

Appendix: Complete Proofs

850	A Proof Overview	22
851	A.1 Learning In-Context Retrieval of Variables	22
852	A.2 Learning the Group Operations	23
853	A.3 Learning the Attention Layer	24
854	B Learning In-Context Retrieval of Variables	26
855	B.1 Preliminaries	26
856	B.2 Induction Hypothesis	26
857	B.3 Gradient Lemma	27
858	B.4 Growth of Gamma	28
859	B.5 Group and Value Correlations Are Not Large	29
860	B.6 Off-diagonal Correlations Are Small	30
861	B.7 Non-target Correlations Are Negligible	31
862	B.8 Convergence	31
863	C Learning The Group Actions: Cyclic Group	32
864	C.1 Preliminaries and Induction Hypotheses	32
865	C.2 Technical Lemmas	35
866	C.3 Phase I: Initial Growth	38
867	C.4 Phase II: Cancellation and Convergence	41
868	D Learning The Group Actions: Symmetry Group	47
869	D.1 Induction Hypothesis and Training Phases	48
870	D.2 Phase I: Emergence of the Feature	50
871	D.3 Phase II: Convergence of the Feature	50
872	E Auxiliary Technical Tools	53
873	E.1 Probability	53
874	E.2 Tensor Power Method Bounds	53
875	F Learning the Attention Layer: Cyclic Case	54
876	F.1 Gradient Computations	54
877	F.2 Stage 2.1: Growth of Gap	57
878	F.3 Stage 2.2: Growth of $Q_{4,3}$ and $Q_{4,4}$	61
879	F.4 Stage 2.3: Decrease of Gap and Convergence	65
880	F.5 Proof of Main Theorem	72
881	G Learning the Attention Layer: Symmetry Case for short-length	73
882	G.1 Gradient Computations	73
883	G.2 Stage 1.2.1: Initial Growth of $Q_{4,3}$	75

884	G.3 Stage 1.2.2: Convergence with Small Wrong Attention	76
885	H Recursive Learning the Attention Layer: Symmetric Case	81
886	H.1 Preliminaries	81
887	H.2 Reducing the Wrong Attention	84
888	H.3 Proof of Main Theorem	87

889 A Proof Overview

890 **Warm-up.** We briefly recall the main settings. We study solving LEGO tasks sampled from the
891 distribution in Assumption 3.2, using a one-layer transformer model defined in Definition 3.5. We
892 consider two types of algebraic structures on group actions: (1) *simply transitive group actions*
893 (Assumption 4.1), and (2) *symmetry group actions* (Assumption 4.2). For simply transitive actions,
894 we train the model on tasks \mathcal{T}^1 and \mathcal{T}^2 via Algorithm 1, using the next-clause prediction loss defined
895 in Definition 4.2. For symmetry group actions, the model is similarly trained on tasks \mathcal{T}^1 and \mathcal{T}^2 ,
896 but additionally employs a recursive self-training procedure (Definition 4.4) for longer tasks \mathcal{T}^{2^k} , as
897 described in Algorithm 2. Our model employs structured attention parameters (Assumption 3.4) and
898 is initialized according to Assumption 3.3. All theoretical results are established in the asymptotic
899 regime specified in Assumption 3.1, where vocabulary and variable set sizes grow large, while action
900 and value sets remain relatively small. We further consider the following assumptions on the model
901 output throughout the proof:

902 **Assumption A.1.** We assume there exists $B = \lambda \log d$ for some sufficiently large constant $\lambda > 0$,
903 such that if the raw model output $[F_i]_j$ exceeds B , it is truncated as follows: $[F_i]_j \leftarrow \min\{[F_i]_j, B\}$.

904 In this section, we provide an overview of the proof of the main theorem. The proof is divided
905 into three parts: (1) learning the one-step reasoning mechanism for solving the LEGO task \mathcal{T}^1 ,
906 including in-context retrieval of variables (Section A.1) and group operations (Section A.2); (2)
907 learning direct short-to-poly CoT length generalization on task \mathcal{T}^2 with simply transitive group
908 actions (Section A.3.1); and (3) recursive length generalization via self-training on task \mathcal{T}^{2^k} with
909 symmetry group actions (Section A.3.2).

910 Our training algorithm proceeds by first training the FFN layer parameter \mathbf{W} to solve the one-step
911 reasoning task \mathcal{T}^1 , then training the attention layer parameter \mathbf{Q} for task \mathcal{T}^2 and tasks \mathcal{T}^{2^k} . The
912 alternating between the two parameter update is to simplify the analysis and to present the core ideas.
913 Such separation of analysis also sheds light on the different roles between the FFN and the attention
914 layer, for the synthetic CoT task we consider.

915 A.1 Learning In-Context Retrieval of Variables

916 We demonstrate that for task \mathcal{T}^1 , the model correctly retrieves the target variable x_1 from the first
917 token of the first predicate clause $Z_{\text{pred},1}$ to accurately predict the 4-th token in the answer clause
918 $Z_{\text{ans},1}$.

919 The central idea is to rigorously track the training dynamics of the weights $\mathbf{W}_{4,j,r,p}$. We prove that
920 weights corresponding to retrieving the correct variable grow significantly, while all others remain
921 small. Specifically, we proceed in four steps:

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

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

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

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

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

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

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

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

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

933 **4. Bounding non-target correlations.** We finally show by induction that all other correlations
 934 $\langle \mathbf{W}_{4,j,r,p}^{(t)}, e_s \rangle$ —including group actions, value tokens, off-diagonal tokens, and non-target
 935 variables—remain negligible throughout the training process.

936 A.2 Learning the Group Operations

937 We briefly summarize the proof of how the model learns to track cyclic and symmetry group
 938 operations. Consider the following definitions.

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

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

941 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
 942 predicting $y = \tau^{-1}(j)$.

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

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

944 **Definition A.3** (ψ, Ψ -Notations, Cyclic Group). For $j \in \tau(\mathcal{Y})$, $r \in [m]$, and $\phi = (g, y) \in \mathfrak{F}_j$, define

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

945 and the composite feature magnitude as

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

946 corresponding to attention weights $\frac{1}{2}$ for $Z_{\text{pred},1}$ and $Z_{\text{ans},0}$.

947 We will analyze the evolution of features $\Psi_{\mathbf{u}}$ according to the initial ordering induced by their
 948 magnitudes at initialization:

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

949 We denote the set of indices \mathcal{U} equipped with this partial order as \mathcal{U}^* .

950 **Training Phases: Cyclic Group.** We shall separate the analysis of the training process into two
 951 phases. The first phase characterizes the emergence of the feature $\Psi_{\mathbf{u}}$ among other features. The
 952 second phase characterizes how the prediction of F_5 sharpens as the feature $\Psi_{\mathbf{u}}$ grows. The third
 953 phase characterizes the convergence of the feature $\Psi_{\mathbf{u}}$ and related quantities.

954 We describe roughly the proof overview of these phases below: for every $\mathbf{u} \in \mathcal{U}^*$, assume (which
 955 we are going to prove) that for the immediate predecessor $\mathbf{u}' \prec \mathbf{u}$ we have achieved its hitting time
 956 $t = T_{\mathbf{u}',3}$. Then the growth of $\Psi_{\mathbf{u}}^{(t)}$ proceeds as follows.

957 (I) Phase I: from $t = 0$ to $t = T_{\mathbf{u},1}$. There are two subphases:

958 (a) In phase I.a $t \in [0, T_{\mathbf{u},1a}]$, the growth of $\Psi_{\mathbf{u}}^{(t)}$ experience competitions with other features,
 959 both intra and inter neurons. By tensor power method, the feature with the largest initial
 960 activation will win and succeed in the learning order.

961 (b) In phase I.b $t \in [T_{u,1a}, T_{u,1}]$, the feature $\Psi_u^{(t)}$ arrives at a level $\Psi_u^{(t)} \geq \Omega(\log d)$, which
 962 increase the prediction logit $\text{logit}_{5,j}^{(t)} \geq \frac{1}{d^{0.01}}$ when this feature appears. After this point,
 963 features started to experience non-monotonic growth and we cannot fully characterize its
 964 process anymore.

965 (II) Phase II: from $t = T_{u,1}$ to $t = T_1$. This phase is further divided into three subphases.

- 966 (a) In phase II.a $t \in [T_{u,1}, T_{u,2a}]$, the feature $\Psi_u^{(t)}$ grows to the point where the incorrect
 967 features must cancel to small activation for small gradient.
 968 (b) In phase II.b $t \in [T_{u,2a}, T_{u,2}]$, the incorrect feature combinations are cancelled to a very
 969 small activation and the loss converged.
 970 (c) In phase II.c $t \in [T_{u,2}, T_1]$, the loss and the features converged.

971 **Symmetry Case.** The proof strategy for symmetry group actions is analogous to the cyclic case
 972 but involves more nuanced control of the training dynamics due to the complex interactions among
 973 symmetric group features. The analysis similarly progresses through emergence, refinement, and
 974 convergence phases, with carefully controlled arguments to handle symmetric structures and their
 975 induced interactions.

976 A.3 Learning the Attention Layer

977 Successful training on \mathcal{T}^1 demonstrates that the model can perform one-step reasoning to compute
 978 $y_1 = g_1(y_0)$. Building on this insight, we now consider the more challenging task \mathcal{T}^2 . Given the
 979 input $Z^{2,1}$, the model must identify both the correct predicate clause $Z_{\text{pred},2}$, which contains g_2 ,
 980 and the answer clause $Z_{\text{ans},1}$, which contains the value y_1 , in order to compute the correct answer
 981 $y_2 = g_2(y_1)$.

982 Since the reasoning mechanism—namely, the group operation—has already been captured by the
 983 trained FFN, the remaining challenge lies in directing attention to the appropriate locations. We
 984 show that this can be achieved by learning an attention pattern with a specific routing structure,
 985 which we refer to as **attention concentration**. Specifically, given an input $Z^{L,\ell}$, the attention
 986 weights concentrate on $\text{Attn}_{\text{ans},\ell \rightarrow \text{pred},\ell+1}$ and $\text{Attn}_{\text{ans},\ell \rightarrow \text{ans},\ell}$, corresponding respectively to the
 987 next predicate clause and the previous answer clause.

988 This behavior is closely tied to the structure of the query matrix \mathbf{Q} and the way clause embeddings
 989 are organized. In particular, $\mathbf{Q}_{4,3}$ governs attention toward the predicate clauses, while $\mathbf{Q}_{4,4}$ governs
 990 attention toward the answer clauses.

991 We quantify the quality of this attention routing via the attention concentration gap:

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

992 which measures the total fraction of attention mass not allocated to the two key clauses. In the
 993 following analysis, we examine how $\epsilon_{\text{attn}}^{L,\ell}$ and its learning dynamics evolve under different types of
 994 group operations.

995 A.3.1 Simply Transitive Group

996 For the simply transitive group, we show that at stage \mathcal{T}^2 , the attention concentration gap ϵ_{attn} can be
 997 reduced below $O(1/\text{poly}(d))$, indicating highly focused attention on the relevant clauses.

998 Throughout the analysis, we use $r_{g \cdot y}$ to denote the neuron index satisfying

$$\frac{1}{2} \langle W_{5,\tau(g \cdot y),r_{g \cdot y},2}, e_g \rangle + \frac{1}{2} \langle W_{5,\tau(g \cdot y),r_{g \cdot y},5}, e_y \rangle \approx B,$$

999 which corresponds to the activation that successfully predicts $j = \tau(g \cdot y)$ in \mathcal{T}^1 .

1000 We further introduce the following notation to quantify the attention gap between the correct predicate
 1001 clause and the answer clause:

$$\Delta^{L,\ell} = \text{Attn}_{\text{ans},\ell \rightarrow \text{pred},\ell+1}(\mathbf{Z}^{(L,\ell)}) - \text{Attn}_{\text{ans},\ell \rightarrow \text{ans},\ell}(\mathbf{Z}^{(L,\ell)}).$$

Roadmap of the Proof. We divide the analysis into three sub-stages. Throughout these stages, the gradients of the diagonal entries $[\mathbf{Q}_{4,3}]_{j,j}$ and $[\mathbf{Q}_{4,4}]_{j,j}$ for $j \in \tau(\mathcal{X})$ are order-wise larger than those of all other entries. As a result, the dynamics are dominated by these diagonal components, and we focus our analysis on them. For simplicity of notation, we denote $[\mathbf{Q}_{4,3}]_{j,j}$ and $[\mathbf{Q}_{4,4}]_{j,j}$ simply as $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ in the remaining discussion.

- **Stage 2.1: Growth of Initial Gap.** We show that $\mathbf{Q}_{4,3}$ grows faster than $\mathbf{Q}_{4,4}$ due to larger gradient contributions from the loss of $\ell = 1$, while the gradient from $\ell = 2$ remains negligible. As a result, a gap of magnitude $\Omega(1/\log d)$ emerges between $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$, leading to an early advantage in attention routing toward the predicate clause $(\text{pred}, 2)$.
- **Stage 2.2: Joint Growth with Controlled Gap.** In this phase, both $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ continue to grow to constant scale. The gradient contributions from $\ell = 2$ gradually become dominant, while the gap is maintained within $[\Omega(1/\log d), O(1)]$. Throughout this stage, the attention gap satisfies $\Delta^{2,1} = \Omega(1/\log d)$.
- **Stage 2.3: Convergence and Gap Reduction.** In the final phase, the continued joint growth of $\mathbf{Q}_{4,4}$ and $\mathbf{Q}_{4,3}$ lead the attention to concentrate near its ideal limit: $\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}(\mathbf{Z}^{(2,1)}) + \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}(\mathbf{Z}^{(2,1)}) \geq 1 - \epsilon_{\text{attn}}^{2,1}$. Throughout this process, the attention gap $\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}$ cannot remain above a certain threshold $o(1)$ for long; otherwise, the incorrect logit $\text{logit}_{5,\tau(g_2(y_0))}$ would receive a stronger gradient signal and drive $\mathbf{Q}_{4,4}$ to grow faster than $\mathbf{Q}_{4,3}$, contradicting the attention dominance assumption. Consequently, at convergence, we guarantee: (i) $\epsilon_{\text{attn}}^{2,1} \leq \frac{1}{\text{poly}d}$; (ii) both $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ reach $\Omega(\log d)$; and (iii) the total loss satisfies $\sum_{\ell=1}^2 \text{Loss}_5^{2,\ell} \leq \frac{1}{\text{poly}(d)}$.

