mathcal{N}^{(\widetilde{T})}_{s,4,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,iii}\big)\mathbb{1}_{\mathcal{E}_1}\Big] \le -\Omega\Big(\tfrac{B}{dn_y}\Big)\cdot\mathbb{E}\big[1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_2} \,\big|\, \tau(x_1) = s,\, \mathcal{E}_1\big].$$

- If $\mathbf{Z}^{2,1} \in \mathcal{E}_2$, then

  $$\Lambda^{(\widetilde{T})}_{5,j_2',r_{g_2\cdot y_0}} = (\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2} + \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},0})2B \le \frac{\log d}{2.01}.$$

  Hence $\mathbf{logit}^{(1)}_{5,j_2} \le O(d^{-1.01/2.01})(1 - \mathbf{logit}^{(t)}_{5,j_2})$. By Lemma G.11,

  $$\mathbb{E}\Big[\big(\mathcal{N}^{(\widetilde{T})}_{s,3,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,iii}\big)\mathbb{1}_{\mathcal{E}_2}\Big] \ge \Omega\Big(\tfrac{B}{dn_y}\Big)\cdot\mathbb{E}\big[1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_2} \,\big|\, \tau(x_1) = s,\, \mathcal{E}_2\big],$$

  and

  $$\mathbb{E}\Big[\big(\mathcal{N}^{(\widetilde{T})}_{s,4,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,iii}\big)\mathbb{1}_{\mathcal{E}_2}\Big] \le -\Omega\Big(\tfrac{B}{dn_y}\Big)\cdot\mathbb{E}\big[1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_2} \,\big|\, \tau(x_1) = s,\, \mathcal{E}_2\big].$$

- If $\mathbf{Z}^{2,1} \in \mathcal{E}_3$, we can apply the crude bounds

  $$\mathbb{E}\Big[\big(\mathcal{N}^{(\widetilde{T})}_{s,3,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,3,2,iii}\big)\mathbb{1}_{\mathcal{E}_3}\Big] \ge 0,$$

  and

  $$\mathbb{E}\Big[\big(\mathcal{N}^{(\widetilde{T})}_{s,4,2,i} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,ii} + \mathcal{N}^{(\widetilde{T})}_{s,4,2,iii}\big)\mathbb{1}_{\mathcal{E}_3}\Big] \le 0.$$

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

$$\begin{aligned}\big[-\nabla_{\mathbf{Q}_{4,3}^{(\widetilde{T})}} \mathsf{Loss}_5^{2,2}\big]_{s,s} \;&\geq\; \Omega\big(\tfrac{B}{d}\big)\cdot \mathbb{E}\big[1-\mathbf{logit}_{5,j_2}^{(\widetilde{T})} \;\big|\; \tau(x_1)=s,\, \mathcal{E}_1\big] \\ &\quad + \Omega\Big(\frac{B}{dn_y}\Big)\cdot \mathbb{E}\big[1-\mathbf{logit}_{5,j_2}^{(\widetilde{T})} \;\big|\; \tau(x_1)=s,\, \mathcal{E}_2\big], \end{aligned} \tag{110}$$

$$\begin{aligned}\big[-\nabla_{\mathbf{Q}_{4,4}^{(\widetilde{T})}} \mathsf{Loss}_5^{2,2}\big]_{s,s} \;&\leq\; -\Omega\big(\tfrac{B}{d}\big)\cdot \mathbb{E}\big[1-\mathbf{logit}_{5,j_2}^{(\widetilde{T})} \;\big|\; \tau(x_1)=s,\, \mathcal{E}_1\big] \\ &\quad - \Omega\Big(\frac{B}{dn_y}\Big)\cdot \mathbb{E}\big[1-\mathbf{logit}_{5,j_2}^{(\widetilde{T})} \;\big|\; \tau(x_1)=s,\, \mathcal{E}_2\big]. \end{aligned} \tag{111}$$

Finally, observe that

$$\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} \;=\; \frac{e^{[\mathbf{Q}_{4,3}^{(t)}]_{s,s}} - e^{[\mathbf{Q}_{4,4}^{(t)}]_{s,s}} - e^{\widetilde{O}(1/d)}}{e^{[\mathbf{Q}_{4,3}^{(t)}]_{s,s}} + e^{[\mathbf{Q}_{4,4}^{(t)}]_{s,s}} + e^{\widetilde{O}(1/d)}}.$$

Hence, (110) and (111) together imply that

$$\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}$$

must decrease at time $\widetilde{T}+1$. □

G.4.4 At the End of Training

Lemma G.35 (At the end of stage 1.2.2)**.** *Induction G.2 holds for all iterations $T_{1,2,1,s} < t \leq T_{1,2,2,s} = O\Big(\frac{\mathsf{poly}(d)}{\eta B}\Big)$. At the end of training, we have*

(a) Attention concentration: $\mathbb{E}[\epsilon_{\mathsf{attn}} \mid \tau(x_1)=s] \leq 2.52c_1$ for some small constant $0 < c_1 = \frac{1.005\log d}{2B}$;

(b) Loss convergence: $\mathsf{Loss}_{5,s}^{2,2} \leq e^{(-\frac{1}{2}+3.01c_1)\cdot 2B} = \frac{1}{\mathsf{poly}d}$.

Proof. Lemma G.24 and Lemma G.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 G.31, Lemma G.32, Lemma G.33, and Lemma G.34, we conclude that there exists a stopping time

$$T_{1,2,2,s} = O\Big(\frac{\mathsf{poly}(d)}{\eta B}\Big),$$

and that Induction G.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

$$\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1}^{(t)}, \quad \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} \;=\; c + \widetilde{O}\big(\tfrac{1}{d}\big).$$

- If $\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} \geq \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}$, then for $\mathbf{Z}^{2,1} \in \mathcal{E}_1$ we have

$$\begin{aligned}\log p_F(\mathbf{Z}_{\mathsf{ans},2,5}|\mathbf{Z}^{(2,1)}) &\geq \Theta(1)\cdot e^{\log d - \left(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\right)2B} \\ &\geq \Theta\left(e^{-\left(\frac{1}{2}+\frac{c_1}{2}-2c-\frac{\log d}{B}\right)2B}\right), \end{aligned}$$

  where the last inequality follows from

$$\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} \leq \tfrac{1}{2}(1-2c+c_1) = \tfrac{1+c_1}{2} - c.$$

- If $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \le \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}$, then for $\mathbf{Z}^{2,1} \in \mathcal{E}_1$ we have

$$\begin{aligned}\log p_F(\mathbf{Z}_{\mathsf{ans},2,5}|\mathbf{Z}^{(2,1)}) &\ge \Theta(1)\cdot e^{\log d - \left(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\right)2B} \\ &\ge \Theta\left(e^{-\left(\frac{1}{2}-2c-\frac{\log d}{B}\right)2B}\right),\end{aligned}$$

  where the last inequality follows from the fact that $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \le \frac{1}{2} - c$.

Putting the above cases together, we obtain

$$\tfrac{1}{2} + \frac{c_1}{2} - 2c - \tfrac{\log d}{B} \;\ge\; \tfrac{1}{2} - 3.01c_1,$$

which implies

$$c \;\le\; \tfrac{1}{2}(3.02c_1 + 0.5c_1 - c_1) \;=\; 1.26c_1.$$

□

H Recursive Learning the Attention Layer: Symmetric Case

In this section, we analyze the recursive learning dynamics of the attention layer in the symmetric 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 H.1 (Greedy data annotator)**.** The greedy language model $\widehat{p}_F$ induced by a network F is defined as

$$\widehat{p}_F(Z_{\mathsf{ans},L'+1} \mid Z^{L,L'}) = \begin{cases} 1, & \text{if } Z_{\mathsf{ans},L'+1} = \operatorname{argmax}_Z p_F(Z \mid Z^{L,L'}), \\ 0, & \text{otherwise.} \end{cases} \tag{112}$$

Definition H.2 (Bootstrapped LEGO distribution)**.** We define $\mathcal{D}_F^{L,L'}$ as the LEGO distribution in Assumption 3.2, except that the answers $Z_{\mathsf{ans},\ell}, 1 \le \ell \le L'$ are generated recursively by sampling $Z_{\mathsf{ans},\ell} \sim \widehat{p}_F(\cdot \mid Z^{L,\ell-1})$ from the greedy language model $\widehat{p}_F$.

Definition H.3 (Self-training loss)**.** Given a (fixed) model $\widetilde{F}$ 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 $\mathcal{D}_{\widetilde{F}}^{L,L'}$ from Definition H.2:

$$\mathsf{Loss}_{\widetilde{F}}^{L,L'}(F) \triangleq \mathbb{E}_{Z^{L,L'}\sim\mathcal{D}_{\widetilde{F}}^{L,L'}}\Big[-\log p_F(Z_{\mathsf{ans},L',i} \mid Z^{L,L'-1})\Big]. \tag{113}$$

Similarly, the per-token loss is defined by

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

At stage $k \ge 2$, 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 $\widehat{p}_{F^{(T_{k-1})}}$, we construct the bootstrapped LEGO distribution $\mathcal{D}^{L_k,L_k}_{F^{(T_{k-1})}}$ with total length $L_k = 2^k$. The corresponding self-training loss $\mathsf{Loss}^{L_k,L'}_{F^{(T_{k-1})}}$ is then defined as in Definition H.3. In this stage, we focus on training via gradient descent on $\mathsf{Loss}^{L_k,L'}_{F^{(T_{k-1})},i}$ 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 $\epsilon^{L_k,2}_{\mathsf{attn}}$ as $\epsilon^{L_k}_{\mathsf{attn}}$ and the attention gap $\Delta^{L_k,2}$ as Δ^{L_k} when the context is clear.

H.1 Preliminaries

We first present preliminaries on the recursive learning attention layer, including its gradient computations and some useful probability lemmas.

H.1.1 Gradient Computations

Fact H.1 (Gradients of $\mathbf{Q}$)**.** Given $F^{(T_{k-1})}$ from the previous stage $\mathcal{T}^{L_{k-1}}$ wiht $L_{k-1} = 2^{k-1}$ for $k \geq 2$, for $(p, q) \in \Big\{(4,3),(4,4)\Big\}$, we have

$$-\nabla_{\mathbf{Q}_{p,q}} \mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5} = -\nabla_{\mathbf{Q}_{p,q}} \mathsf{Loss}^{L_k,2}_{5}.$$

Remark H.1. *This fact is straightforward to verify. Since the attention error $\epsilon^{L_{k-1},2}_{\mathsf{attn}}$ 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*

$$\epsilon^{L_k}_{\mathsf{attn}} \;\leq\; 2\epsilon^{L_{k-1}}_{\mathsf{attn}},$$

which remains small. Moreover, the attention gap $\Delta^{L_k} = |\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}|$ 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 $Z_{\mathsf{ans},2,5}$ given $Z^{L_k,1}$ remains significantly larger than that of any incorrect prediction. Consequently, under the greedy data annotator, the self-training loss $\mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5}$ coincides with the original loss $\mathsf{Loss}^{L_k,2}_{5}$ on the ground-truth LEGO distribution.

Lemma H.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*

$$\Big[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5}\Big]_{s,s} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot \Big(\big\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},2}\big\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Bigg)\mathbb{1}_{s=\tau(x_1)}\Bigg].$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\Big[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5}\Big]_{s,s'} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot \Big(\big\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{pred},1}\big\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Bigg)\mathbb{1}_{s=\tau(x_1),s'=\tau(x_0)}\Bigg].$$

Lemma H.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*

$$\Big[-\nabla_{\mathbf{Q}_{4,4}} \mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5}\Big]_{s,s} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot \Big(\big\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{ans},1}\big\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Bigg)\mathbb{1}_{s=\tau(x_1)}\Bigg].$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,4}]_{s,s'}$ with $s \neq s'$, we have

$$\Big[-\nabla_{\mathbf{Q}_{4,4}} \mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5}\Big]_{s,s'} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0} \cdot \Bigg(\sum_{j\in[d]} \mathcal{E}_{5,j}(\mathbf{Z}^{2,1}) \sum_{r\in[m]} \mathbf{sReLU}'\big(\Lambda_{5,j,r}\big)\cdot \Big(\big\langle \mathbf{W}_{5,j,r}, \mathbf{Z}_{\mathsf{ans},0}\big\rangle - \Lambda_{5,j,r} + b_{5,j,r}\Big)\Bigg)\mathbb{1}_{s=\tau(x_1),s'=\tau(x_0)}\Bigg].$$

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

$$\widehat{\mathfrak{A}}_j(\mathbf{Z}^{L,\ell-1}) = \Big\{ r_{g\cdot y} \;\Big|\; g \in \mathrm{Fiber}_{j,y},\ g \in \widehat{\mathcal{G}}(\mathbf{Z}^{L,\ell-1}) \ \vee\ y \in \widehat{\mathcal{Y}} \Big\}.$$

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 G.2. With these preliminaries in place, we now state the following lemma, which also relies on the feature-magnitude bounds established therein.

Lemma H.3 (Characterizations of Lambda)**.** *Given $\mathbf{Z}^{L_k,1}$ with $\{\mathbf{Attn}_{\mathsf{ans},1\to\mathbf{k}}\}_{\mathbf{k}\in\mathcal{I}^{L_k,1}}$, then*

(a) for $j \in \tau(\mathcal{Y})$, for activated neuron $r \in \mathfrak{A}_j$, we have

$$\Lambda_{5,j,r} = \sum_{\ell'=1}^{2} \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},\ell'} \psi_{j,r}(g_{\ell'}) + \sum_{\ell'=1}^{L_k} \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},\ell'-1} \psi_{j,r}(y_{\ell'-1}) \pm \widetilde{O}(\sigma_0).$$

(b) for $j \in \tau(\mathcal{Y})$, for any non-activated neuron $r \notin \mathfrak{A}_j$ we have

$$\Big|\Lambda_{5,j,r}\Big| \le O(\delta).$$

(c) for $j \notin \tau(\mathcal{Y})$, for any $r \in [m]$, we have

$$\Big|\Lambda_{5,j,r}\Big| \le \widetilde{O}(\sigma_0).$$

Lemma H.4. *Given $j \in \tau(\mathcal{Y})$, for $r \in \mathfrak{A}_j \setminus \widehat{\mathfrak{A}}_j$, we have $\mathbf{sReLU}'\big(\Lambda_{5,j,r}\big) = 0$.*

We are now ready to derive the gradients of the attention layer, building on Lemmas H.1 and H.2 and the properties established above.