1024 A.3.2 Symmetry Group

1025 In the following discussion, we focus on symmetry-group tasks $\mathcal{T}^{(L)}$, where only the input $Z^{(L,1)}$ is
 1026 used to predict the value token in $Z_{\text{ans},2}$. For notational simplicity, we drop the superscript ℓ from
 1027 $\epsilon_{\text{attn}}^{L,\ell}$ and $\Delta^{L,\ell}$ when no ambiguity arises.

1028 Roadmap of the Proof for \mathcal{T}^2

- **Stage 1.2.1: Growth of Initial Gap.** Similar to earlier phases, we show that $\mathbf{Q}_{4,3}$ grows faster than $\mathbf{Q}_{4,4}$, but this time due to larger gradient contributions from the loss at $\ell = 2$. As a result, a gap $\Delta^{2,1}$ of magnitude $\Omega(1/\log d)$ emerges between $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$, which induces an early advantage in attention routing toward the predicate clause $(\text{pred}, 2)$.
- **Stage 1.2.2: Convergence and Controlled Gap.** After the initial phase, we show that as long as $\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}$ remains below $1/2$, the total gradient on $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ remains positively lower bounded. As a result, $\mathbf{Q}_{4,3} + \mathbf{Q}_{4,4}$ will continue to grow until $\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}$ approaches $1/2$. Furthermore, we prove that the attention gap Δ^2 cannot stay above a fixed small constant. Therefore, when $\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}$ nears $1/2$, the attention to the answer clause, $\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}$, must also be close to $1/2$. This results in the off-target attention mass ϵ_{attn}^2 being reduced to a small constant. Combining this attention behavior with the FFN structure established in the previous stage, we obtain that $\Lambda_{5,\tau(g_2(y_1)),r_{g_2,y_1}} \approx B$ while all other logits remain small. This ensures correct prediction, and the total loss satisfies $\text{Loss}_5^{2,2} \leq \frac{1}{\text{poly}(d)}$.

1043 **Roadmap of the Proof for \mathcal{T}^{2^k} for $k \geq 2$** Since $\epsilon_{\text{attn}}^{2^{k-1}}$ has already been reduced to a small constant
 1044 in the previous stage, we begin the current stage with nearly concentrated attention. That is, at the
 1045 start of \mathcal{T}^{2^k} , we still have $\epsilon_{\text{attn}}^{2^k} \leq 2\epsilon_{\text{attn}}^{2^{k-1}}$, which remains small. Moreover, $\Delta^{2^k} \leq \Delta^{2^{k-1}}$. As a
 1046 result, the attention structure does not deviate significantly from that in \mathcal{T}^2 . This further implies
 1047 that the input $Z^{(2^k,2)}$ follows a bootstrapped LEGO distribution generated by the greedy language
 1048 model $\hat{p}_{F^{T_{k-1}}}$, which coincides with the original LEGO distribution. In particular, the answers y_1
 1049 and y_2 are correct, satisfying $y_1 = g_1(y_0)$ and $y_2 = g_2(y_1)$. This allows us to directly apply a similar
 1050 convergence analysis as in \mathcal{T}^2 , and show that $\epsilon_{\text{attn}}^{2^k}$ decreases further to a small constant.

1051 B Learning In-Context Retrieval of Variables

1052 B.1 Preliminaries

1053 First we define some notations for the presentation of gradients.

1054 **Notations for gradient expressions** For each $i \in [5], m \in [L], j \in [d]$, we denote

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

1055 where $\text{logit}_{i,j}(F, \mathbf{Z}^{(L,\ell-1)})$ are defined as

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

1056 **Fact B.1.** For any $i \in [5], j \in [d], r \in [m]$

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

1057 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
1058 $\mathcal{E}_{i,j}$ when the context is clear.

1059 Given $\mathbf{Z}^{(L)}$, we use $\hat{\mathcal{X}}^{(L)}$ to denote the appeared variables in the context clauses, i.e. $\hat{\mathcal{X}}^{(L)} =$
1060 $\{x_0, x_1, \dots, x_L\}$. We write $\hat{\mathcal{X}}^{(L)}$ as $\hat{\mathcal{X}}$ for simplicity. Throughout this section, we write $[F_i]_j$ as $F_{i,j}$
1061 for simplicity.

1062 B.2 Induction Hypothesis

1063 In this stage, we consider the learning process for $\mathbf{W}_{4,\cdot}$.

1064 **Induction B.1.** For $t \leq T = \frac{\text{poly}d}{\eta}$, all of the following holds:

- 1065 (a). for $j \in \tau(\mathcal{X})$, $\tilde{\Omega}(\sigma_0) \leq \langle \mathbf{W}_{4,j,r,1}^{(t)}, e_j \rangle + \mu \leq \tilde{O}(1)$, where $\langle \mathbf{W}_{4,j,r,1}^{(t)}, e_j \rangle$ is non-decreasing;
(b). for $j \in \tau(\mathcal{X})$, $g \in \mathcal{G}$

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

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

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

- 1066 (d). Else, $|\langle \mathbf{W}_{4,j,r,p}^{(t)}, e_s \rangle| \leq \tilde{O}(\sigma_0)$;

1067 **Claim B.1.** If Induction B.1 holds at iteration t , then for a sequence \mathbf{Z}

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

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

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

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

1068 • otherwise, $0 \leq \Lambda_{4,j,r}^{(t)} \leq \frac{5}{2}\mu + \tilde{O}(\sigma_0)$.

1069 **Claim B.2.** If Induction B.1 holds at iteration t , then for a sequence \mathbf{Z} ,

1070 • if $j = \tau(x_1)$, $\text{logit}_{4,j}^{(t)} = \frac{e^{O(\Phi_{4,j}^{(t)})}}{e^{O(\Phi_{4,j}^{(t)})} + d}$;

1071 • otherwise, $\text{logit}_{4,j}^{(t)} = O(\frac{1}{d}) \left(1 - \text{logit}_{\tau(x_1)}^{(t)}\right)$.

1072 *Proof.* If $j = \tau(x_1)$, by Induction B.1 and Claim B.1, we have

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

1073 for $j \in \tau(\mathcal{X}) \neq \tau(x_1)$, $F_{4,j}^{(t)}(\mathbf{Z}) \leq \tilde{O}(\sigma_0) + O(\frac{\max\{\Phi_{4,j}^{(t)}, \Phi_{4,j^*}^{(t)}\}}{|\mathcal{G}|})$; else $F_{4,j}^{(t)}(\mathbf{Z}) \leq \tilde{O}(\sigma_0)$. Combining
1074 together, we prove the result. \square

1075 B.3 Gradient Lemma

1076 Starting with the gradient computation:

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

1077 We first consider the gradient for $j \in \tau(\mathcal{X})$

1078 **Lemma B.1.** For $j \in \tau(\mathcal{X})$, we have

1079 (a) for $\mathbf{W}_{4,j,r,1}$, $s \in \tau(\mathcal{X})$

1080 (1) if $s = j$, $\langle -\nabla_{\mathbf{W}_{4,j,r,1}}^{(t)} \text{Loss}^1, e_s \rangle = \frac{1}{2} \mathbb{E} \left[(1 - \text{logit}_{4,j}^{(t)}) \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_1)=j} \right]$

1081 (2) $s \neq j$, $\langle -\nabla_{\mathbf{W}_{4,j,r,1}}^{(t)} \text{Loss}^1, e_s \rangle = \frac{1}{2} \mathbb{E} \left[-\text{logit}_{4,j}^{(t)} \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_1)=s} \right]$

1082 (b) for $\mathbf{W}_{4,j,r,2}$, $s = \tau(g)$ for $g \in \mathcal{G}$

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

1083 (c) for $\mathbf{W}_{4,j,r,p}$ with $p \in \{3, 4\}$, $s \in \tau(\mathcal{X})$

1084 (1) $s = j$, $\langle -\nabla_{\mathbf{W}_{4,j,r,3}}^{(t)} \text{Loss}^1, e_j \rangle = \frac{1}{2} \mathbb{E} \left[-\text{logit}_{4,j}^{(t)} \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{\tau(x_0)=j} \right]$

1085 (2) $s \neq j$

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

1086 (d) for $\mathbf{W}_{4,j,r,5}$, $s = \tau(y)$ for $y \in \mathcal{Y}$

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

1087 Then for $j \notin \tau(\mathcal{X})$, we can obtain

1088 **Lemma B.2.** For $j \notin \tau(\mathcal{X})$, we have

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

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

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

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

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

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

1089 (d) for $\mathbf{W}_{4,j,r,5}$, $s = \tau(y)$ for $g \in \mathcal{Y}$

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

1090 B.4 Growth of Gamma

1091 **Lemma B.3** (Growth). Given $j \in \tau(\mathcal{X})$, suppose Induction B.1 holds at iteration t , when $\Phi_{4,j}^{(t)} \leq$
 1092 $0.01 \log d$ or $\Gamma_{4,j}^{(t)} \leq \frac{0.01 \log d}{m}$, then it satisfies

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

1093 *Proof.* By Lemma B.1, we have

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

1094 By Claim B.2, when $\Phi_{4,j}^{(t)} \leq 0.01 \log d$, $\text{logit}_{4,j}^{(t)} = \frac{O(e^{0.01 \log d})}{O(e^{0.01 \log d}) + d} \ll 1$ when $j = \tau(x_1)$; and
 1095 combining with the fact that the event $\{\tau(x_1) = j\}$ happens with probability $\frac{1}{|\mathcal{X}|}$, we complete the
 1096 proof. \square

1097 Lemma B.3, combined with the growth of the tensor power method, immediately gives the following
 1098 corollary.

1099 **Lemma B.4.** Suppose Induction B.1 holds for all iterations. Define threshold $\Lambda^- = \Theta(\frac{1}{m})$. Let $T_{1,j}$
 1100 be the first iteration so that $\Gamma_{4,j}^{(t)} \geq \Lambda^-$, and $T_1 \stackrel{\text{def}}{=} \Theta(\frac{d}{\eta \sigma_0^{q-2}})$. Then we have $T_1 \geq T_{1,j}$ for every
 1101 $j \in \tau(\mathcal{X})$, i.e., for $t \geq T_1$, it satisfies $\Gamma_{4,j}^{(t)} \geq \Lambda^-$.

1102 **Lemma B.5** (Upper bound). Suppose Induction B.1 holds for all iterations $< t$, we have $\Phi_{4,j}^{(t)} \leq \tilde{O}(1)$,
 1103 for $j \in \tau(\mathcal{X})$.

1104 *Proof.* We only need to consider the time $t \geq T_1$. Notice that the gradient descent update in
 1105 Lemma B.1 gives

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

1106 Therefore, for sufficiently small η , we have

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

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

1107 When there exists \tilde{T} , s.t., $\max_{j \in \tau(\mathcal{X})} \Phi_{4,j}^{(\tilde{T})} > \Omega(\log^{1.5} d)$, by Induction B.1 and Claim B.1, given a
 1108 sequence \mathbf{Z} with $\tau(x_1) = \tilde{j} = \arg \max_{j \in \tau(\mathcal{X})} \Phi_{4,j}^{(\tilde{T})}$, we have

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

1109 Following the similar analysis as Claim B.2, $F_{4,j'}^{(\tilde{T})}(\mathbf{Z}) \leq O(\frac{\Phi_{4,j'}^{(\tilde{T})}}{|\mathcal{G}|})$ for other $j' \in \tau(\mathcal{X})$, and
 1110 $F_{4,j'}^{(\tilde{T})}(\mathbf{Z}) \leq o(1)$ for $j' \notin \tau(\mathcal{X})$, which implies $1 - \mathbf{logit}_{4,j}^{(\tilde{T})} = e^{-\Omega(\log^{1.5} d)}$. Therefore, we derive
 1111 that for $t \in [\tilde{T} + 1, \frac{\text{poly}d}{\eta}]$,

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

1112 since $\varrho \ll \frac{1}{m \log d}$ which implies $\Phi_{4,\tilde{j}}^{(t)} \leq O(\log^{1.5} d)$. □

1113 B.5 Group and Value Correlations Are Not Large

1114 **Lemma B.6.** Suppose Induction B.1 holds for all iterations $< t$, then for any $j \in \tau(\mathcal{X})$ and
 1115 $s = \tau(g)$, $g \in \mathcal{G}$, we have

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

1116 *Proof.* By Lemma B.1, we have

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

1117 Clearly, the positive gradient can be upper bounded by $O(\frac{1}{|\mathcal{G}|} \langle -\nabla_{\mathbf{W}_{4,j,r,1}^{(t)}} \text{Loss}^1, e_j \rangle)$. Moreover, for
 1118 the negative gradient, by Claim B.1, we have a naive bound

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

1119 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 $\mathbf{logit}_{4,j}^{(t)}|_{j \neq \tau(x_1)} \leq O(\frac{1}{d})$,
 1120 which implies

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

1121 Therefore, for $t \leq T_1$, we have

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

1122 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 $r' \in \mathcal{A}_{4,j}^{(t)}$

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

1123 furthermore, $\mathbf{logit}_{4,j}^{(t)}|_{j \neq \tau(x_1)} \leq O(\frac{1}{d})(1 - \mathbf{logit}_{4,j^*}^{(t)}|_{j^*=\tau(x_1)})$, which implies

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

1124 Due to the arbitrary of r' , we have

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

1125 □

1126 **Lemma B.7.** Suppose Induction B.1 holds for all iterations $< t$, then for any $j \in \tau(\mathcal{X})$ and
 1127 $s = \tau(y)$, $y \in \mathcal{Y}$, we have

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

1128 *Proof.* The proof is similar as Lemma B.6. □

1129 B.6 Off-diagonal Correlations Are Small

1130 **Lemma B.8** (off-diagonal bound). Given $j \in \tau(\mathcal{X})$, suppose Induction B.1 holds at all iterations
 1131 $< t$, for $s \in \tau(\mathcal{X}) \neq j$

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

1132 *Proof.* By Lemma B.1

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

1133 Notice that by Claim B.1,

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

1134 combined with Claim B.2, $\text{logit}_{4,j} \leq O(\frac{1}{d})(1 - \text{logit}_{4,s}^{(t)})$ when $s = \tau(x_1)$, thus

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

From Induction B.1, we have

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

1135 □

1136 **Lemma B.9.** Given $j \in \tau(\mathcal{X})$, suppose Induction B.1 holds at all iterations $< t$, we have

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

1137 *Proof.* When $s = j$, we have

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

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

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

1138 When $s \neq j$,

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

1139 Noticing that $\{\tau(x_0) = s, \tau(x_1) = j\}$ happens with probability $\frac{1}{|\mathcal{X}|(|\mathcal{X}|-1)}$, thus the positive gradient

1140 can be upper bounded by $O(\frac{1}{d}) \cdot |\langle -\nabla_{\mathbf{W}_{4,j,r,1}} \text{Loss}^1, e_j \rangle|$. Furthermore, the negative part can be
 1141 upper bounded in the similar way as previous off-diagonal negative gradient. Putting it together, we
 1142 complete the proof. □

1143 B.7 Non-target Correlations Are Negligible

1144 **Lemma B.10.** Suppose Induction B.1 holds at all iterations $< t$, for $j' \notin \tau(\mathcal{X})$, for $p \in [5]$ and
 1145 $s \in [d]$

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

1146 *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
 1147 way previous off-diagonal negative gradient.