Lemma H.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*

$$\begin{aligned}
\Big[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5} \Big]_{s,s} = \mathbb{E}\Bigg[& \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2} \cdot \\
& \Big((1 - \mathbf{logit}_{5,j_2}) \cdot \Big(\sum_{r \in \widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \Big(\psi_{j_2,r}(g_2) - \Lambda_{5,j_2,r} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \\
& - \sum_{j \neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \Big(\sum_{r \in \widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \Big(\psi_{j,r}(g_2) - \Lambda_{5,j,r} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \\
& \pm \sum_{j \notin \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \widetilde{O}(\sigma_0^q) \Big) \mathbb{1}_{\tau(x_1)=s} \Bigg]
\end{aligned}$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\Big[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5} \Big]_{s,s'} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},1} \cdot \Big((1-\mathbf{logit}_{5,j_2}) \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \Big(\psi_{j_2,r}(g_1) - \Lambda_{5,j_2,r} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) - \sum_{j\neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \Big(\psi_{j,r}(g_1) - \Lambda_{5,j,r} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \pm \sum_{j\notin\tau(\mathcal{Y})} \mathbf{logit}_{5,j} \widetilde{O}(\sigma_0^q) \Big) \mathbb{1}_{\tau(x_1)=s,\tau(x_0)=s'} \Bigg]$$

Lemma H.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*

$$\Big[-\nabla_{\mathbf{Q}_{4,4}} \mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5} \Big]_{s,s} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1} \cdot \Big((1-\mathbf{logit}_{5,j_2}) \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \Big(\psi_{j_2,r}(y_1) - \Lambda_{5,j_2,r} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) - \sum_{j\neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \Big(\psi_{j,r}(y_1) - \Lambda_{5,j,r} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \pm \sum_{j\notin\tau(\mathcal{Y})} \mathbf{logit}_{5,j} \widetilde{O}(\sigma_0^q) \Big) \mathbb{1}_{\tau(x_1)=s} \Bigg].$$

Moreover, for the off-diagonal entries $[\mathbf{Q}_{4,3}]_{s,s'}$ with $s \neq s'$, we have

$$\Big[-\nabla_{\mathbf{Q}_{4,4}} \mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5} \Big]_{s,s'} = \mathbb{E}\Bigg[\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0} \cdot \Big((1-\mathbf{logit}_{5,j_2}) \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \Big(\psi_{j_2,r}(y_0) - \Lambda_{5,j_2,r} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) - \sum_{j\neq j_2 \in \tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot \Big(\psi_{j,r}(y_0) - \Lambda_{5,j,r} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big) \pm \sum_{j\notin\tau(\mathcal{Y})} \mathbf{logit}_{5,j} \widetilde{O}(\sigma_0^q) \Big) \mathbb{1}_{\tau(x_1)=s,\tau(x_0)=s'} \Bigg].$$

Following the above calculations, we can further obtain the gradient summation of $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ as follows:

Lemma H.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*

$$\begin{aligned}
&\Big[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5} \Big]_{s,s} + \Big[-\nabla_{\mathbf{Q}_{4,3}} \mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5} \Big]_{s,s} \\
&= \mathbb{E}\Bigg[\Big((1-\mathbf{logit}_{5,j_2}) \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}} \mathbf{sReLU}'(\Lambda_{5,j_2,r}) \cdot \\
&\qquad \Big(-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0} \cdot \psi_{j_2,r}(y_0) - \sum_{\ell\neq 2} \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},\ell} \cdot \psi_{j_2,r}(g_\ell) \\
&\qquad\quad + \big(1 - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\big) \Lambda_{5,j_2,r} \pm \widetilde{O}(\sigma_0) \Big) \pm \widetilde{O}(\delta^q) \Big)
\end{aligned}$$

$$
\begin{aligned}
&+ \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}_{5,j}\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\\
&\Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}\cdot\psi_{j,r}(y_0)+\sum_{\ell\neq 2}\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},\ell}\cdot\psi_{j_2,r}(g_\ell)\\
&\quad-\big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\\
&\pm\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot\big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}+\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\widetilde{O}(\sigma_0^q)\Big)\mathbb{1}_{\tau(x_1)=s}\Big].
\end{aligned}
$$

Notations for gradient decompositions. Next we define some useful notations to further simplify the expressions of gradient.

Lemma H.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 $\big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5}\big]_{s,s}=\mathbb{E}\big[\mathcal{N}_{s,3,L_k,i}+\mathcal{N}_{s,3,L_k,ii}+\mathcal{N}_{s,3,L_k,iii}\big]$, where*

$$
\begin{aligned}
\mathcal{N}_{s,3,L_k,i}&=\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot(1-\mathbf{logit}_{5,j_2})\cdot \qquad (114)\\
&\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot\Big(\psi_{j_2,r}(g_2)-\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s},\\
\mathcal{N}_{s,3,L_k,ii}&=-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot \qquad (115)\\
&\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(g_2)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s},\\
\mathcal{N}_{s,3,L_k,iii}&=\pm\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\mathbb{1}_{\tau(x_1)=s}. \qquad (116)
\end{aligned}
$$

2. *for $[\mathbf{Q}_{4,4}]_{s,s}$, we have $\big[-\nabla_{\mathbf{Q}_{4,4}}\mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5}\big]_{s,s}=\mathbb{E}\big[\mathcal{N}_{s,4,L_k,i}+\mathcal{N}_{s,4,L_k,ii}+\mathcal{N}_{s,4,L_k,iii}\big]$, where*

$$
\begin{aligned}
\mathcal{N}_{s,4,L_k,i}&=\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot(1-\mathbf{logit}_{5,j_2})\cdot \qquad (117)\\
&\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot\Big(\psi_{j_2,r}(y_1)-\Lambda_{5,j_2,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s},\\
\mathcal{N}_{s,4,L_k,ii}&=-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot\sum_{j\neq j_2\in\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\cdot \qquad (118)\\
&\Big(\sum_{r\in\widehat{\mathfrak{A}}_j}\mathbf{sReLU}'(\Lambda_{5,j,r})\cdot\Big(\psi_{j,r}(y_1)-\Lambda_{5,j,r}\pm\widetilde{O}(\sigma_0)\Big)\pm\widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s},\\
\mathcal{N}_{s,4,L_k,iii}&=\pm\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}\cdot\sum_{j\notin\tau(\mathcal{Y})}\mathbf{logit}_{5,j}\widetilde{O}(\sigma_0^q)\mathbb{1}_{\tau(x_1)=s}. \qquad (119)
\end{aligned}
$$

3. *for the summation of $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$, we have $\big[-\nabla_{\mathbf{Q}_{4,3}}\mathsf{Loss}^{2,2}_5\big]_{s,s}+\big[-\nabla_{\mathbf{Q}_{4,4}}\mathsf{Loss}^{2,2}_5\big]_{s,s}=\mathbb{E}\big[\mathcal{N}_{s,2,i}+\mathcal{N}_{s,2,ii}+\mathcal{N}_{s,2,iii}\big]$, where*

$$
\mathcal{N}_{s,L_k,i}=(1-\mathbf{logit}_{5,j_2})\cdot\Big(\sum_{r\in\widehat{\mathfrak{A}}_{j_2}}\mathbf{sReLU}'(\Lambda_{5,j_2,r})\cdot
$$

$$\Big(- \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0} \cdot \psi_{j_2,r}(y_0) - \sum_{\ell\neq 2} \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},\ell} \cdot \psi_{j_2,r}(g_\ell)$$

$$+ \big(1 - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda_{5,j_2,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}$$

$$\mathcal{N}_{s,L_k,ii} = \sum_{j\neq j_2\in\tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \Big(\sum_{r\in\widehat{\mathfrak{A}}_j} \mathbf{sReLU}'(\Lambda_{5,j,r}) \cdot$$

$$\Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0} \cdot \psi_{j,r}(y_0) + \sum_{\ell\neq 2} \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},\ell} \cdot \psi_{j,r}(g_\ell)$$

$$- \big(1 - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda_{5,j,r} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big)\mathbb{1}_{\tau(x_1)=s}$$

$$\mathcal{N}_{s,L_k,iii} = \pm \sum_{j\notin\tau(\mathcal{Y})} \mathbf{logit}_{5,j} \cdot \big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1} + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\widetilde{O}(\sigma_0^q)\mathbb{1}_{\tau(x_1)=s}.$$

H.1.2 Probabilistic Events

We introduce the following events:

$$\mathcal{E}_{L_k,1} = \Big\{ \frac{1}{L_k} \sum_{\ell\in[L_k]\setminus\{2\}} \mathbb{1}\big\{g_\ell(y_1) = g_2(y_1)\big\} \;\le\; U_k \Big\},$$

$$\mathcal{E}_{L_k,2} = \big\{y_0 = y_1\big\},$$

$$\mathcal{E}_{L_k,3} = \Big\{ \max_{y\in\mathcal{Y}} \sum_{\ell\in[L_k]\setminus\{2\}} \mathbb{1}\big\{g_\ell(y) = g_2(y)\big\} \;\le\; W_k \Big\}.$$

We set

$$U_k \;=\; \frac{1}{L_k}\Big\lceil 1 + \frac{L_k}{8}\Big\rceil \;=\; \Theta(1), \qquad W_k \;=\; \begin{cases} \Theta\Big(\dfrac{\log n_y}{\log\big(\frac{4n_y}{L_k}\,\log n_y\big)}\Big), & L_k \le n_y \log n_y, \\ \Theta\big(\dfrac{L_k}{n_y}\big), & L_k \ge n_y \log n_y. \end{cases}$$

By standard balls-into-bins tail bounds and maximum-load estimates (e.g., [88]), we obtain

$$\mathbf{Pr}\big(\mathbf{Z}^{L_k,1} \notin \mathcal{E}_{L_k,1}\big) \;=\; \begin{cases} O\big(n_y^{-\,U_k L_k + 1}\big), & L_k \ll n_y, \\ \exp\big(-\Theta(n_y \log n_y)\big), & L_k = \Theta(n_y), \\ \exp\big(-\Theta(L_k \log n_y)\big), & L_k \gg n_y, \end{cases}$$

and

$$\mathbf{Pr}\big(\mathbf{Z}^{L_k,1} \in \mathcal{E}_{L_k,3}\big) \;=\; 1 - n_y^{-\Omega(1)}.$$

H.2 Reducing the Wrong Attention

As discussed in Remark H.1, both $\epsilon_{\mathsf{attn}}^{L_k}$ 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 $\mathsf{Loss}_{5,s}^{L_k,2}$ in what follows.

Induction H.1. *Given $F^{(T_{k-1})}$ from the previous stage $\mathcal{T}^{L_{k-1}}$ wiht $L_{k-1} = 2^{k-1}$ for $k \ge 2$, let T_k denote the first time that $\mathsf{Loss}_{5,s}^{L_k,2}$ decreases below $e^{(-\frac{1}{2}+3.01c_1)\cdot 2B}$. For all iterations $t \le T_k$, we have the following holds*

1. $\big[\mathbf{Q}_{4,3}^{(t)}\big]_{s,s} + \big[\mathbf{Q}_{4,4}^{(t)}\big]_{s,s}$ *monotonically increases;*
2. *for any sample $\mathbf{Z}^{2,1}$, we have $\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} \le c_1$ for some sufficiently small constant $c_1 = \frac{1.005\log d}{2B} > 0$;*

3. *for any* $\mathbf{Z}^{2,1}$, *we have* $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \le \min\Big\{\frac{L_k-1}{L_k}\epsilon^{L_k,2}_{\mathsf{attn}} - c_2, 0\Big\}$, *where* $c_2 = \frac{\log d}{4B} > 0$ *is some sufficiently small constant.*

4. *for* $p \in \{3,4\}$, *for* $s' \in \tau(\mathcal{X}) \neq s$, $|\big[\mathbf{Q}^{(t)}_{4,p}\big]_{s,s'}| \le O\Big(\frac{[\mathbf{Q}^{(t)}_{4,p}]_{s,s}}{d}\Big)$ *; otherwise* $[\mathbf{Q}^{(t)}_{4,p}]_{s,s'} = 0$.

H.2.1 Attention and Lambda Preliminaries

Lemma H.9. *If Induction H.1 holds for all iterations* $[T_{k-1}, t)$,*then for any sample* $\mathbf{Z}^{L_k,1}$, *we have*

1. $\epsilon^{L_k}_{\mathsf{attn}} \in \Big[\frac{4}{3}c_1, 2\epsilon^{L_{k-1}}_{\mathsf{attn}}\Big]$;
2. $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathbf{k}} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathbf{k}'} \ge \Omega(1)$ *for any* $\mathbf{k},\mathbf{k}' \in \mathcal{I}^{L_k,1}\setminus\{(\mathsf{ans},1),(\mathsf{pred},2)\}$;
3. $\big|\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathbf{k}} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathbf{k}'}\big| \le \widetilde{O}\big(\frac{1}{d}\big)$ *for any* $\mathbf{k},\mathbf{k}' \in \mathcal{I}^{L_k,1}\setminus\{(\mathsf{ans},1)\}$;

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 $y_0 \neq y_1$.

$$
\begin{aligned}
&\log p_F(\mathbf{Z}_{\mathsf{ans},2,5}|\mathbf{Z}^{(L_k,1)}) \\
&\le \Theta(1)\cdot\max\Bigg\{e^{\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}+\frac{L_k-1}{L_k}\epsilon^{L_k,2}_{\mathsf{attn}}-\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)2B-\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}-\frac{1}{L_k}\epsilon^{L_k,2}_{\mathsf{attn}}\Big)2B}, \\
&\qquad e^{\log d-\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}-\frac{1}{L_k}\epsilon^{L_k,2}_{\mathsf{attn}}\Big)2B}\Bigg\} \\
&\le \Theta\Bigg(e^{(\frac{L_k-1}{L_k}\epsilon^{L_k,2}_{\mathsf{attn}}+c_1)2B-\Big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}-\frac{1}{L_k}\epsilon^{L_k,2}_{\mathsf{attn}}\Big)2B}\Bigg) \\
&\le \Theta\Bigg(e^{-\Big(\frac{1}{2}-\frac{3}{2}\epsilon^{L_k,2}_{\mathsf{attn}}-c_1\Big)2B}\Bigg).
\end{aligned}
$$

Thus if $\epsilon^{L_k}_{\mathsf{attn}} \le \frac{4}{3}c_1$, we must have

$$
\mathsf{Loss}^{2,2}_{5,s} \;\le\; \Theta\Big(e^{(-\frac{1}{2}+3\cdot\frac{2c_1}{3}+c_1)2B}\Big) = \Theta\Big(e^{-(\frac{1}{2}+3c_1)2B}\Big),
$$

which contradicts the definition of $T_{1,2,2,s}$. Therefore, it must hold that $\epsilon^{L_k}_{\mathsf{attn}} \ge \frac{4}{3}c_1$ for all $t \le T_k$.

□

Lemma H.10. *If Induction H.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*

$$
\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} = \Big(\sum_{\ell\in[L_k]}\mathbb{1}_{g_\ell(y_1)=g_2(y_1)}\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},\ell} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\Big)\cdot 2B + O\Big(\frac{B}{n_y}\Big) + O(\delta).
$$

2. *for the prediction* $j_2' = \tau\big(g_2(y_0)\big)$, *we have*

$$
\Lambda^{(t)}_{5,j_2',r_{g_2\cdot y_0}} = \Big(\sum_{\ell\in[L_k]}\mathbb{1}_{g_\ell(y_0)=g_2(y_0)}\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},\ell} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}\Big)\cdot 2B + O\Big(\frac{B}{n_y}\Big) + O(\delta).
$$

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*

$$
\begin{aligned}
\Lambda^{(t)}_{5,j,r_{g_2\cdot y}} = &\Big(\sum_{\ell\in[L_k]}\mathbb{1}_{g_\ell(y)=g_2(y)}\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},\ell} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\Big)\cdot 2B \\
&+ O\Big(\frac{B}{n_y}\Big) + O(\delta).
\end{aligned}
$$

Lemma H.11. *If Induction H.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*

$$\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)} = \sum_{\ell\in[L_k]} \mathbb{1}_{g_\ell(y_1)=g_2(y_1)} \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},\ell}^{(t)} \cdot 2B + O\Big(\frac{B}{n_y}\Big) + O(\delta).$$