1148 We can observe that for $j' \notin \tau(\mathcal{X})$, all the non-zero gradient on the different directions are negative
 1149 gradient, which implies $\langle \mathbf{W}_{4,j',r,p}^{(t)}, e_s \rangle \leq \langle \mathbf{W}_{4,j',r,p}^{(0)}, e_s \rangle = \tilde{O}(\sigma_0)$. Moreover, $\Lambda_{4,j',r}^{(t)} \leq \tilde{O}(\sigma_0)$ is
 1150 also non-increasing.

1151 For $s = \tau(g)$, $g \in \mathcal{G}$, whenever $\langle \mathbf{W}_{4,j',r,2}^{(t)}, e_s \rangle$ reaches -3μ , we have $\Lambda_{4,j',r}^{(t)}|_{g_1=g} \leq -3\mu + \frac{5}{2}\mu +$
 1152 $\tilde{O}(\sigma_0) \leq 0$, and thus $\langle -\nabla_{\mathbf{W}_{4,j',r,2}^{(t)}} \text{Loss}^1, e_s \rangle = \frac{1}{2} \mathbb{E} \left[-\text{logit}_{4,j}^{(t)} \text{sReLU}'(\Lambda_{4,j,r}^{(t)}) \mathbb{1}_{g_1=g} \right] = 0$, which
 1153 implies $\langle \mathbf{W}_{4,j',r,2}^{(t)}, e_s \rangle \geq -3\mu$. Hence, $|\langle \mathbf{W}_{4,j',r,2}^{(t)}, e_s \rangle| \leq \tilde{O}(\sigma_0)$. Following the similar argument,
 1154 we can prove the result for $\langle \mathbf{W}_{4,j',r,5}^{(t)}, e_s \rangle$ for $s \in \tau(\mathcal{Y})$. \square

1155 B.8 Convergence

1156 **Lemma B.11.** For $|\mathcal{G}| \geq |\mathcal{Y}| \geq \Omega(\frac{\log \log d}{\log \log \log d})$, $\text{polylog} d \geq m \geq |\mathcal{Y}|$, $\varrho \ll \frac{1}{m \log d}$ and sufficiently
 1157 small $\eta \leq \frac{1}{\text{poly} d}$, Induction B.1 holds for all iterations $t \leq T = \frac{\text{poly} d}{\eta}$.

1158 *Proof.* Putting the results in Lemmas B.5 to B.8 and B.10, we can directly establish the results in
 1159 Induction B.1. \square

1160 **Lemma B.12 (Convergence).** For sufficiently large $T_1 \leq t = \frac{\text{poly} d}{\eta}$, we have

1161 (a) Objective convergence: $\text{Loss}^1 \leq \frac{1}{\text{poly} d}$;

1162 (b) Successful learning of diagonal feature: $\Phi_{4,j}^{(t)} \geq \Omega(\log d)$ for any $j \in \tau(\mathcal{X})$.

1163 *Proof.* Assuming for some sufficiently large constant $n > 0$, $\mathbb{E}[(1 - \text{logit}_{4,j^*}^{(t)}) \mid \tau(x_1) = j^*] \geq$
 1164 $\Omega(\frac{1}{d^n})$ for $t \in (T_1, T_1 + \frac{d^{n+1} \log^2 d}{\eta}]$ then by Lemma B.1, we have

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

1165 which contradicts with $\Gamma_{4,j^{(*)}}^{(t)} \leq \Phi_{4,j^{(*)}}^{(t)} \leq O(\log^{1.5} d) = \tilde{O}(1)$ in the polynomial time. This
 1166 implies after sufficiently large iteration t , we must have $\mathbb{E}[(1 - \text{logit}_{4,j}^{(t)}) \mid \tau(x_1) = j] \leq O(\frac{1}{d^n})$ for
 1167 $j \in \tau(\mathcal{X})$. Hence

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

1168 By Claim B.2, at the time of convergence, we must have $\Phi_{4,j}^{(t)} \geq \Omega(\log d)$. \square

1169 C Learning The Group Actions: Cyclic Group

1170 In the LEGO language, one step of the state transition corresponds predicting the next answer clause
 1171 from the current sequence $Z^{(L,L')} \sim \mathcal{D}^{L,L'}$. As in Algorithm 1 and 2, we start with training the
 1172 model F on length-1 sequences, which only requires the model to predict one answer $Z_{\text{ans},1}$ given
 1173 input clauses $Z_{\text{pred},1}$ and $Z_{\text{ans},0}$. In this appendix section, we show how the model learns to predict
 1174 the 5-th token of $Z_{\text{ans},1}$, that is, the value of x_1 , in the LEGO sentence $Z^{(1)}$ in (3).

1175 Let's recall the structure of the uniform case setting. Note that the group action defined in Assump-
 1176 tion 4.1 is equivalent to the following group action by group isomorphism:

1177 **Assumption C.1** (Assumption 4.1, restated). Let $\mathbf{LEGO}(\mathcal{X}, \mathcal{G}, \mathcal{Y})$ be the LEGO language. We
 1178 assume $\mathcal{Y} = \{0, 1, \dots, n_y - 1\}$ and $\mathcal{G} = C_{|\mathcal{Y}|}$, i.e., the cyclic group of order $|\mathcal{Y}|$, and $n_y \in$
 1179 $[\Omega(\log \log d), \log d]$.

1180 We define some notations for this section here.

1181 **Notations.** Let \mathcal{D}^1 be the LEGO distribution of length 1 under the language $\mathbf{LEGO}(\mathcal{X}, \mathcal{G}, \mathcal{Y})$. We
 1182 define $\mathcal{D}_{\mathcal{X}}^1$, $\mathcal{D}_{\mathcal{G}}^1$ and $\mathcal{D}_{\mathcal{Y}}^1$ be the distribution of (x_0, x_1) , g_0 and (y_0, y_1) in \mathcal{D}^1 respectively. That is,
 1183 given a LEGO sentence

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

1184 The sampling distribution of (x_0, x_1) is $\mathcal{D}_{\mathcal{X}}^1$, and similarly for g_0 and (y_0, y_1) .

1185 C.1 Preliminaries and Induction Hypotheses

1186 First we compute the expression of gradients for \mathbf{W} here. With slight abuse of notation, we write
 1187 $\text{Loss}^{(t)} \equiv \text{Loss}(F^{(t)}) \equiv \text{Loss}^{1,0}(F^{(t)})$ in this stage. The gradients are given by:

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

1188 where $\mathcal{E}_{i,j}(\mathbf{Z}^1) = \mathbb{1}_{\tau(Z_{\text{ans},1,i})=j} - \text{logit}_{i,j}(F, \mathbf{Z}^{1,0})$ is the

1189 To analyze the learning of the group actions, we need to first define the set of features that the model
 1190 will learn, corresponding to the group action.

1191 **Definition C.1** (feature combinations, cyclic group). Assuming the group \mathcal{G} follows Assumption C.1.
 1192 For each $j \in \tau(\mathcal{Y})$, let

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

1193 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
 1194 feature combinations for predicting $y = \tau^{-1}(j)$. Furthermore, for any $\phi = (g, y) \in \mathfrak{F}$, we write

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

1195 as the set of **confounding features** for ϕ , that is, the features that share exactly one component with
 1196 ϕ , either g or y .

1197 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
 1198 predicting the corresponding token index $j = \tau(y')$. It is the set of features we want our network to
 1199 learn in the neurons of output coordinate j , while sets $\mathfrak{F}_{j'}, j' \neq j$ are the sets of features we do not
 1200 want to learn in the neurons of output coordinate j' .

1201 The set of confounding features $\mathfrak{F}_{\text{conf}}(\phi)$ for a given feature $\phi = (g, y) \in \mathfrak{F}$ contains the features
 1202 that share exactly one component with ϕ , either g or y . Confounding features are the combination of
 1203 features that are similar to ϕ but are incorrect for predicting j -th output.

1204 Features in sets \mathfrak{F}_j exist in the neurons. We define a short notation for the set of all indices of feature
 1205 combinations at coordinate $j \in \tau(\mathcal{Y})$ and neuron $r \in [m]$.

1206 **Definition C.2** (neuron feature indices). We define

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

1207 be the set of all indices of compositional features.

1208 Now we present the some notations that could help us determine which features are learned first.

1209 **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
1210 following notation:

$$\psi_{j,r}(g) := \langle \mathbf{W}_{5,j,r,2}, e_g \rangle, \quad \psi_{j,r}(y) := \langle \mathbf{W}_{5,j,r,5}, e_y \rangle \quad (10)$$

1211 When cG follows Assumption C.1, we define an index $\mathbf{u} = (j, r, \phi)$ and a feature magnitude
1212 $\Psi_{\mathbf{u}} \equiv \Psi_{j,r}(g, y)$ as follows:

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

1213 which is the combination of features we need to predict the correct answer $y' = g \cdot y$ with fixed
1214 attention weights $\frac{1}{2}$ for both $Z_{\text{pred},1}$ and $Z_{\text{ans},0}$ during training phase I.

1215 A key technical ingredient of our proof is the characterization of the learning order of the features.
1216 By leveraging the smoothness of the **sReLU** activationfunction, we can show that the features are
1217 learned in a specific order that relates to the feature magnitude $\Psi_{\mathbf{u}}^{(0)}$ at initialization. We define the
1218 learning order, encoded by the order \prec on \mathcal{U}^* as follows:

1219 **Definition C.4** (learning order). The *learning order* is the ordered set \mathcal{U}^* that we obtain from the
1220 following process: Define a total order on \mathcal{U} as follows:

$$\mathbf{u} \prec \mathbf{u}' \iff \Psi_{\mathbf{u}}^{(0)} \geq \Psi_{\mathbf{u}'}^{(0)} \quad \forall \mathbf{u}, \mathbf{u}' \in \mathcal{U} \quad (11)$$

1221 We construct the sets \mathcal{U}^* by the following procedure: initialize an empty neuron set $\mathcal{W}_{tmp}^{(0)} = \emptyset$, and
1222 an empty feature set $\mathcal{R}_{tmp}^{(0)} = \emptyset$, and the initial index set $\mathcal{U}^{(0)} = \emptyset$. Starting from $k = 1$, we do the
1223 following:

- 1224 (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
1225 feature $\phi \in \mathfrak{F} \setminus \mathcal{R}_{tmp}^{(k-1)}$ and $(j, r) \in \tau(\mathcal{Y}) \times [m] \setminus \mathcal{W}_{tmp}^{(k-1)}$.
- 1226 (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 $\mathcal{U}^{(k)} \leftarrow \mathcal{U}^{(k-1)} \cup \{\mathbf{u}\}$.
- 1227 (3) Iterate the (1) and (2) steps until $k = n_y^2$, then yield $\mathcal{U}^* \equiv \mathcal{U}^{(n_y^2)}$.

1228 This process yields the ordered set \mathcal{U}^* , equipped with the total order \prec defined in (11).

1229 Note that \mathcal{U}^* is a smaller subset of \mathcal{U} . In fact, \mathcal{U}^* encodes the order that the neural network F learn
1230 the features ϕ in the neurons $(j, r) \in [d] \times [m]$, and leave out the indices $\mathbf{u} \in \mathcal{U} \setminus \mathcal{U}^*$ that are not
1231 learned. The order \prec is induced by the feature magnitude $\Psi_{j,r}^{(0)}(\phi)$. In the proof we shall show with
1232 high probability, the order that each ϕ is learned is according to its position in \mathcal{U}^* .

1233 In order to characterize the feature updates before they start to become significant, we define the
1234 following notion of *pseudo weights*.

1235 **Definition C.5** (pseudo weights). Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}$. We define the pseudo weight $\widetilde{\mathbf{W}}^{(t)}(\mathbf{u})$
1236 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
1237 $\widetilde{\mathbf{W}}_{5,j,r}^{(0)}(\mathbf{u}) \equiv \mathbf{W}_{5,j,r}^{(0)}$, and let $\widetilde{\Lambda}_{5,j,r}^{(t)}$ be the corresponding activation with weights $\widetilde{\mathbf{W}}_{5,j,r}^{(t)}$. We define
1238 the update rule of $\widetilde{\mathbf{W}}_{5,j,r}^{(t)}(\mathbf{u})$ in the following manner:

- 1239 • if $(p, v) \notin \{(2, g), (5, y)\}$, we define $\langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t+1)}(\mathbf{u}), e_v \rangle \equiv \langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v \rangle$;
- 1240 • if $(p, v) \in \{(2, g), (5, y)\}$, then we let the update rule to be:

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

1241 This means that we are updating the pseudo weights for the 2nd and 5th components of the weight
1242 vectors differently, while keeping all other components' updates unchanged.

1243 In order to introduce the induction hypothesis, we first define the probability events of feature
1244 appearance.

1245 **Definition C.6** (probability events of feature appearance). Let $j \in \tau(\mathcal{Y})$, $r \in [m]$ and $\phi = (g, y) \in$
1246 \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 $\tilde{\mathcal{B}}_\phi, \tilde{\mathcal{B}}_j(g), \tilde{\mathcal{B}}_j(y)$
1247 as follows:

- 1248 1. $\mathcal{B}_\phi \equiv \mathcal{B}(g, y) \equiv \mathcal{B}_j(g) \equiv \mathcal{B}_j(y) := \{g_1 = g, y_0 = y\}$;
- 1249 2. $\tilde{\mathcal{B}}_j(g) := \{g_1 = g, y_0 \neq y\}$, the event that g is the incorrect feature for predicting j -th
1250 output;
- 1251 3. $\tilde{\mathcal{B}}_j(y) := \{g_1 \neq g, y_0 = y\}$, the event that y is the incorrect feature for predicting j -th
1252 output;
- 1253 4. $\tilde{\mathcal{B}}_\phi := \tilde{\mathcal{B}}_j(g) \cup \tilde{\mathcal{B}}_j(y)$, the event that $\phi = (g, y)$ is the incorrect feature for predicting j -th
1254 output.