2. *for other $j = \tau(g_2(y)) \in \tau(\mathcal{Y})$ with $y \neq y_1$, then we have*

$$\Lambda_{5,j,r_{g_2\cdot y}}^{(t)} = \Big(\sum_{\ell\in[L_k]} \mathbb{1}_{g_\ell(y)=g_2(y)} \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},\ell}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\Big)\cdot 2B + O\Big(\frac{B}{n_y}\Big) + O(\delta).$$

H.2.2 Gradient Lemma

Lemma H.12. *If Induction H.1 holds for all iterations $t \in [T_{k-1}, T_k)$, given $s \in \tau(\mathcal{X})$, we have*

$$\Big[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \mathsf{Loss}_5^{2,2}\Big]_{s,s} + \Big[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \mathsf{Loss}_5^{2,2}\Big]_{s,s} \geq \Omega(B/d)\mathbb{E}\Big[\epsilon_{\mathsf{attn}}^{L_k}\cdot(1-\mathbf{logit}_{5,j_2}^{(t)})\big|\tau(x_1)=s, \mathcal{E}_{L_k,1}\cap\mathcal{E}_{L_k,2}^c\Big].$$

Proof. By Induction H.1, Lemma H.10, and Lemma H.11, we have

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}_{s,L_k,i}^{(t)}\Big] \\
&= \mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\cdot\Big(-\sum_{\ell\neq 2}\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},\ell}^{(t)}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(g_\ell) - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(y_0) \\
&\quad + \big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big] \qquad (120) \\
&\quad \mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)}\leq B}.
\end{aligned}$$

Similarly,

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}_{s,L_k,ii}^{(t)}\Big] \\
&= \mathbb{E}\Big[\sum_{y\neq y_1}\mathbf{logit}_{5,\tau(g_2(y))}^{(t)}\cdot\mathbf{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2\cdot y}}^{(t)}) \\
&\qquad \Big(\sum_{\ell\neq 2}\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},\ell}^{(t)}\cdot\psi_{\tau(g_2(y)),r_{g_2\cdot y}}(g_\ell) + \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{\tau(g_2(y)),r_{g_2\cdot y}}(y_0) \\
&\qquad\quad - \big(1-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}\big)\Lambda_{5,\tau(g_2(y)),r_{g_2\cdot y}}^{(t)} \pm \widetilde{O}(\sigma_0)\Big) \pm \widetilde{O}(\delta^q)\Big]\mathbb{1}_{\tau(x_1)=s}.
\end{aligned}$$

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

$$\begin{aligned}
\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)} &= \sum_{\ell\neq 2}\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},\ell}^{(t)}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(g_\ell)\mathbb{1}_{g_\ell(y_1)=j_2} - \mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},0}^{(t)}\cdot\psi_{j_2,r_{g_2\cdot y_1}}(y_0) \\
&\leq 2\Big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)} + (U_k - 1/L_k)\cdot\epsilon_{\mathsf{attn}}^{L_k}\Big)B \\
&\leq 2\Big(\tfrac{1-\epsilon_{\mathsf{attn}}^{L_k}+c_1}{2} + (U_k-1/L_k)\cdot\epsilon_{\mathsf{attn}}^{L_k}\Big)B \\
&= 2\Big(\tfrac{1+c_1}{2} - (\tfrac{1}{2}+1/L_k-U_k)\cdot\epsilon_{\mathsf{attn}}^{L_k}\Big)B.
\end{aligned}$$

Thus, provided $\epsilon_{\mathsf{attn}}^{L_k} \geq 4c_1/3$ and choosing $U_k = \big\lfloor 1 + \frac{L_k}{8}\big\rfloor/L_k$, we indeed have $\Lambda_{5,j_2,r_{g_2\cdot y_1}}^{(t)} \leq B$.

- For $j = \tau(g_2(y))$ in $\mathcal{N}^{(t)}_{s,L_k,ii}$:
  - $y = y_0 \neq y_1$,

$$\sum_{\ell\neq 2} \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},\ell} \cdot \psi_{\tau(g_2(y)),r_{g_2\cdot y}}(g_\ell) + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} \cdot \psi_{\tau(g_2(y)),r_{g_2\cdot y}}(y_0)$$
$$\geq \sum_{\ell\neq 2} \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},\ell} \cdot \psi_{\tau(g_2(y)),r_{g_2\cdot y}}(g_\ell) \mathbb{1}_{g_\ell(y)\neq g_2(y)}$$
$$\geq -O\Big(\tfrac{B}{n_y}\Big) \cdot \big(1 - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big). \tag{121}$$

  - $y \neq y_0, y_1$. If there exists at least one $\ell \neq 2$ such that $g_\ell(y) = g_2(y)$, then by cancellation,

$$\sum_{\ell\neq 2} \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},\ell} \cdot \psi_{\tau(g_2(y)),r_{g_2\cdot y}}(g_\ell) + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} \cdot \psi_{\tau(g_2(y)),r_{g_2\cdot y}}(y_0)$$
$$\geq -O\Big(\tfrac{B}{n_y}\Big) \cdot \big(1 - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big). \tag{122}$$

    Else, pick $y' \neq y, y_1$ so that $\Lambda^{(t)}_{5,\tau(g_2(y')),r_{g_2\cdot y'}} \geq \Lambda^{(t)}_{5,\tau(g_2(y)),r_{g_2\cdot y}}$. Then

$$\mathbf{logit}_{5,\tau(g_2(y))} \leq O\Big(\tfrac{1}{n_y}\Big)\big(1 - \mathbf{logit}_{5,j_2}\big),$$

    which implies

$$\mathbf{logit}^{(t)}_{5,j}\Big(\sum_{\ell\neq 2} \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},\ell} \cdot \psi_{\tau(g_2(y)),r_{g_2\cdot y}}(g_\ell) + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} \cdot \psi_{\tau(g_2(y)),r_{g_2\cdot y}}(y_0)\Big)$$
$$\geq -O\Big(\tfrac{B}{L_k n_y}\Big) \cdot \big(1 - \mathbf{logit}_{5,j_2}\big) \cdot \big(1 - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big). \tag{123}$$

- For $\mathcal{N}^{(t)}_{s,L_k,i}$:
  - If $\mathbf{Z}^{L_k,1} \in \mathcal{E}_{L_k,1} \cap \mathcal{E}^c_{L_k,2}$, then

$$-\sum_{\ell\neq 2} \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},\ell} \cdot \psi_{j_2,r_{g_2\cdot y_1}}(g_\ell) - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} \cdot \psi_{j_2,r_{g_2\cdot y_1}}(y_0)$$
$$+ \big(1 - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}}$$
$$\geq -\sum_{\ell\neq 2} \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},\ell} \mathbb{1}_{g_\ell = j_2} \cdot \psi_{j_2,r_{g_2\cdot y_1}}(g_\ell) - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} \cdot \psi_{j_2,r_{g_2\cdot y_1}}(y_0)$$
$$+ \big(1 - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}}$$
$$\geq \Big(-(U_k - \tfrac{1}{L_k}) \cdot 2B + \Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}}\Big) \cdot \epsilon^{L_k}_{\mathsf{attn}} \quad \geq \quad \big(\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}} - \tfrac{B}{4}\big) \cdot \epsilon^{L_k}_{\mathsf{attn}}. \tag{124}$$

  - Else, for $\mathbf{Z}^{L_k,1} \in \mathcal{E}^c_{L_k,1} \cup \mathcal{E}_{L_k,2}$, note that having more identical predictions $g_\ell(y_1) = j_2$ or less distraction (i.e., $y_0 = y_1$) increases $\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}}$. Hence

$$\mathbb{E}\big[(1 - \mathbf{logit}_{5,j_2}) \mid \mathcal{E}^c_{L_k,1} \cup \mathcal{E}_{L_k,2}\big] \;\leq\; O(1) \cdot \mathbb{E}\big[(1 - \mathbf{logit}_{5,j_2}) \mid \mathcal{E}_{L_k,1} \cap \mathcal{E}^c_{L_k,2}\big],$$

    and

$$\Big| -\sum_{\ell\neq 2} \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},\ell} \cdot \psi_{j_2,r_{g_2\cdot y_1}}(g_\ell) - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0} \cdot \psi_{j_2,r_{g_2\cdot y_1}}(y_0)$$
$$+ \big(1 - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2}\big)\Lambda^{(t)}_{5,j_2,r_{g_2\cdot y_1}}\Big| \;\leq\; O(1) \cdot \epsilon^{L_k}_{\mathsf{attn}} \cdot B. \tag{125}$$

Putting the above together, we have

$$\begin{aligned}
&\Big[-\nabla_{\mathbf{Q}_{4,3}^{(t)}}\mathsf{Loss}_5^{2,2}\Big]_{s,s}+\Big[-\nabla_{\mathbf{Q}_{4,4}^{(t)}}\mathsf{Loss}_5^{2,2}\Big]_{s,s}\\
&=\mathbb{E}\Big[\mathcal{N}_{s,L_k,1}^{(t)}\mathbb{1}_{\mathcal{E}_{L_k,1}\cap\mathcal{E}_{L_k,2}^c}\Big]+\mathbb{E}\Big[\mathcal{N}_{s,L_k,1}^{(t)}\mathbb{1}_{\Lambda_{5,j_2,r_{g_2}\cdot y_1}^{(t)}\le B}\mathbb{1}_{\mathcal{E}_{L_k,1}^c\cup\mathcal{E}_{L_k,2}}\Big]\\
&\quad+\mathbb{E}\Big[\mathcal{N}_{s,L_k,2}^{(t)}\Big]\pm\widetilde{O}(\sigma_0^q)\\
&=\mathbb{E}\Big[(\mathcal{N}_{s,L_k,1}^{(t)}+\mathcal{N}_{s,L_k,2}^{(t)})\mathbb{1}_{\mathcal{E}_{L_k,1}\cap\mathcal{E}_{L_k,2}^c}\Big]\\
&\quad+\mathbb{E}\Big[\mathcal{N}_{s,L_k,1}^{(t)}\mathbb{1}_{\Lambda_{5,j_2,r_{g_2}\cdot y_1}^{(t)}\le B}\mathbb{1}_{\mathcal{E}_{L_k,1}^c\cup\mathcal{E}_{L_k,2}}\Big]+\mathbb{E}\Big[\mathcal{N}_{s,L_k,2}^{(t)}\mathbb{1}_{\mathcal{E}_{L_k,1}^c\cup\mathcal{E}_{L_k,2}}\Big]\pm\widetilde{O}(\sigma_0^q).
\end{aligned}$$

First, by (121), (122), (124),

$$\begin{aligned}
&\mathbb{E}\Big[(\mathcal{N}_{s,L_k,1}^{(t)}+\mathcal{N}_{s,L_k,2}^{(t)})\mathbb{1}_{\mathcal{E}_{L_k,1}\cap\mathcal{E}_{L_k,2}^c}\Big]\\
&\ge\mathbb{E}\Big[(1-\mathbf{logit}_{5,j_2}^{(t)})\Big((\Lambda_{5,j_2,r_{g_2}\cdot y_1}^{(t)}-\tfrac{B}{4})\epsilon_{\mathsf{attn}}^{L_k}\Big)-(1-\mathbf{logit}_{5,j_2}^{(t)})\Big(O(\tfrac{B}{n_y})\epsilon_{\mathsf{attn}}^{L_k}+\max_{y\neq y_1}\Lambda_{5,\tau(g_2(y)),r_{g_2}\cdot y}^{(t)}\Big)\\
&\qquad\qquad\cdot\mathbb{1}_{\tau(x_1)=s}\mathbb{1}_{\mathcal{E}_{L_k,1}\cap\mathcal{E}_{L_k,2}^c}\Big]\\
&\ge\Omega(B/d)\cdot\mathbb{E}\Big[\epsilon_{\mathsf{attn}}^{L_k}\cdot(1-\mathbf{logit}_{5,j_2}^{(t)})\mid\tau(x_1)=s,\mathcal{E}_{L_k,1}\cap\mathcal{E}_{L_k,2}^c\Big].
\end{aligned}$$

Moreover, by (125),

$$\begin{aligned}
&\mathbb{E}\Big[\mathcal{N}_{s,L_k,1}^{(t)}\mathbb{1}_{\Lambda_{5,j_2,r_{g_2}\cdot y_1}^{(t)}\le B/2}\mathbb{1}_{\mathcal{E}_{L_k,1}^c\cup\mathcal{E}_{L_k,2}}\Big]\\
&\ge-\mathbf{Pr}(\mathbf{Z}^{L_k,1}\notin\mathcal{E}_{L_k,1}^c)\cdot O(B/d)\cdot\mathbb{E}\Big[\epsilon_{\mathsf{attn}}^{L_k}\cdot(1-\mathbf{logit}_{5,j_2}^{(t)})\mid\tau(x_1)=s,\mathcal{E}_{L_k,1}^c\cup\mathcal{E}_{L_k,2}\Big]\\
&\ge-\Big(O(1/n^{\Omega(1)})+1/n_y\Big)\cdot O(B/d)\cdot\mathbb{E}\Big[\epsilon_{\mathsf{attn}}^{L_k}\cdot(1-\mathbf{logit}_{5,j_2}^{(t)})\mid\tau(x_1)=s,\mathcal{E}_{L_k,1}\cap\mathcal{E}_{L_k,2}^c\Big],
\end{aligned}$$

and, by (121), (122), (123),

$$\mathbb{E}\Big[\mathcal{N}_{s,L_k,2}^{(t)}\mathbb{1}_{\mathcal{E}_{L_k,1}^c\cup\mathcal{E}_{L_k,2}}\Big]\ \ge\ \tfrac{1}{n_y}\cdot O(B/d)\cdot\mathbb{E}\Big[\epsilon_{\mathsf{attn}}^{L_k}\cdot(1-\mathbf{logit}_{5,j_2}^{(t)})\mid\tau(x_1)=s,\mathcal{E}_{L_k,1}^c\cup\mathcal{E}_{L_k,2}\Big].$$

Therefore,

$$\begin{aligned}
&\Big[-\nabla_{\mathbf{Q}_{4,3}^{(t)}}\mathsf{Loss}_5^{2,2}\Big]_{s,s}+\Big[-\nabla_{\mathbf{Q}_{4,4}^{(t)}}\mathsf{Loss}_5^{2,2}\Big]_{s,s}\\
&\ge\Omega(B/d)\cdot\mathbb{E}\Big[\epsilon_{\mathsf{attn}}^{L_k}\cdot(1-\mathbf{logit}_{5,j_2}^{(t)})\mid\tau(x_1)=s,\mathcal{E}_{L_k,1}\cap\mathcal{E}_{L_k,2}^c\Big].
\end{aligned}$$

□

Lemma H.13. *If Induction H.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*

$$\Big|\Big[-\nabla_{\mathbf{Q}_{4,p}^{(t)}}\mathsf{Loss}_5^{2,2}\Big]_{s,s'}\Big|\le O(\frac{1}{d})\Big|\Big[-\nabla_{\mathbf{Q}_{4,p}^{(t)}}\mathsf{Loss}_5^{2,2}\Big]_{s,s}\Big|.$$

H.2.3 Attention gap is small

Lemma H.14. *If Induction H.1 holds for all iterations $t\in[T_{k-1},T_k)$, then for any sample $\mathbf{Z}^{L_k,1}$, we have*

$$\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}\le c_1,$$

where $c_1>0$ is a small constant.