1255 We then define the notion of the gradient conditions, which will be used in the proof of the induction.

1256 **Definition C.7** (gradient criterion). Let $\mathbf{u} = (j, r, \phi) \in \mathcal{O}^*$ and $t \leq T_1$, we define the following two
1257 conditions:

- 1258 • Given $\delta > 0$, the positive gradient criterion $\mathcal{K}_{\text{pos}}(\mathbf{u}, \delta)$:

$$\mathcal{K}_{\text{pos}}(\mathbf{u}, \delta) \text{ is true} \iff \left| \mathbb{E}[(1 - \text{logit}_{5,j}) \cdot \text{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\mathcal{B}_\phi} \mathbb{1}_{F_{5,j} \leq B}] \right| \leq \delta \quad (12)$$

- 1259 • Given $\delta > 0$, the negative gradient criterion $\mathcal{K}_{\text{neg}}(\mathbf{u}, \delta)$:

$$\mathcal{K}_{\text{neg}}(\mathbf{u}, \delta) \text{ is true} \iff \left| \mathbb{E}[\text{logit}_{5,j} \cdot \text{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\tilde{\mathcal{B}}_\phi} \mathbb{1}_{F_{5,j} \leq B}] \right| \leq \delta \quad (13)$$

1260 These conditions control the magnitude of the gradient of the feature \mathbf{u} at iteration t . Now we define
1261 the following intermediate time-steps:

1262 **Definition C.8** (phase decomposition). Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$, we define the following intermediate
1263 time-steps:

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

1264 Below we shall introduce some induction hypotheses that will be used in the proof. We first introduce
1265 a induction hypothesis for the pseudo weights.

1266 **Induction C.1** (induction on pseudo weight bounds). Let $j \in \tau(\mathcal{Y})$ and $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$,
1267 $\phi = (g, y) \in \mathfrak{F}_j$. Let $\mathbf{u}' \prec \mathbf{u} \in \mathcal{U}^*$ be the immediate predecessor of \mathbf{u} in \mathcal{U}^* . Then at $t = T_{\mathbf{u}',2}$, it
1268 holds that the pseudo weights $\tilde{\mathbf{W}}_{5,j,r}^{(t)}(\mathbf{u})$ defined in Definition C.5 satisfies

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

1269 We maintain the following induction hypotheses for the case of Assumption 4.1.

1270 **Induction C.2** (induction on weight bounds). Assuming Assumption 4.1, for $t \leq T_1$, the following
1271 properties hold:

$$1272 \quad (A) \quad |\langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v \rangle - \langle \mathbf{W}_{5,j,r,p}^{(0)}, e_v \rangle| \leq \tilde{O}(1/d) \text{ for all } v \in \mathcal{X} \text{ and } p \in \{1, 3, 4\};$$

1273 (B) $|\langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v \rangle - \langle \mathbf{W}_{5,j,r,p}^{(0)}, e_v \rangle| \leq \tilde{O}(\sigma_0/d)$ for all $v \in \mathcal{V}$ and $p \in [5], r \in [m]$ if $j \notin \tau(\mathcal{Y})$;

1274 The last induction hypothesis is about the checkpoints defined in Definition C.8.

1275 **Induction C.3** (induction on cyclic group actions). *Under Assumption C.1, for $t \leq T_1$, in addition to*
 1276 *the induction hypotheses in Induction C.1 and C.2, the following properties hold:*

1277 (A) For any $\mathbf{u}' \prec \mathbf{u} \in \mathcal{U}^*$, it holds that $T_{\mathbf{u},1a} \geq T_{\mathbf{u}',2} + \tilde{\Omega}(1/\eta\sigma_0^{q-2})$;

1278 (B) Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$, 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

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

1279 C.2 Technical Lemmas

1280 Here is a simple fact for the logits of 5-th token predictors. The proofs are trivial.

1281 **Fact C.1** (non-updating weights). For $i = 5, j \in [d], r \in [m]$ and $p \in [5]$, the following components
 1282 of \mathbf{W}_5 , i.e., $\langle \mathbf{W}_{i,j,r,p}, e_v \rangle$ would not be updated:

- 1283 • $j \in [d], p \in \{1, 3, 4\}, v \notin \mathcal{X}$;
- 1284 • $j \in [d], p = 2, v \notin \mathcal{G}$;
- 1285 • $j \in [d], p \in [5], v \notin \mathcal{Y}$.

1286 We also have some unconditional bounds on logits.

1287 **Fact C.2** (unconditional logit bounds). Due to Assumption A.1, for the logits of 5-th token predictors,
 1288 we have the following simple facts:

1289 (a) Let $j \in [d]$ and $i \in [5]$, and suppose $F_{i,j}(\cdot) \leq B$, then

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

1290 (b) Let $j \in \tau(\mathcal{Y})$ and input $\mathbf{Z} \in \text{supp}(\mathcal{D}^1(\mathcal{Z}))$. Suppose there are no more than n coordinates in
 1291 $[d]$ such that $F_{5,j}(\mathbf{Z}) \geq \frac{1}{\log d}$, then as long as $e^B \gg d$, it holds that

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

1292 This is the logit lower bound for the prediction of the j -th head of the language model.

1293 **Lemma C.1** (gradient bounds). Let $j \in \tau(\mathcal{Y})$, and $\phi = (g, y) \in \mathfrak{F}$. Suppose Induction C.3 is
 1294 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

1295 (a) $\mathbb{E}[(1 - \text{logit}_{5,j}^{(t)}) \mid \mathcal{B}_\phi] \geq 1 - d^{-0.49+\delta}$

1296 (b) $\mathbb{E}[\text{logit}_{5,j}^{(t)} \mid \tilde{\mathcal{B}}_\phi] \leq d^{-0.49+\delta}$

1297 (c) for any $r \in [m]$, it holds that for $(p, v) \in \{(2, g), (5, y)\}$:

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

1298 *Proof.* We prove the statements separately.

1299 • **Part (a):** By Induction C.2, we have that All the irrelevant features $|\langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v \rangle| \leq \tilde{O}(\sigma_0/d)$,
 1300 so conditioned on $\mathcal{B}_{g,y}$ we have that

$$F_j(\mathbf{Z}) \leq \sum_{r \in [m]} \text{sReLU}(\Lambda_{5,j,r}^{(t)}(\mathbf{Z})) \leq \sum_{r \in [m]} \Psi_{j,r}^{(t)}(g, y) + \tilde{O}(\sigma_0 m/d)$$

$$\begin{aligned} &\leq (0.5 + \delta) \log d + \tilde{O}(\sigma_0) \\ &\leq 0.501 \log d \end{aligned}$$

1301 Then since all $F_j \geq 0$, we have that conditioned on $\mathcal{B}_{g,y}$, we have that

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

1302 which concludes the proof of (a). (b) can also be similarly proved.

1303 • **Part (b):** Firstly when Induction C.2 holds, for any $\phi = (g, y) \in \mathfrak{F}_j$, we have for all $\phi' = (g', y')$
1304 such that exactly one of $g' = g$ or $y' = y$ is satisfied, it holds

$$\Psi_{j,r}(\phi') \leq \Psi_{j,r}(\phi) + \tilde{O}(\sigma_0)$$

1305 Therefore by taking a sum over $r \in [m]$ it holds that

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

1306 So by (a), we have the logit upper bound for all $\phi' = (g', y')$ such that it shares a compo-
1307 nent with ϕ as $\mathbb{E}[\text{logit}_{5,j} \mid \mathcal{B}_{\phi'}] \leq d^{-0.49+\delta}$. Note that since $\tilde{\mathcal{B}}_{\phi} = (\bigcup_{(g',y), g' \neq g} \mathcal{B}_{g',y}) \cup$
1308 $(\bigcup_{(g,y'), y' \neq y} \mathcal{B}_{g,y'})$, we can obtain the desired result.

1309 • **Part (c):** By combining (a) and Induction C.3, we can compute

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

1310 where the last inequality is because $\Lambda_{5,j,r}^{(t)}$ conditioned on $\tilde{\mathcal{B}}_{\phi}$ is smaller than $\tilde{O}(1) \mathbb{E}[\Lambda_{5,j,r}^{(t)} \mid \mathcal{B}_{\phi}]$
1311 from Induction C.3, and the application of (a). Note that we ignore the event $\{F_{5,j} \leq B\}$ because
1312 it always happens when $\sum_r \Psi_{j,r}(\phi) \leq \log d$ (coupled with our induction about irrelevant
1313 features).

1314 Now we have finished all proofs. □

1315 **Lemma C.2** (initialization gap between features). *Assuming Assumption 4.1. Let $j \in \tau(\mathcal{Y})$, for all*
1316 *$r \in [m]$ and any two $\mathbf{u}, \mathbf{u}' \in \mathcal{U}$, we have with prob $\geq 1 - o(1)$ over the randomness at initialization*
1317 *that*

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

1318 *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
1319 $\Psi_{\mathbf{u}}$ of different $\mathbf{u} \in \mathcal{U}$ are independent and identically distributed on the randomness of $\mathbf{W}^{(0)}$, due
1320 to the orthogonality of embeddings $e_v, v \in \mathcal{V}$. Then, by the basic property of a Gaussian variable
1321 (notice that $\Psi_{\mathbf{u}}^{(0)} - \Psi_{\mathbf{u}'}^{(0)}$ is also Gaussian with variance $2\sigma_0$) (all though different pairs could be
1322 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
1323 pair. Then by a union bound over $O(m^2 n_y^4)$ -many all possible pairs we can conclude the proof. □

1324 **Lemma C.3** (consistency of gradients). *Let $\phi = (g, y) \in \mathfrak{F}$, for any $j \in [d]$ and $r \in [m]$, we have*
1325 *with probability $1 - \frac{1}{d^{\Omega(\log d)}}$ over $\mathbf{W}_{5,j,r}^{(0)}$ that:*

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

1326 *Proof.* We can view the expectation

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

1327 at iteration $t = 0$ as

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

1328 where

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

1329 Now we can apply Lemma E.1 by viewing the RHS as a U-statistics where the randomness of each
 1330 $h(v_1, v_2)$ 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
 1331 are identically distributed and jointly independent with any $h(v'_1, v'_2)$ if the sets $\{v_1, v_2\}$ is disjoint
 1332 with the set $\{v'_1, v'_2\}$. So by choosing $n = |\mathcal{X}| = \Theta(d)$ and $m = 2$, $M = (\tilde{O}(\Psi_{j,r}^{(0)}(\phi))^{q-1} \log d)^{q-1}$
 1333 when $\Psi_{j,r}^{(0)}(\phi) \leq d^{-0.01}$ and $M = \tilde{O}(\Psi_{j,r}^{(0)}(\phi) \log d)^{q-1}$ when $\Psi_{j,r}^{(0)}(\phi) > d^{-0.01}$ and corresponding
 1334 $t = nM \log d$ in Lemma E.1 we have the desired result. \square

1335 **Fact C.3** (conditional expectation of logit). Let $\phi = (g, y) \in \mathfrak{F}_j$, then we have the decompositions:
 1336 For the negative gradient $\mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbb{1}_{\tilde{\mathcal{B}}_\phi}]$, we have

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

1337 **Lemma C.4** (irrelevant features). Suppose Induction C.3 holds for all $t < T_1$, then Induction C.2a
 1338 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
 1339 following holds:

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

1340 *Proof.* We shall be proving that the total feature growth of any $x \in \mathcal{X}$ at the end of training
 1341 should be negligible, that is, $\langle \mathbf{W}_{5,j,r,p}^{(t)}, e_x \rangle - \langle \mathbf{W}_{5,j,r,p}^{(0)}, e_x \rangle \leq \tilde{O}(\frac{1}{d}) \ll \sigma_0$. In fact, suppose that
 1342 something weaker happened before t , for example $\langle \mathbf{W}_{5,j,r,p}^{(t)}, e_x \rangle - \langle \mathbf{W}_{5,j,r,p}^{(0)}, e_x \rangle \leq \sigma_0/d^{0.1}$, then
 1343 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
 1344 $\langle \mathbf{W}_{5,j,r,p}, e_x \rangle$ is:

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

1345 By Induction C.3, we know that there is at most one feature $(g, y) \in \mathfrak{F}_j$ such that $\Psi_{j,r}^{(t)}(g, y) > d^{0.1} \sigma_0$,
 1346 so we can decompose the gradient of e_x at any iteration $s \leq t$ to:

$$\begin{aligned} &\langle \nabla_{\mathbf{W}_{5,j,r,p}} \text{Loss}^{(s)}, e_x \rangle \tag{14} \\ &= \mathbb{E}[\mathbb{1}_{\mathcal{A}_x} (1 - \mathbf{logit}_{5,j}^{(s)}) \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j} \mathbb{1}_{F_{5,j} \leq B} - \mathbf{logit}_{5,j}^{(s)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\tilde{\mathcal{B}}_j} \mathbb{1}_{F_{5,j} \leq B}] \tag{15} \\ &= \mathbb{E}[(1 - \mathbf{logit}_{5,j}^{(s)}) \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(g)} - \mathbf{logit}_{5,j}^{(s)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\tilde{\mathcal{B}}_j(g)} \mathbb{1}_{F_{5,j} \leq B} \mathbb{1}_{\mathcal{A}_x}] \\ &\quad + \Pr(\mathcal{A}_x) \sum_{g' \in \mathcal{G}, g' \neq g} \mathbb{E}[(1 - \mathbf{logit}_{5,j}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(g')} \mathbb{1}_{F_{5,j} \leq B} \mid \mathcal{A}_x] \\ &\quad + \Pr(\mathcal{A}_x) \sum_{g' \in \mathcal{G}, g' \neq g} \mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mathbb{1}_{\tilde{\mathcal{B}}_{g'}} \mathbb{1}_{F_{5,j} \leq B} \mid \mathcal{A}_x] \end{aligned}$$