Proof. The proof follows the same high-level idea as Lemma G.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

$$\Lambda_{5,j,r_{g_2}\cdot y}^{(t)}\ \ge\ 2\big(\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{pred},2}^{(t)}-\mathbf{Attn}_{\mathsf{ans},1\to\mathsf{ans},1}^{(t)}\big)B.$$

Let $\widehat{\mathcal{J}}$ denote the set of indices j achieving $\arg\max_{y\notin\{y_1\}} \Lambda^{(t)}_{5,j,r_{g_2\cdot y}}$, and pick one such index $j' = g_2(y')$. Define $\widetilde{T}$ to be the first time when $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} \ge c_1$. At this time, we have

$$\sum_{j\in\widehat{\mathcal{J}}} \mathbf{logit}^{(\widetilde{T})}_{5,j} \;=\; \Big(1 - O(d^{-\Omega(1)})\Big)\big(1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_2}\big).$$

Plugging this into the gradient expressions yields

$$\begin{aligned}
&\mathcal{N}^{(\widetilde{T})}_{s,3,L_k,i} + \mathcal{N}^{(\widetilde{T})}_{s,3,L_k,ii} \\
&\quad= \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2} \cdot \Bigg(\Big(1 - O(d^{-\Omega(1)})\Big)\big(1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_2}\big) \cdot \Big(\Lambda^{(\widetilde{T})}_{5,j',r_{g_2\cdot y'}} - \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\Big) \\
&\qquad + O(d^{-\Omega(1)})\big(1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_2}\big) \cdot \Big(\psi_{5,j_2,r_{g_2\cdot y_1},2}(g_2) - \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\Big)\Bigg)\mathbb{1}_{\tau(x_1)=s},
\end{aligned}$$

and similarly

$$\begin{aligned}
&\mathcal{N}^{(\widetilde{T})}_{s,4,L_k,i} + \mathcal{N}^{(\widetilde{T})}_{s,4,L_k,ii} \\
&\quad= \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2} \cdot \Bigg(\Big(1 - O(d^{-\Omega(1)})\Big)\big(1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_2}\big) \cdot \Big(-\psi_{j'_2,r_{g_2\cdot y_0}}(y_1) + \Lambda^{(\widetilde{T})}_{5,j',r_{g_2\cdot y'}} - \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\Big) \\
&\qquad + O(d^{-\Omega(1)})\big(1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_2}\big) \cdot \Big(\psi_{5,j_2,r_{g_2\cdot y_1},2}(y_1) - \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\Big)\Bigg)\mathbb{1}_{\tau(x_1)=s}.
\end{aligned}$$

Since $\Lambda^{(\widetilde{T})}_{5,j',r_{g_2\cdot y'}} - \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}} \le 0$ and

$$-\psi_{j_2,r_{g_2\cdot y_1}}(y_1) - \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}} + \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y'}} \;\ge\; \Omega(B),$$

we obtain

$$\Big[-\nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,4}}\mathsf{Loss}^{2,2}_5\Big]_{s,s} + \Big[\nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,3}}\mathsf{Loss}^{2,2}_5\Big]_{s,s} \;\ge\; \Omega\big(\tfrac{B}{d}\big)\cdot\mathbb{E}\big[1 - \mathbf{logit}^{(\widetilde{T})}_{5,j_2} \;\big|\; \tau(x_1) = s\big].$$

Therefore, the quantity $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}$ cannot further increase, and we conclude that

$$\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} \;\le\; c_1.$$

□

Lemma H.15. *If Induction H.1 holds for all iterations $t \in [T_{k-1}, T_k)$, then for any sample $\mathbf{Z}^{L_k,1}$, we have*

$$\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \le \frac{L_k - 1}{L_k}\epsilon^{L_k}_{\mathsf{attn}} - c_2,$$

where $c_2 = \frac{\log d}{2B} > 0$ is a small constant.

Proof. The argument parallels Lemma G.34, but in the recursive setting we first note that for every $\mathbf{Z}^{L_k,1} \in \mathcal{E}^c_{L_k,2}$,

$$\Lambda^{(t)}_{5,\tau(g_2(y_0)),\,r_{g_2\cdot y_0}} \;\le\; \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} + \frac{L_k - 1}{L_k}\,\epsilon^{L_k}_{\mathsf{attn}}.$$

Let $\widetilde{T}$ be the first time such that

$$\mathbb{E}\Big[\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} + \frac{L_k - 1}{L_k}\,\epsilon^{L_k}_{\mathsf{attn}} \;\Big|\; \tau(x_1) = s\Big] \;\le\; \frac{\log d}{4.02\,B}.$$

Then, for any y,

$$\Lambda^{(\widetilde{T})}_{5,\tau(g_2(y)),\,r_{g_2\cdot y}}$$

$$= 2\big(\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2} + \sum_{\ell\neq 2}\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},\ell}\,\mathbb{1}_{g_\ell(y)=g_2(y)} - \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{ans},0}\big)\,B \;\le\; \frac{\log d}{2.01}.$$

Hence the corresponding wrong-class logit mass is small:

$$\mathbf{logit}^{(1)}_{5,j_2} \;\le\; O\big(d^{-1.01/2.01}\big)\,\big(1-\mathbf{logit}^{(t)}_{5,j_2}\big).$$

Plugging this into the gradient decomposition, we obtain

$$\begin{aligned}
&\mathcal{N}^{(\widetilde{T})}_{s,3,L_k,i} + \mathcal{N}^{(\widetilde{T})}_{s,3,L_k,ii} \\
&\quad= \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2}\Big(\big(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2}\big)\big(\psi_{5,j_2,r_{g_2\cdot y_1},2}(g_2) - \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\big) \\
&\qquad\quad - O\big(d^{-1.01/2.01}\big)\,\big(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2}\big)\sum_{y\neq y_0}\big(\psi_{5,j_2,r_{g_2\cdot y},2}(g_2) - \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y}}\big)\Big)\,\mathbb{1}_{\tau(x_1)=s} \\
&\quad\ge \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\Omega(1)\cdot\big(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2}\big)\cdot\big(\psi_{5,j_2,r_{g_2\cdot y_1},2}(g_2) - \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\big)\,\mathbb{1}_{\tau(x_1)=s},
\end{aligned}$$

and

$$\begin{aligned}
&\mathcal{N}^{(\widetilde{T})}_{s,4,L_k,i} + \mathcal{N}^{(\widetilde{T})}_{s,4,L_k,ii} \\
&\quad= \mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2}\Big(\big(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2}\big)\big(\psi_{5,j_2,r_{g_2\cdot y_1},2}(y_1) - \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\big) \\
&\qquad\quad - O\big(d^{-1.01/2.01}\big)\,\big(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2}\big)\sum_{y\neq y_0}\big(\psi_{5,j_2,r_{g_2\cdot y},2}(y_1) - \Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y}}\big)\Big)\,\mathbb{1}_{\tau(x_1)=s} \\
&\quad\le -\mathbf{Attn}^{(\widetilde{T})}_{\mathsf{ans},1\to\mathsf{pred},2}\cdot\Omega(1)\cdot\big(1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2}\big)\cdot\Lambda^{(\widetilde{T})}_{5,j_2,r_{g_2\cdot y_1}}\,\mathbb{1}_{\tau(x_1)=s}.
\end{aligned}$$

Therefore,

$$\big[-\nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,3}}\mathsf{Loss}^{2,2}_5\big]_{s,s} \;\ge\; \Omega\Big(\frac{B}{d}\Big)\cdot\mathbb{E}\big[1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2}\,\big|\,\tau(x_1)=s,\,\mathcal{E}^c_{L_k,2}\big], \tag{126}$$

$$\big[-\nabla_{\mathbf{Q}^{(\widetilde{T})}_{4,4}}\mathsf{Loss}^{2,2}_5\big]_{s,s} \;\le\; -\Omega\Big(\frac{B}{d}\Big)\cdot\mathbb{E}\big[1-\mathbf{logit}^{(\widetilde{T})}_{5,j_2}\,\big|\,\tau(x_1)=s,\,\mathcal{E}^c_{L_k,2}\big]. \tag{127}$$