1347 Now since $\langle \mathbf{W}_{5,j,r,p}^{(s)}, e_{x'} \rangle \leq \sigma_0 \log d, \forall x' \in \mathcal{X}$, it holds that the sum of all updates of $\langle \mathbf{W}_{5,j,r,p}, e_x \rangle$
 1348 before t is bounded by

$$\begin{aligned} & \sum_{s \leq t} O\left(\frac{\eta}{d}\right) \mathbb{E}[(1 - \mathbf{logit}_{5,j}^{(s)}) \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(g,y)} \mathbb{1}_{F_{5,j} \leq B} - \mathbf{logit}_{5,j}^{(t)} \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\tilde{\mathcal{B}}_j(g)} \mathbb{1}_{F_{5,j} \leq B} \mid \mathcal{A}_x] \\ & + \sum_{s \leq t} \sum_{g' \neq g, g' \in \mathcal{G}} O\left(\frac{\eta}{d}\right) \mathbb{E}[(1 - \mathbf{logit}_{5,j}^{(s)}(F)) \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(g')} \mathbb{1}_{F_{5,j} \leq B} \mid \mathcal{A}_x] \\ & + \sum_{s \leq t} \sum_{g' \neq g, g' \in \mathcal{G}} O\left(\frac{\eta}{d}\right) \mathbb{E}[\mathbf{logit}_{5,j}^{(s)} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\tilde{\mathcal{B}}_j(g')} \mathbb{1}_{F_{5,j} \leq B} \mid \mathcal{A}_x] \\ & =: I_1 + I_2 + I_3 \end{aligned}$$

1349 Since by Induction C.3, we can bound $I_2 + I_3$ by

$$\begin{aligned} |I_2 + I_3| & \leq \sum_{g' \neq g, g' \in \mathcal{G}} O\left(\frac{1}{d}\right) \left| \sum_{s \leq t} \eta \langle \nabla_{\mathbf{W}_{5,j,r,2}} \text{Loss}^{(s)}, e_{g'} \rangle \right| \\ & \lesssim \frac{n}{d} \times \tilde{O}(\sigma_0) \ll \sigma_0 / d^{0.99} \end{aligned}$$

1350 and for I_1 , we also have

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

1351 Thus the total update to $\langle \mathbf{W}_{5,j,r,p}, e_x \rangle$ for all iterations before t is bounded by $\tilde{O}(\frac{1}{d})$ which proves
 1352 the induction hypothesis and also the final desired result. The second statement can be obtained by
 1353 looking at (15). \square

1354 C.3 Phase I: Initial Growth

1355 First suppose $\mathbf{u} = (j, r, (g, y))$, where j denotes the token index and r is the neuron index. In phase
 1356 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
 1357 while competing with other features. In phase I.2, we show that the feature \mathbf{u} grows faster than other
 1358 features and reach a certain magnitude.

1359 C.3.1 Phase I.a: Emergence of the Feature

1360 Before the feature \mathbf{u} starts to grow, we know that the gradient for it could change, due to the
 1361 confounding feature which affects the logits of the gradient. When the feature $\Psi_{\tilde{\mathbf{u}}}$ grows, the
 1362 erroneous prediction probability $\mathbf{logit}_{5,\tilde{j}}$ could rise and therefore decrease the gradient of \mathbf{u} . We
 1363 shall show below that this two effects will not affect the growth of \mathbf{u} much, and thus we can safely
 1364 ignore them.

1365 In fact, we show the following bound on the "optimistic growth" by pseudo weights in Definition C.5
 1366 almost match the actual growth by the true weights, in phase I.a.

$$|\langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t)}, e_v \rangle - \langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v \rangle| \leq \tilde{O}(\sigma_0 / d^{0.1}) \quad (16)$$

1367 To show this, we need to bound the trajectory of $|\widetilde{\mathbf{W}}_{5,j,r,p}^{(t)} - \mathbf{W}_{5,j,r,p}^{(t)}|$ during the time peirod
 1368 $t \in [0, T_{\mathbf{u}',2}]$, which is the point where the immediate predecessor has been learned almost optimally.
 1369 We need to argue the two following conditions are satisfied:

- 1370 • \mathbf{C}_1 : The total amount of iterations where $\mathbb{E}[\mathbf{logit}_{5,j}^{(t)} \mid \mathcal{B}_{g,y}] \geq \frac{1}{d^{0.1}}$ is smaller than $O(\frac{d^{0.2}}{\eta})$
 1371 (here the $d^{0.2}$ can be losen to almost $1/\sigma_0^{q-2}$).
- 1372 • \mathbf{C}_2 : Throughout $t \in [0, T_{\mathbf{u}',2}]$, we have $\langle \widetilde{\mathbf{W}}_{5,j,r,p}^{(t)}, e_v \rangle \leq \tilde{O}(\sigma_0)$.

1373 So we have the following proposition for $\Psi_{\mathbf{u}}$'s feature growth during the feature learning process of
 1374 predecessor features.

1375 **Proposition C.1** (Phase I.a). Let \mathbf{u}' be the immediate predecessor of \mathbf{u} in \mathcal{U}^* , then Induction C.1 is
 1376 satisfied for all iterations $t \leq T_{\mathbf{u}',2}$. Moreover, (16) holds at $t = T_{\mathbf{u}',2}$.

1377 To prove this proposition, we first prove the following lemma.

1378 **Lemma C.5** (gradient approximation). *Let $j \in [d]$ and $r \in [m]$. Suppose at iteration $t \leq \mathcal{T}_1$ there is*
 1379 *a $\beta \in [0, O(\sigma_0)]$ such that $|\langle \mathbf{W}_{5,j,r,q}^{(t)} - \mathbf{W}_{5,j,r,q}^{(0)}, e_x \rangle| \leq \beta$ for all $q \in \{1, 3, 4\}$ and all $x \in \mathcal{X}$, then*
 1380 *if $\Lambda_{5,j,r}^{(t)} \geq 0$, we shall have*

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

1381 *Proof.* Note that by Induction C.2(b) we have that

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

1382 which allows us to bound the difference between $\Lambda_{5,j,r}^{(t)}$ and $\tilde{\Lambda}_{5,j,r}^{(t)}$ as any iteration t :

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

1383 Now we are able to bound the difference between $\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)})$ and $\mathbf{sReLU}'(\tilde{\Lambda}_{5,j,r}^{(t)})$ as follows:

1384 • When $\Lambda_{5,j,r}^{(t)}(\mathbf{Z}) \geq \frac{1}{2}\varrho$, because of the monotonically increasing slope of $\mathbf{sReLU}'(x)$ for
 1385 $x \in [0, \varrho]$ and flat slope when $x \geq \varrho$, we have that $|\mathbf{sReLU}'(x) - \mathbf{sReLU}'(x + \epsilon)| \leq \epsilon$ for
 1386 $\epsilon > 0$. Therefore

$$\left| \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}(\mathbf{Z})) - \mathbf{sReLU}'(\tilde{\Lambda}_{5,j,r}^{(t)}(\mathbf{Z})) \right| \leq O(|\Lambda_{5,j,r}^{(t)}(\mathbf{Z}) - \tilde{\Lambda}_{5,j,r}^{(t)}(\mathbf{Z})|) \leq \beta$$

1387 • When $\Lambda_{5,j,r}^{(t)}(\mathbf{Z}) \leq \frac{1}{2}\varrho$, we have that $|\mathbf{sReLU}'(x) - \mathbf{sReLU}'(x + \epsilon)| \leq O(q^2 \epsilon \cdot (|x| + \epsilon)^{q-2})$
 1388 for $x \leq \frac{1}{2}\varrho$ and $\epsilon \ll \varrho$, therefore

$$\begin{aligned} \left| \mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}(\mathbf{Z})) - \mathbf{sReLU}'(\tilde{\Lambda}_{5,j,r}^{(t)}(\mathbf{Z})) \right| &\leq O(|(\Lambda_{5,j,r}^{(t)}(\mathbf{Z}))^{q-1} - (\tilde{\Lambda}_{5,j,r}^{(t)}(\mathbf{Z}))^{q-1}|) \\ &\leq O\left(\beta \cdot (\Lambda_{5,j,r}^{(t)}(\mathbf{Z}) + \beta)^{q-2}\right) \end{aligned}$$

1389 This concludes the proof. \square

1390 *proof of Proposition C.1.* Actually, for all $\tilde{\mathbf{u}} \in \Phi$ such that $\tilde{\mathbf{u}} \prec \mathbf{u}$ and $\tilde{\mathbf{u}}_3$ share at least one feature
 1391 component g or y with \mathbf{u}_3 , we know that there time duration of $t \in [\mathcal{T}_{\tilde{\mathbf{u}},1}, \mathcal{T}_{\tilde{\mathbf{u}},2}]$ is at most $\mathcal{T}_{\tilde{\mathbf{u}},2} - \mathcal{T}_{\tilde{\mathbf{u}},1} \leq$
 1392 $\tilde{O}(\frac{d^{0.1}}{\eta})$. Therefore the total number of iterations where $\mathbb{E}[\mathbf{logit}_{5,j} | \mathcal{B}_{g,y}] \geq \frac{1}{d^{0.1}}$ before $t = T_{\mathbf{u}',2}$
 1393 is at most $nO(\frac{d^{0.1}}{\eta} \log^2 d) \leq O(\frac{d^{0.1}}{\eta} \log^4 d)$. Then for each step t , we have

1394 • When $\mathbb{E}[\mathbf{logit}_{5,j} | \mathcal{B}_{g,y}] \geq \frac{1}{d^{0.1}}$, we have that

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

1395 where the last inequality is because both $\mathbf{logit}_{5,j}^{(t)} \leq O(1)$ and $\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) = \tilde{O}(\sigma_0^{q-1})$ is
1396 much smaller conditioned on $\tilde{\mathcal{B}}_{g,y}$. In fact, by Induction C.3 at each single iteration $t \leq T_{\mathbf{u}',2}$
1397 there can only be one $\tilde{\mathbf{u}} = (\tilde{j}, \tilde{r}, \tilde{\phi}) \in \mathcal{U}^*$ such that $\mathbf{logit}_{5,j}^{(t)} \geq \frac{1}{d^{0.1}}$ conditioned on $(g_1, y_0) = \tilde{\phi}$.
1398 Therefore we achieve the same bound.

1399 • When $\mathbb{E}[\mathbf{logit}_{5,j} \mid \mathcal{B}_{g,y}] \leq \frac{1}{d^{0.1}}$, by similar calculations, we have

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

1400 Since we know that $T_{\mathbf{u}',2} \leq \tilde{O}\left(\frac{1}{\eta \sigma_0^{q-2}}\right)$, we have that the total growth of difference between
1401 $\langle \tilde{\mathbf{W}}_{5,j,r,p}^{(t)}, e_v \rangle$ and $\langle \mathbf{W}_{5,j,r,p}^{(t)}, e_v \rangle$ is bounded by

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

1402 Therefore, we have shown that the (16) holds for all $t \in [0, T_{\mathbf{u}',2}]$. This provides a good initialization
1403 for future learning. \square

1404 Now we are set to prove the proposition that concludes Phase I.a.

1405 **Proposition C.2** (Phase I.a, feature competition). Let $\mathbf{u}' \prec \mathbf{u} \in \mathcal{U}^*$, then Induction C.3 holds for all
1406 iterations $t \in [0, T_{\mathbf{u},1a}]$. Moreover, we have the following properties:

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

1407 *Proof.* By Proposition C.1, we can deduce

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

1408 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
1409 of $\mathbf{u} = (j, r, (g, y))$ in \mathcal{U}^* . At this point, by Lemma C.1, we have a basic gradient lower bound for
1410 both g and y in neuron r :

$$\langle \nabla_{\mathbf{W}_{5,j,r,p}} \text{Loss}^{(t)}, e_v \rangle \geq \frac{1}{n^2} (1 - O(n/d)) \mathbb{E}[\mathbf{sReLU}'(\Lambda_{5,j,r}^{(t)}) \mid \mathcal{B}_{g,y}] \quad \text{for } (p, v) \in \{(2, g), (5, y)\} \quad (17)$$

1411 Using these lower bound, we can prove that the feature $\Psi_{\mathbf{u}}^{(t)}$ will outgrow all other feature $\Psi_{\hat{\mathbf{u}}}^{(t)}$ where
1412 $\mathbf{u} \prec \hat{\mathbf{u}} \in \mathcal{U}^*$ with $\hat{\mathbf{u}}_1 = j$. In fact, let $\hat{\mathbf{u}} = (\hat{j}, \hat{r}, (\hat{g}, \hat{y}))$ be an index from \mathcal{U} where $(\hat{g}, \hat{y}) \in \mathfrak{F}_{\hat{j}}$. Now
1413 assume the following induction hypotheses ① and ② during $t \in [T_{\mathbf{u}',2}, T_{\mathbf{u},1a}]$:

1414 ① $\Psi_{\mathbf{u}}^{(t)} \geq \Psi_{\hat{\mathbf{u}}}^{(t)} + \Omega\left(\frac{\sigma_0}{\log^{O(1)} d}\right) \leq \Psi_{\mathbf{u}}^{(t)}$ for all $\hat{\mathbf{u}} \in \Sigma$ where $\hat{\mathbf{u}}_1 = j$.

1415 ② The condition (17) is satisfied for all $\hat{\mathbf{u}} \in \Sigma$ where $\hat{\mathbf{u}}_1 = j$.

1416 We shall prove that if both ① and ② are satisfied at $t = T_{\mathbf{u}',2}$, then they are also satisfied at each
 1417 $t \in (T_{\mathbf{u}',2}, T_{\mathbf{u},1a}]$. Let us assume they are satisfied at some iteration $t \in (T_{\mathbf{u}',2}, T_{\mathbf{u},1a}]$, we show
 1418 below it remains the case at $t + 1$.

1419 • **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}
 1420 for all iteration $t \leq T_{\mathbf{u},1a}$. Moreover, by our induction hypothesis, ① holds for all iteration s
 1421 where $T_{\mathbf{u}',2} \leq s \leq t$, i.e. $\Psi_{\hat{\mathbf{u}}}^{(s)} \leq \Psi_{\mathbf{u}}^{(s)}$ for all $s \in [T_{\mathbf{u}',2}, t]$. Therefore we can obtain that for all
 1422 $s \in [T_{\mathbf{u}',2}, t]$, the gradient of $\Psi_{\hat{\mathbf{u}}}$ satisfies

$$\langle \nabla_{\mathbf{W}_{5,j,r',p}} \text{Loss}^{(s)}, e_v \rangle \leq \frac{1}{n^2} \mathbb{E}[\text{sReLU}'(\Lambda_{5,j,\hat{r}}^{(s)} \mid \mathcal{B}_{\hat{g},\hat{y}})] \text{ for } (p, v) \in \{(2, \hat{g}), (5, \hat{y})\} \quad (18)$$

1423 By applying both (17) and (18), couple with Induction C.3, we can show that the gradient of $\Psi_{\mathbf{u}}$ is
 1424 always larger than that of $\Psi_{\hat{\mathbf{u}}}$ as long as which proves ① for $t + 1$.