Finally, recall that

$$\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},1} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1} \;=\; \frac{e^{[\mathbf{Q}^{(t)}_{4,3}]_{s,s}} - e^{[\mathbf{Q}^{(t)}_{4,4}]_{s,s}} - e^{\widetilde{O}(1/d)}}{e^{[\mathbf{Q}^{(t)}_{4,3}]_{s,s}} + e^{[\mathbf{Q}^{(t)}_{4,4}]_{s,s}} + e^{\widetilde{O}(1/d)}}.$$

Combining the two gradient inequalities shows that the quantity

$$\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + \frac{L_k-1}{L_k}\,\epsilon^{L_k}_{\mathsf{attn}} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},1}$$

must strictly decrease at time $\widetilde{T}+1$. □

H.2.4 At the end of stage

Lemma H.16 (At the end of stage)**.** *Induction H.1 holds for all iterations $T_{k-1} < t \le T_k = O(\frac{\mathsf{poly}(d)}{\eta B})$. At the end of stage k, we have*

(a) Attention concentration: $\epsilon^{L_k}_{\mathsf{attn}} \le 5.04c_1$ for some small constant $c_1 = \frac{1.005\log d}{B} > 0$;

(b) Loss convergence: $\mathsf{Loss}^{L_k,2}_{F^{(T_{k-1})},5} \le e^{(-\frac{1}{2}+3.01c_1)2B} = \frac{1}{\mathsf{poly}d}$.

Proof. By Lemmas H.9 and H.12, the diagonal entries $[\mathbf{Q}^{(t)}_{4,3}]_{s,s}$ and $[\mathbf{Q}^{(t)}_{4,4}]_{s,s}$ keep increasing until the total wrong-attention mass satisfies $\epsilon_{\mathsf{attn}} \ge \frac{4c_1}{3}$. Combining Lemmas H.12 to H.15, there exists a stopping time

$$T_k \;=\; O\Big(\frac{\mathsf{poly}(d)}{\eta B}\Big),$$

such that Induction H.1 holds for all $t \in [T_1,\, T_k)$.

To obtain an upper bound on $\epsilon_{\mathsf{attn}}^{L_k}$ 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,

$$\begin{aligned}
&\log p_F\big(Z_{\mathsf{ans},2,5} \mid Z^{(2,1)}\big) \\
&\geq \Theta(1) \cdot \exp\left(\log d - \big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + \sum_{\ell\neq 2} \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},\ell}\, \mathbb{1}_{g_\ell(y_1)=g_2(y_1)} - \mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{ans},0}\big)\, 2B\right) \\
&\geq \Theta(1) \cdot \exp\Big(\log d - \big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + (1/L_k - 1/L_k)\, \epsilon_{\mathsf{attn}}^{L_k}\big)\, 2B\Big) \\
&\geq \Theta\Big(\exp\big(-(\tfrac{1}{2} + \tfrac{c_1}{2} - \tfrac{\epsilon_{\mathsf{attn}}^{L_k}}{2} - \tfrac{\log d}{B})\, 2B\big)\Big),
\end{aligned}$$

where the last line uses $\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} \leq \frac{1}{2}(1 - \epsilon_{\mathsf{attn}}^{L_k} + c_1) = \frac{1+c_1}{2} - \frac{\epsilon_{\mathsf{attn}}^{L_k}}{2}$.

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

$$\begin{aligned}
\log p_F\big(Z_{\mathsf{ans},2,5} \mid Z^{(2,1)}\big) &\geq \Theta(1) \cdot \exp\Big(\log d - \big(\mathbf{Attn}^{(t)}_{\mathsf{ans},1\to\mathsf{pred},2} + (o(1) - 1/L_k)\, \epsilon_{\mathsf{attn}}^{L_k}\big)\, 2B\Big) \\
&\geq \Theta\Big(\exp\big(-(\tfrac{1}{2} + \tfrac{c_1}{2} - \tfrac{\epsilon_{\mathsf{attn}}^{L_k}}{2} - \tfrac{\log d}{B})\, 2B\big)\Big).
\end{aligned}$$

At convergence we therefore have

$$\tfrac{1}{2} + \tfrac{c_1}{2} - \tfrac{\epsilon_{\mathsf{attn}}^{L_k}}{2} - \tfrac{\log d}{B} \;\geq\; \tfrac{1}{2} - 3.01\, c_1,$$

which rearranges to

$$\epsilon_{\mathsf{attn}}^{L_k} \leq 2(3.02c_1 + 0.5c_1 - c_1) = 5.04c_1.$$

In particular, since $B/\log d$ can be taken as a sufficiently large constant, we obtain the clean stage-end bound

$$\epsilon_{\mathsf{attn}}^{L_k} \;=\; O(c_1).$$

□

H.3 Proof of Main Theorem

Theorem H.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 4.2, and assume the transformer network satisfies Assumption 3.3 and A.2. Then for any $2 \leq k \leq \log_2 |\mathcal{X}|$, the transformer $F^{(T_k)}$ trained via Algorithm 2 up to length $L_k = 2^k$ and $T_k = O(\frac{\mathsf{poly}(d)}{\eta})$, is able to solve the task $\mathcal{T}^{L_{k+1}}$, $L_{k+1} = 2^{k+1}$ with accuracy:*

$$\mathrm{Acc}_{L_{k+1}}\big(F^{(T_k)}\big) = 1 - O(1/\mathsf{poly}(d)).$$

Proof. By Lemma H.16, at the time T_k, we have $\epsilon_{\mathsf{attn}}^{L_k} \leq 5.04c_1$, combining with Induction H.1, non-diagonal entry of $\mathbf{Q}_{p,q}$ remains close to 0, thus moving to the next stage, we have $\epsilon_{\mathsf{attn}}^{L_{k+1},\ell} \leq 4\epsilon_{\mathsf{attn}}^{L_k} \leq 20.16c_1$ (notice that $20.16c_1$ could be sufficently 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

$$\begin{aligned}
&\mathbb{E}_{Z^{L_{k+1}} \sim \mathcal{D}^{L_{k+1}}}\left[\mathbb{E}_{\widehat{Z}_{\mathsf{ans},\ell+1} \sim p_F^{(T_k)}(\cdot \mid Z^{L_{k+1},\ell})}\big[\mathbb{1}_{\{\widehat{Z}_{\mathsf{ans},\ell+1} \neq Z_{\mathsf{ans},\ell+1}\}}\big]\right] \\
&\leq O(1) \cdot \mathbb{E}_{Z^{L_{k+1}} \sim \mathcal{D}^{L_{k+1}}}\left[1 - \mathbf{logit}_{5,\tau(\mathbf{Z}_{\mathsf{ans},\ell+1,5})}\right] \\
&\leq O(1) \cdot e^{\left(-(\frac{1}{2} - \epsilon_{\mathsf{attn}}^{L_{k+1}} - (\epsilon_{\mathsf{attn}}^{L_{k+1}} - c_2)/2 - \epsilon_{\mathsf{attn}}^{L_{k+1}})\right) \cdot 2B}
\end{aligned}$$

$$\leq O(1) \cdot e^{(-\frac{1}{2} - \frac{\log d}{4B} + \frac{5}{2} \cdot 0.01)2B} = O\Big(\frac{1}{\mathsf{poly}d}\Big).$$

Thus $\mathrm{Acc}_{L_{k+1}}(F^{T_k}) \geq 1 - O\big(\frac{1}{\mathsf{poly}d}\big)$. □