1425 • **Proof of ① at $t + 1$ by ②** we have $\Psi_{\hat{\mathbf{u}}}^{(t+1)} \leq \Psi_{\mathbf{u}}^{(t+1)}$, which, combined with Induction C.3, implies
 1426 that

$$\sum_{r \in [m]} \Psi_{j,r}^{(t)}(\hat{g}, \hat{y}) \leq \tilde{O}(\sigma_0 m) \leq O(d^{0.1} \sigma_0)$$

1427 then by Lemma C.1, we have the negative gradient for $\Psi_{\mathbf{u}}^{(t+1)}$ satisfies (17), which concludes the
 1428 induction step.

1429 Moreover, from Lemma E.2, we obtain that at $t = cT_{\mathbf{u},1a}$, we have the following result: $\Psi_{\mathbf{u}}^{(t)} \geq$
 1430 $d^{0.01} \sigma_0$ while all $\Psi_{\hat{\mathbf{u}}}^{(t)} \leq \tilde{O}(\sigma_0)$ for $\hat{\mathbf{u}} \succ \mathbf{u}$ in Σ ; So the proof of Phase Ia is complete. \square

1431 C.3.2 Phase I.b: Feature Growth

1432 In this stage we prove the following result:

1433 **Proposition C.3** (Phase I, cyclic group). Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$, then Induction C.3 holds for all
 1434 $t \leq T_{\mathbf{u},1}$, and we have the following results at $t = T_{\mathbf{u},1}$:

1435 (A) $\Psi_{\mathbf{u}}^{(T_{\mathbf{u},1})} \geq \Omega(\log d)$

1436 (B) For any $\tilde{\mathbf{u}} \in \mathcal{U}$ such that $\tilde{\mathbf{u}} \succ \mathbf{u}$, we have $\Psi_{\tilde{\mathbf{u}}}^{(T_{\mathbf{u},1})} \leq \tilde{O}(\sigma_0)$.

1437 *Proof.* The same result for phase Ia is proven in Proposition C.2. We only need to prove for
 1438 the iterations $t \in [T_{\mathbf{u},1a}, T_{\mathbf{u},1}]$. Indeed, The total number of iterations in this stage is at most
 1439 $\tilde{O}(\frac{1}{\eta d^{0.01} \sigma_0^{q-2}})$. By Lemma C.1 and Lemma C.1, we know that as long as $\Psi_{\mathbf{u}}^{(t)} \leq 0.5 \log d$, it holds
 1440 that

$$\nabla \psi_{j,r}^{(t)}(v) \geq \frac{1}{n_y^2} (1 - O(\frac{n_y}{d^{0.49}})) \mathbb{E}[\text{sReLU}'(\Lambda_{5,j,r}^{(t)} \mid \mathcal{B}_{g,y})], \quad v \in \{g, y\} \quad (19)$$

1441 This ensured that we can use Corollary E.1 for $\tilde{O}(\frac{1}{\eta d^{0.01} \sigma_0^{q-2}})$ many iterations starting from $T_{\mathbf{u},1a}$ to
 1442 reach $T_{\mathbf{u},1}$, where $\Psi_{\mathbf{u}}^{(t)} \geq \varrho/2$. Moreover, since $\text{sReLU}'(x)$ is a smooth polynomial for $x \in [0, \varrho]$
 1443 and a constant for $x \geq \varrho$. The same lower bound (19) can be applied and used to calculate the
 1444 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
 1445 $\Psi_{\mathbf{u}}^{(t)} \leq 0.5 \log d$. \square

1446 C.4 Phase II: Cancellation and Convergence

1447 In this stage, we shall show that the incorrect feature combination will move close to zero when the
 1448 FFN weights reach a certain level of convergence.

1449 **Lemma C.6** (activeness of a neuron). Let $\mathbf{u} = (j, r, (g, y)) \in \mathcal{U}^*$. If Induction C.3 holds, then for
 1450 all $t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$, it holds that $\text{sReLU}'(\Lambda_{5,j,r}^{(t)}) = 1$ conditioned on $\mathcal{B}_{g,y}$.

1451 *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
 1452 $\Psi_{\mathbf{u}}^{(t)}$ falls slightly below $\frac{1}{2} \log d$ (which have to happen before it reaches even lower), by Lemma C.1
 1453 and Lemma C.1 we have that

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

1454 Therefore $\Psi_{\mathbf{u}}^{(t)}$ is increasing once it surpass a certain threshold, and that $\Psi_{\mathbf{u}}^{(t)} \geq 0.49 \log d$ for all
 1455 $t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$ and therefore the neuron r is active. \square

1456 **Lemma C.7** (bounds on same feature in different neurons). *Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}$, where $\phi = (g, y)$,
 1457 and $\mathbf{u}' = (j, r', \phi) \in \mathcal{U}$ for some $r' \neq r$ such that $\mathbf{u}' \succ \mathbf{u}$, we have that $\Psi_{\mathbf{u}'}^{(t)} \leq \tilde{O}(\sigma_0)$ for all
 1458 $t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$.*

1459 *Proof.* We shall prove this by induction. It is true at $t = T_{\mathbf{u},1}$ by Proposition C.3. Since $T_{\mathbf{u},2} - T_{\mathbf{u},1} =$
 1460 $O(\frac{1}{\eta d^{0.001} \sigma_0^{q-2}})$, we can simply bound the total growth of $\Psi_{\mathbf{u}'}^{(t)}$ as follows: let $v = y$ or g , we have
 1461 that

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

1462 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.

1463 Therefore we have that $\psi_{j,r'}^{(t)} \leq \tilde{O}(\sigma_0)$ for all $t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$. This concludes the induction. \square

1464 **Lemma C.8** (logits in phase II). *Consider a $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$, let $\mathfrak{F}^{(t)}$ contains all the feature
 1465 combinations from \mathfrak{F} that are learned at iteration t , that is, for each $\phi' \in \mathfrak{F}^{(t)}$, there exists a $\mathbf{u}' \in \mathcal{U}^*$
 1466 such that $\mathbf{u}' \preceq \mathbf{u}$, which also means $t \geq T_{\mathbf{u}',2}$. Then for any $\phi' \in \mathfrak{F}^{(t)} \cap \mathfrak{F}_{\text{conf}}(\phi)$, we have that for
 1467 all $t \geq T_{\mathbf{u},2}$,*

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

1468 *otherwise, we have that $\mathbf{logit}_{5,j}^{(t)} \mathbb{1}_{\mathcal{B}_{\phi'}} = \Theta(\exp(\Psi_{j,r}^{(t)}(\phi') / d))$.*

1469 *Proof.* It is direct to verify the above claim following Induction C.3(b) since there is only one neuron
 1470 that learned ϕ' and it would be the only active neuron besides the current learning neuron. \square

1471 The above lemma also classified the different $\mathbf{logit}_{5,j}^{(t)}$ into two categories: one is the suppressed
 1472 logits which are small and proportional to λ/d , and the other is the unsuppressed logits which are
 1473 proportional to $\frac{1}{d}$. This leads to the following lemma:

1474 **Lemma C.9** (different cancellation conditions). *Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$, let $\mathfrak{F}^{(t)}$ contains all the
 1475 feature combinations from \mathfrak{F} that are learned at iteration t as in Lemma C.8.*

- 1476 • If $\phi' \in \mathfrak{F}_j \cap \mathfrak{F}^{(t)} \cap \mathfrak{F}_{\text{conf}}(\phi)$, we have that before $\Psi_{\mathbf{u}}^{(t)}$ exceeds $B/2$, $\Psi_{j,r}^{(t)}(\phi') \geq \Omega(\log d)$.
- 1477 • Otherwise, we have that $\Psi_{j,r}^{(t)}(\phi') \leq \frac{1}{d^{\Omega(1)}}$ once $\Psi_{\mathbf{u}}^{(t)} \geq 2.01 \log d$.

1478 The above lemma showed how different logits interact with the feature cancellation condition. Now
 1479 we are ready to prove the following lemma:

1480 **Lemma C.10** (convergence of positive gradient). *Let $j \in \tau(\mathcal{Y})$ and $\phi = (g, y) \in \mathfrak{F}_j$. For any level
 1481 $\delta > d \sigma_0^{q-2}$, the total number of iterations of which the condition $\mathcal{K}_{\text{pos}}(\phi, \delta)$ does not hold must be
 1482 smaller than $O(\frac{n_y^3 \log^2 d}{\eta \delta})$.*

1483 *Proof.* This is due to the upper bound assumption we put on $F_{5,j}$. In fact, suppose conversely that
 1484 the pair $(g, y) \in \mathcal{F}_j$ satisfies

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

1485 for more than $O(\frac{n_y^3 \log^2 d}{\delta})$ many iterations. We compute the update by

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

1486 Let $g' \in \mathcal{G} \setminus \{g\}$ be such that $(g', y) \notin \mathfrak{F}_j$, we have that

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

1487 By a telescoping sum, we have that

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

1488 Due to the contradiction assumption, we know that for $T = \omega(\frac{n_y^3 \log^2 d}{\delta})$ many iterations, we have that

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

1489 which is impossible because $\psi_{j,r}^{(t)}(y)$ and $\psi_{j,r}^{(t)}(g)$ is both absolutely bounded by $B + O(1)$. Therefore
 1490 there is a contradiction, $\mathcal{K}_{\text{pos}}(\phi, \delta)$ does not hold for at most $O(\frac{n_y^3 \log^2 d}{\delta})$ many iterations. This
 1491 concludes the proof. \square

1492 The above proof also produced a corollary.

1493 **Corollary C.1** (monotonicity of cancellations). *Let $j \in \tau(\mathcal{Y})$ and $\phi = (g, y) \in \mathfrak{F}_j$. The following*
 1494 *quantity is non-decreasing:*

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

1495 for all $t \geq T_{\mathbf{u},1}$.

1496 Using the above corollary, we can now prove the following lemma:

1497 **Lemma C.11** (convergence of negative gradient). *Let $j \in \tau(\mathcal{Y})$ and $\phi = (g, y) \in \mathfrak{F}_j$. We have that*

1498 (a) *For all iterations $t \geq T_{\mathbf{u},1}$, it holds that $\Lambda_{5,j,r}^{(t)}(\mathbf{Z}) \mathbb{1}_{\mathcal{B}_{\tilde{\phi}}} \in (-\frac{\rho}{2}, \frac{\rho}{2})$ for any $\tilde{\phi} \in \mathfrak{F}_{\text{conf}}(\phi)$.*

1499 (b) *For any level $\delta \in (\frac{\lambda}{d^{1.1}}, (d^{0.01} \sigma_0)^{q-2})$, there exists an iteration $t \geq T_{\mathbf{u},1} + O(\frac{n_y^3 \log^2 d}{\delta})$ such*
 1500 *that $\mathcal{K}_{\text{neg}}(\phi, \delta)$ holds.*

1501 *Proof.* We prove this by contradiction. By Corollary C.1, we know that the difference $\psi_{j,r}^{(t)}(g) -$
 1502 $\sum_{y' \in \mathcal{Y} \setminus \{y\}} \psi_{j,r}^{(t)}(y')$ is non-decreasing with a growth speed

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

1503 Since by Lemma C.10, we know that the iterations of which the condition $\mathcal{K}_{\text{pos}}(\phi, \delta)$ does not hold
 1504 must be smaller than $O(\frac{n_y^3 \log^2 d}{\eta \delta})$, we have that for some $\delta < \frac{\lambda}{d^{1.1}}$ and $t \geq T_{\mathbf{u},1} + \Omega(\frac{n_y^3 \log^2 d}{\eta \delta})$, it
 1505 must hold that $\Psi_{\mathbf{u}}^{(t)} \geq B + o(1)$. However, for any $\delta < \frac{\lambda}{d}$, if $\psi_{j,r}^{(t)}(g) + \psi_{j,r}^{(t)}(y') \leq -\Omega(\frac{\varrho}{n_y \log d})$ for
 1506 some $y' \neq y$, we have that

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

1507 such that the gradient of $\psi_{j,r}^{(t)}(y')$ would greatly exceed the average growth speed of $\psi_{j,r}^{(t)}(g) -$
 1508 $\sum_{y' \in \mathcal{Y} \setminus \{y\}} \psi_{j,r}^{(t)}(y')$. This would result in a fast increase of $\psi_{j,r}^{(t)}(y')$ to $-\psi_{j,r}^{(t)}(g)$. Due to this effect,
 1509 we know that $\psi_{j,r}^{(t)}(y') + \psi_{j,r}^{(t)}(g) \geq \varrho/2$ for any $y' \neq y$ and $t \geq T_{\mathbf{u},1} + \Omega(\frac{n_y^3 \log^2 d}{\eta \delta})$.

1510 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 $\mathcal{K}_{\text{neg}}(\phi, \delta)$ does not hold.
 1511 Suppose this is not the case, by the continuity of the gradient term

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

1512 we know that it must either be larger than δ or smaller than $-\delta$. Suppose the former is the case,
 1513 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
 1514 iterations, which is impossible as that would make $\Psi_{j,r}^{(t)}(\phi) \leq O(1)$ after sufficient iterations. A
 1515 similar argument can be applied to the case where the gradient is always negative. Therefore
 1516 there must exist a iteration $t \geq T_{\mathbf{u},1} + \Omega(\frac{n_y^3 \log^3 d}{\eta \delta})$ such that $\mathcal{K}_{\text{neg}}(\phi, \delta)$ holds. This concludes the
 1517 proof. \square

1518 **Corollary C.2** (incorrect feature cancellation). *Let $j \in \tau(\mathcal{Y})$ and $\phi = (g, y) \in \mathfrak{F}_j$ and $\delta =$*
 1519 *$(d^{0.01} \sigma_0)^{q-2}$, 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*

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

1520 *Proof.* By choosing $\delta = d^{0.01} \sigma_0^{q-2}$ in Lemma C.11, we have that there exists an iteration $t \geq$
 1521 $T_{\mathbf{u},1} + \Omega(\frac{n_y^3 \log^2 d}{\delta})$ such that

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

1522 accounting for the fact that there is an absolute logit lower bound $\text{logit}_{5,j}^{(t)} \geq \Omega(\frac{\lambda}{n_y d})$, we have that

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

1523 By looking into the features in $\Lambda_{5,j,r}^{(t)} \mathbb{1}_{\tilde{\mathcal{B}}_\phi}$ we can conclude that

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

1524 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
 1525 $g' \neq g, y' \neq y$ contains only the terms with indicator function $\mathbb{1}_{\tilde{\mathcal{B}}_\phi}$. Moreover, at this point
 1526 $\Psi_{\mathbf{u}}^{(t)} \geq B - \tilde{O}(\sigma_0)$ with gradient $\Omega(\lambda/n_y^2)$ as long as it's all activated, thus the gradient of $\psi_{j,r}^{(t)}(g')$
 1527 and $\psi_{j,r}^{(t)}(y')$ cannot move beyond $(d^{0.01} \sigma_0)^{q-2}$ in the following iterations. This concludes the
 1528 proof. \square

1529 Combining Lemma C.10 and Corollary C.2, we can now prove the following proposition:

1530 **Proposition C.4.** Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$, then the following holds:

- 1531 • At $t \geq T_{\mathbf{u},2} = T_{\mathbf{u},1} + O(\frac{1}{\eta d^{0.01} \sigma_0^{q-2}})$, we have $\Psi_{\mathbf{u}}^{(t)} \geq B - O(d^{0.01} \sigma_0)$.
- 1532 • For any $\phi' \in \mathfrak{F}_{\text{conf}}(\phi)$, we have $\Psi_{j,r}^{(t)}(\phi') \leq O((d^{0.01} \sigma_0)^{q-2} d / \lambda)^{1/(q-1)}$ if $t \geq T_{\mathbf{u},2}$.

1533 This proved Induction C.3 for $t \in [T_{\mathbf{u},1}, T_1]$.

1534 C.4.1 Feature Shape at Convergence

1535 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
1536 of training.

1537 **Lemma C.12** (symmetry of ϕ in the end). *Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$ where $\phi = (g, y) \in \mathfrak{F}_j$. Suppose*
1538 *$\Psi_{\mathbf{u}}^{(t)} \geq B - O((d^{0.01}\sigma_0))$ and $\max_{\phi' \in \mathfrak{F}_{\text{conf}}(\phi)} |\Psi_{j,r}^{(t)}(\phi')| \leq O(\delta)$ for all $t \geq T_{\mathbf{u},2}$, then we have*

$$\mathcal{J}_{\phi}^{(t)} := |\psi_{j,r}^{(t)}(g) - \psi_{j,r}^{(t)}(y)| \leq O(\delta) \quad (20)$$

1539 *Moreover, plugging in $\delta = O((d^{0.01}\sigma_0)^{q-2}d/\lambda)^{1/(q-1)}$ as in Proposition C.4, we have that $\mathcal{J}_{\phi}^{(t)} \leq$*
1540 *$O((d^{0.01}\sigma_0)^{q-2}d/\lambda)^{1/(q-1)}$.*

1541 *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
1542 $\psi_{j,r}^{(t)}(y) = \langle \mathbf{W}_{5,j,r,5}^{(t)}, e_y \rangle$ for every $g \in \mathcal{G}$ and $y \in \mathcal{Y}$ over the during the period $s \in [0, t]$. First for
1543 any $\phi = (g, y) \in \mathfrak{F}_j$, we have

$$\begin{aligned} & \psi_{j,r}^{(t)}(g) - \psi_{j,r}^{(0)}(g) \\ &= \sum_{s=0}^{t-1} \eta \langle \nabla \mathbf{W}_{5,j,r,2} \text{Loss}^{(s)}, e_g \rangle \\ &= \sum_{s=0}^{t-1} \eta \left(\mathbb{E}[(1 - \text{logit}_{5,j}^{(s)}) \text{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(g)} \mathbb{1}_{F_{5,j}^{(s)} \leq B}] - \mathbb{E}[\text{logit}_{5,j}^{(s)} \text{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\tilde{\mathcal{B}}_j(g)} \mathbb{1}_{F_{5,j}^{(s)} \leq B}] \right) \\ &= \sum_{s=0}^{t-1} \eta \left(\mathbb{E}[(1 - \text{logit}_{5,j}^{(s)}) \cdot \text{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(g)} \mathbb{1}_{F_{5,j}^{(s)} \leq B}] \right. \\ & \quad \left. - \sum_{s=0}^{t-1} \sum_{y' \neq y} \eta \mathbb{E}[\text{logit}_{5,j}^{(s)} \cdot \text{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}(g,y')} \mathbb{1}_{F_{5,j}^{(s)} \leq B}] \right) \\ &= U_{g,y} - \sum_{y' \neq y} R_{g,y'} \end{aligned} \quad (21)$$

1544 where the terms $U_{g,y}$ and $R_{g,y'}$ are defined as follows:

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

1545 Similarly, the total growth of $\psi_{j,r}^{(t)}(y) = \langle \mathbf{W}_{5,j,r,5}^{(t)}, e_y \rangle$ is given by

$$\begin{aligned} & \psi_{j,r}^{(t)}(y) - \psi_{j,r}^{(0)}(y) \\ &= \sum_{s=0}^{t-1} \eta \left(\mathbb{E}[(1 - \text{logit}_{5,j}^{(s)}) \text{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(y)} \mathbb{1}_{F_{5,j}^{(s)} \leq B}] - \mathbb{E}[\text{logit}_{5,j}^{(s)} \text{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\tilde{\mathcal{B}}_j(y)} \mathbb{1}_{F_{5,j}^{(s)} \leq B}] \right) \\ &= \sum_{s=0}^{t-1} \eta \left(\mathbb{E}[(1 - \text{logit}_{5,j}^{(s)}) \cdot \text{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}_j(y)} \mathbb{1}_{F_{5,j}^{(s)} \leq B}] \right. \\ & \quad \left. - \sum_{s=0}^{t-1} \sum_{g' \neq g} \eta \mathbb{E}[\text{logit}_{5,j}^{(s)} \cdot \text{sReLU}'(\Lambda_{5,j,r}^{(s)}) \mathbb{1}_{\mathcal{B}(g',y)} \mathbb{1}_{F_{5,j}^{(s)} \leq B}] \right) \\ &= U_{g,y} - \sum_{g' \neq g} R_{g',y} \end{aligned} \quad (22)$$

1546 When $\tilde{\phi} = (\tilde{g}, \tilde{y}) \in \mathfrak{F}_j$ but $\tilde{\phi} \neq \phi$, suppose $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$, then we have $\Lambda_{5,j,r}^{(s)} \mathbf{1}_{\mathcal{B}_{\tilde{g}, \tilde{y}}} \leq \tilde{O}(\sigma_0)$
 1547 by Induction C.3 for all $s \in [0, \mathcal{T}_1]$. Therefore, we can further compute

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

1548 Similarly, we have $R_{\tilde{g}, \tilde{y}} = O(d^{0.002} \sigma_0)$ for all $t \leq \mathcal{T}_1$. Thus for any $\phi' = (g', y') \neq \phi \in \mathfrak{F}_j$, we
 1549 have

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

1550 Now combined with Equation (21) and Equation (22), we have

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

1551 However, since we assumed the condition $\mathcal{K}_{\text{neg}}(\mathbf{u}, \delta)$ holds at t , which implies that $|\psi_{j,r}^{(t)}(g) +$
 1552 $\psi_{j,r}^{(t)}(y')| \leq O(\delta)$ for any $y' \neq y$, and similarly $|\psi_{j,r}^{(t)}(y) + \psi_{j,r}^{(t)}(g')| \leq O(\delta)$. Therefore, we can infer
 1553 that

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

1554 which implies $\psi_{j,r}^{(t)}(g) = U_{g,y}/n \pm O(\delta)$, we can similarly derive that $\psi_{j,r}^{(t)}(y) = U_{g,y}/n \pm O(\delta)$.
 1555 Therefore, we have

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

1556 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
 1557 $\psi_{j,r}^{(t)}(y)$, and similarly $\psi_{j,r}^{(t)}(y') \leq -\Omega(\log d)$ for any $y' \neq y$. \square

1558 At the end of induction, we have the following theorem as the training result:

1559 **Theorem C.1** (learning cyclic group action). *At iteration $t = \mathcal{T}_1$, the following properties hold:*

1560 (A) *Optimal loss: The training loss is optimal for $i = 5$:*

$$\text{Loss}_5^1(F^{(t)}) \leq \frac{1}{\text{poly}(d)}$$

1561 (B) *Sparse activations: For $j \in \tau(\mathcal{Y})$, let $\phi = (g, y) \in \mathfrak{F}_j$, then there exists exactly one activated*
 1562 *neuron $r \in [m]$ such that when $g_1 = g, y_0 = y$ happens:*

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

1563 (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*
 1564 *the activated neuron in (B), then for any $g' \neq g \in \mathcal{G}$, and any $y' \in \mathcal{Y}$, we have*

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

1565 *for some $\delta = O((d^{0.01} \sigma_0)^{q-2} d / \lambda)^{1/(q-1)}$*

1566 *Proof.* Once all features are learned, that is, the last $\mathbf{u} \in \mathcal{U}^*$ is learned and $t = T_1$, the correct
 1567 $\mathbf{logit}_{5,j}(F^{(t)})$ 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)
 1568 is correct following Induction C.3 and (C) is proven in Proposition C.4, for every $\mathbf{u} \in \mathcal{U}^*$. \square

1569 D Learning The Group Actions: Symmetry Group

1570 Similar to the analysis of the cyclic group actions, we first need to define the features for the symmetry
 1571 group. We first introduce the same set of notations.

1572 **Notations.** Let \mathcal{D}^1 be the LEGO distribution of length 1 under the language $\mathbf{LEGO}(\mathcal{X}, \mathcal{G}, \mathcal{Y})$. We
 1573 define $\mathcal{D}_{\mathcal{X}}^1$, $\mathcal{D}_{\mathcal{G}}^1$ and $\mathcal{D}_{\mathcal{Y}}^1$ be the distribution of (x_0, x_1) , g_0 and (y_0, y_1) in \mathcal{D}^1 respectively. That is,
 1574 given a LEGO sentence

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

1575 The sampling distribution of (x_0, x_1) is $\mathcal{D}_{\mathcal{X}}^1$, and similarly for g_0 and (y_0, y_1) .

1576 We restate the assumption on the symmetry group \mathcal{G} and its action here.

1577 **Assumption D.1** (Assumption 4.2, restated). Let $\mathbf{LEGO}(\mathcal{X}, \mathcal{G}, \mathcal{Y})$ be the LEGO language. We
 1578 assume $\mathcal{Y} = \{0, 1, \dots, n_y - 1\}$ and $\mathcal{G} = \mathbf{Sym}(\mathcal{Y})$, i.e., the symemtry group of order $n_y!$. We
 1579 assume $n_y = \Theta(\frac{\log \log d}{\log \log \log d})$ and $n_y! = \Theta(\text{polylog}(d)) \gg \frac{1}{\rho}$.

1580 We also redefine the feature combinations in the symmetry group case. For symmetric group \mathcal{G} , we
 1581 shall define a new notion called *fiber* of a value.

1582 **Definition D.1** (fiber of values). Assuming the group \mathcal{G} follows Assumption D.1. For each $j \in \tau(\mathcal{Y})$
 1583 and each $y \in \mathcal{Y}$, we denote the **fiber**

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

1584 denotes all group elements g that sends y to $y' = \tau^{-1}(j)$.

1585 **Definition D.2** (feature combinations, symmetric group). Assuming the group \mathcal{G} follows Assump-
 1586 tion D.1. Let $y, y' \in \mathcal{Y}$ be a pair of values, we define the following set $G_{y \rightarrow y'}$: and we call it the set
 1587 of **transition from y to y'** . Now for each $j \in \tau(\mathcal{Y})$, let

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

1588 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**
 1589 **combinations** for predicting $y' = \tau^{-1}(j)$. Similar to Definition C.1, for any $\phi = (\text{Fiber}_{j,y}, y) \in \mathfrak{F}$,
 1590 we write

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

1591 as the set of **confounding features** for $\phi = (\text{Fiber}_{j,y}, y)$.

1592 The ψ and Ψ notations are redefined as follows.

1593 **Definition D.3** (ψ , Ψ notations, symmetry group). Using the same notation of ψ in Definition C.3.
 1594 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
 1595 follows:

$$\begin{aligned} \Psi_{\mathbf{u}} &\equiv \Psi_{j,r}(\phi) := \frac{1}{(n_y - 1)!} \sum_{g \in \text{Fiber}_{j,y}} \frac{1}{2}(\psi_{j,r}(g) + \psi_{j,r}(y)) \\ \Psi_{\mathbf{u},\max} &:= \max_{g \in \text{Fiber}_{j,y}} \frac{1}{2}(\psi_{j,r}(g) + \psi_{j,r}(y)), \quad \Psi_{\mathbf{u},\min} := \min_{g \in \text{Fiber}_{j,y}} \frac{1}{2}(\psi_{j,r}(g) + \psi_{j,r}(y)) \end{aligned}$$

1596 Here $\text{Fiber}_{j,y} = \{g \in \mathcal{G} \mid \tau(g \cdot y) = j\}$ is the fiber of y under the action of \mathcal{G} .

1597 The learning order in the symmetry group case is the same as in Definition C.4, but with $\Psi_{\mathbf{u}}$ defined
 1598 in Definition D.3, which we do not repeat here. The definition of pseudo weights remains the same.
 1599 We restate the learning order here.

1600 **Definition D.4** (learning order). The *learning order* is the ordered set \mathcal{U}^* that we obtain from the
 1601 following process: Define a total order on \mathcal{U} as follows:

$$\mathbf{u} \prec \mathbf{u}' \iff \Psi_{\mathbf{u}}^{(0)} \geq \Psi_{\mathbf{u}'}^{(0)} \quad \forall \mathbf{u}, \mathbf{u}' \in \mathcal{U} \quad (23)$$

1602 We construct the sets \mathcal{U}^* by the following procedure: initialize an empty neuron set $\mathcal{W}_{tmp}^{(0)} = \emptyset$, and
 1603 an empty feature set $\mathcal{R}_{tmp}^{(0)} = \emptyset$, and the initial index set $\mathcal{U}^{(0)} = \emptyset$. Starting from $k = 1$, we do the
 1604 following:

- 1605 (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
 1606 feature $\phi \in \mathfrak{F} \setminus \mathcal{R}_{tmp}^{(k-1)}$ and $(j, r) \in \tau(\mathcal{Y}) \times [m] \setminus \mathcal{W}_{tmp}^{(k-1)}$.
- 1607 (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 $\mathcal{U}^{(k)} \leftarrow \mathcal{U}^{(k-1)} \cup \{\mathbf{u}\}$.
- 1608 (3) Iterate the (1) and (2) steps until $k = n_y^2$, then yield $\mathcal{U}^* \equiv \mathcal{U}^{(n_y^2)}$.

1609 This process yields the ordered set \mathcal{U}^* , equipped with the total order \prec defined in (23).

1610 D.1 Induction Hypothesis and Training Phases

1611 Again we define some useful probabilistic events.

1612 **Definition D.5** (probability events of feature appearance). Let $j \in \tau(\mathcal{Y})$, $r \in [m]$ and
 1613 $\phi = (\text{Fiber}_{j,y}, y) \in \mathfrak{F}_j$ be a feature combination that predicts j -th output. Denote events
 1614 $\mathcal{B}_\phi, \mathcal{B}(g, y), \mathcal{B}_j(g), \mathcal{B}_j(y)$ and $\tilde{\mathcal{B}}_\phi, \tilde{\mathcal{B}}_j(g), \tilde{\mathcal{B}}_j(y)$ as follows:

- 1615 1. $\mathcal{B}_\phi \equiv \mathcal{B}(\text{Fiber}_{j,y}, y) \equiv \mathcal{B}_j(y) \equiv \mathcal{B}_j(\text{Fiber}_{j,y})$, which is defined as

$$\mathcal{B}_\phi := \{g_1 \in \text{Fiber}_{j,y}, y_0 = y\}$$
- 1616 2. For individual $g \in \text{Fiber}_{j,y}$, we let $\mathcal{B}_{g,y} := \{g_1 = g, y_0 = y\}$;
- 1617 3. $\tilde{\mathcal{B}}_j(g) := \{g_1 \in \text{Fiber}_{j,y}, y_0 \neq y\}$, the event that $g \in \text{Fiber}_{j,y}$ did not appear together with
 1618 y for predicting j -th output. Moreover, we define $\tilde{\mathcal{B}}_j(\text{Fiber}_{j,y}) = \bigcup_{g \in \text{Fiber}_{j,y}} \tilde{\mathcal{B}}_j(g)$;
- 1619 4. $\tilde{\mathcal{B}}_j(y) := \{g_1 \notin \text{Fiber}_{j,y}, y_0 = y\}$, the event that y did not appear together with $g \in$
 1620 $\text{Fiber}_{j,y}$ for predicting j -th output;
- 1621 5. $\tilde{\mathcal{B}}_\phi := \tilde{\mathcal{B}}_j(\text{Fiber}_{j,y}) \cup \tilde{\mathcal{B}}_j(y)$, the event that the appeared feature combination ϕ' is wrong
 1622 for predicting j -th output.

1623 Similar to above, we give several induction hypotheses, each characterize different aspects of the
 1624 process.

1625 We then define the notion of the gradient conditions similar to the cyclic group case.

1626 **Definition D.6** (gradient criterion, symmetry group). Let $\mathbf{u} = (j, r, \phi) \in \mathcal{O}^*$ and $t \leq T_1$, we define
 1627 the following two conditions:

- 1628 • Given $\delta > 0$, the positive gradient criterion $\mathcal{K}_{\text{pos}}(\mathbf{u}, \delta)$:

$$\mathcal{K}_{\text{pos}}(\mathbf{u}, \delta) \text{ is true} \iff \left| \mathbb{E}[(1 - \text{logit}_{5,j}) \cdot \text{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\mathcal{B}_\phi} \mathbb{1}_{F_{5,j} \leq B}] \right| \leq \delta \quad (24)$$

- 1629 • Given $\delta > 0$, the negative gradient criterion $\mathcal{K}_{\text{neg}}(\mathbf{u}, \delta)$:

$$\mathcal{K}_{\text{neg}}(\mathbf{u}, \delta) \text{ is true} \iff \left| \mathbb{E}[\text{logit}_{5,j} \cdot \text{sReLU}'(\Lambda_{5,j,r}) \mathbb{1}_{\tilde{\mathcal{B}}_\phi} \mathbb{1}_{F_{5,j} \leq B}] \right| \leq \delta \quad (25)$$

1630 These conditions control the magnitude of the gradient of the feature \mathbf{u} at iteration t . Now we define
 1631 the following intermediate time-steps:

1632 **Definition D.7** (phase decomposition, symmetry group). Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$, we define the
 1633 following intermediate time-steps:

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

1634 Below we shall introduce some induction hypotheses that will be used in the proof. We first introduce
 1635 a induction hypothesis for the pseudo weights.

1636 **Induction D.1** (induction on pseudo weight bounds). Let $j \in \tau(\mathcal{Y})$ and $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$. Let
 1637 $\mathbf{u}' \prec \mathbf{u} \in \mathcal{U}^*$ be the immediate predecessor of \mathbf{u} in \mathcal{U}^* . Then at $t = T_{\mathbf{u}',2}$, it holds that the pseudo
 1638 weights $\widetilde{\mathbf{W}}_{5,j,r}^{(t)}(\mathbf{u})$ defined in Definition C.5 satisfies

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

1639 where $p = 2, v \in \mathcal{G}$ or $p = 5, v \in \mathcal{Y}$.

1640 We maintain the following induction hypotheses for the case of Assumption D.1.

1641 **Induction D.2** (induction on weight bounds). Assuming Assumption 4.1, for $t \leq T_1$, the following
 1642 properties hold:

- 1643 (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 $p \in \{1, 3, 4\}$;
 1644 (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 $j \notin \tau(\mathcal{Y})$;

1645 The last induction hypothesis is about the checkpoints defined in Definition D.7.

1646 **Induction D.3** (induction on cyclic group actions). Under Assumption D.1, for $t \leq T_1$, in addition
 1647 to the induction hypotheses in Induction D.1 and D.2, the following properties hold:

- 1648 (A) For any $\mathbf{u}' \prec \mathbf{u} \in \mathcal{U}^*$, it holds that $T_{\mathbf{u},1a} \geq T_{\mathbf{u}',2} + \widetilde{\Omega}(1/\eta\sigma_0^{q-2})$;
 1649 (B) Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$, 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

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

1650 The proof will proceed in a similar way as in the cyclic group case, with some additional technical
 1651 details. We first introduce some lemmas that will be used in the proof, many of them are similar to
 1652 the ones in the cyclic group case.

1653 **Lemma D.1** (initialization gap between features). Assuming Assumption 4.1. Let $j \in \tau(\mathcal{Y})$, for all
 1654 $r \in [m]$ and any two $\mathbf{u}, \mathbf{u}' \in \mathcal{U}$, we have with prob $\geq 1 - o(1)$ over the randomness at initialization
 1655 that

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

1656 *Proof.* the proof is similar to that in Lemma C.2, one simply needs to modify the feature g that
 1657 appears in the proof. \square

1658 **Lemma D.2** (gradient bounds). Let $j \in \tau(\mathcal{Y})$, and $\phi = (\text{Fiber}_{j,y}, y) \in \mathfrak{F}$. Suppose Induction D.3 is
 1659 satisfied at $t \leq T_1$, then for any $\delta \in [0, 0.4]$, if $\sum_{r \in [m]} \Psi_{j,r,\max}^{(t)}(\phi) \leq (0.5 + \delta) \log d$, it holds that

1660 (a) $\mathbb{E}[(1 - \text{logit}_{5,j}^{(t)}) \mid \mathcal{B}_\phi] \geq 1 - d^{-0.49+\delta}$

1661 (b) $\mathbb{E}[\text{logit}_{5,j}^{(t)} \mid \widetilde{\mathcal{B}}_\phi] \leq d^{-0.49+\delta}$

1662 (c) for any $r \in [m]$, for any $g \in \text{Fiber}_{j,y}$, it holds that with $v = g$ or $v = y$,

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

1663 Also, we can control the irrelevant features in the same way as in the cyclic group case.

1664 **Lemma D.3** (irrelevant features). *Suppose Induction D.3 holds for all $t < T_1$, then Induction D.2a*
 1665 *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*
 1666 *following holds:*

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

1667 The proofs of the above lemmas are similar to that in the cyclic group case, we omit them here.

1668 D.2 Phase I: Emergence of the Feature

1669 We need to show that the feature growth in Phase I is again following the pseudo weight trajectory.
 1670 We give the same version of the proposition as in the cyclic group case.

1671 **Proposition D.1** (Phase I.a). Let \mathbf{u}' be the immediate predecessor of \mathbf{u} in \mathcal{U}^* , then Induction D.1 is
 1672 satisfied for all iterations $t \leq T_{\mathbf{u}',2}$. Moreover, (16) holds at $t = T_{\mathbf{u}',2}$ for the symmetry group case.

1673 The proof of Proposition D.1 follows the same line as in the cyclic group case, we omit it here. This
 1674 proposition guarantees that the feature growth before $T_{\mathbf{u}',2}$ is small and rather negligible.

1675 **Proposition D.2** (Phase I). Let \mathbf{u}' be the immediate predecessor of \mathbf{u} in \mathcal{U}^* , then Induction D.3 is
 1676 satisfied for all iterations $t \leq T_{\mathbf{u}',2}$.

1677 Now we analyze the feature competition in Phase Ia.

1678 **Proposition D.3** (Phase I.a, feature competition). Let $\mathbf{u}' \prec \mathbf{u} \in \mathcal{U}^*$, then Induction D.3 holds for all
 1679 iterations $t \in [0, T_{\mathbf{u},1a}]$. Moreover, we have the following properties:

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

1680 *Proof.* The proof is similar to that in the cyclic group case. The caveat here is that we need to consider
 1681 a different version of TPM by applying Lemma E.2 with Lemma D.1. In fact, since each $\text{Fiber}_{j,y}$
 1682 contains $(n_y - 1)!$ elements, we can transform the problem into the tensor power method of the
 1683 growth of feature $\psi_{j,r}(y)$, averaged over all combinations with $g \in \text{Fiber}_{j,y}$. Since the constant
 1684 is much smaller for group feature g , we only need to control the growth of the feature $\psi_{j,r}(y)$ for
 1685 comparison. \square

1686 Finally, we show the result for Phase I.

1687 **Proposition D.4** (Phase I, symmetry group). Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$, then Induction D.3 holds for all
 1688 $t \leq T_{\mathbf{u},1}$, and we have the following results at $t = T_{\mathbf{u},1}$:

1689 (A) $\Psi_{\mathbf{u},\min}^{(T_{\mathbf{u},1})}, \Psi_{\mathbf{u},\max}^{(T_{\mathbf{u},1})} \geq \Omega(\log d);$

1690 (B) For any $\tilde{\mathbf{u}} \in \mathcal{U}$ such that $\tilde{\mathbf{u}} \succ \mathbf{u}$, we have $\Psi_{\tilde{\mathbf{u}}}^{(T_{\mathbf{u},1})} \leq \tilde{O}(\sigma_0).$

1691 *Proof.* The proof is similar to that in the cyclic group case but with some caveats. We still mainly
 1692 employ Lemma E.2 along with gradient approximation in Lemma D.2. \square

1693 D.3 Phase II: Convergence of the Feature

1694 Beginning with Phase II, we need to show that the feature $\Psi_{\mathbf{u}}^{(t)}$ continues to grow despite very
 1695 complex landscape.

1696 Firstly, we have some similar results as in the cyclic group case.

1697 **Lemma D.4** (activeness of a neuron). *Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$ and $\phi = (\text{Fiber}_{j,y}, y)$. If Induction D.3*
 1698 *holds, then for all $t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$, it holds that $\text{sReLU}'(\Lambda_{5,j,r}^{(t)}) = 1$ conditioned on \mathcal{B}_ϕ .*

1699 *Proof.* The proof is similar to that in the cyclic group case but relies on Lemma D.2 in a slightly
 1700 different way. Here we need to consider the gradient lower bound when $\Psi_{\mathbf{u},\min}^{(t)} \leq \frac{1}{2} \log d$. This
 1701 already provides enough gradient to keep the neuron active for specific feature combination (g, y)
 1702 where $g \in \text{Fiber}_{j,y}$. \square

1703 also, we have the following lemma for the same feature in different neurons.

1704 **Lemma D.5** (bounds on same feature in different neurons). *Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}$, where $\phi = (g, y)$,*
 1705 *and $\mathbf{u}' = (j, r', \phi) \in \mathcal{U}$ for some $r' \neq r$ such that $\mathbf{u}' \succ \mathbf{u}$, we have that $\Psi_{\mathbf{u}'}^{(t)} \leq \tilde{O}(\sigma_0)$ for all*
 1706 *$t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$.*

1707 *Proof.* The proof is similar, only requiring some bounds on the gradient trajectory for $t \in [T_{\mathbf{u},1}, T_{\mathbf{u},2}]$.
 1708 \square

1709 Similarly, we classify the logits in phase II into two categories.

1710 **Lemma D.6** (logits in phase II). *Consider a $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$, let $\mathfrak{F}^{(t)}$ contains all the feature*
 1711 *combinations from \mathfrak{F} that are learned at iteration t , that is, for each $\phi' \in \mathfrak{F}^{(t)}$, there exists a $\mathbf{u}' \in \mathcal{U}^*$*
 1712 *such that $\mathbf{u}' \preceq \mathbf{u}$, which also means $t \geq T_{\mathbf{u}',2}$. Then for any $\phi' \in \mathfrak{F}^{(t)} \cap \mathfrak{F}_{\text{conf}}(\phi)$, we have that for*
 1713 *all $t \geq T_{\mathbf{u},2}$,*

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

1714 *otherwise, we have that $\text{logit}_{5,j}^{(t)} \mathbb{1}_{\mathcal{B}_{\phi'}} \geq \Omega(\exp(\Psi_{j,r,\min}^{(t)}(\phi') / d))$.*

1715 Now based on the classification of logits, we can derive the following lemma, which concerns with
 1716 different cancellation conditions.

1717 **Lemma D.7** (different cancellation conditions). *Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$ and $\phi = (\text{Fiber}_{j,y}, y)$, let*
 1718 *$\mathfrak{F}^{(t)}$ contains all the feature combinations from \mathfrak{F} that are learned at iteration t as in Lemma D.6.*
 1719 *Let $g \in \text{Fiber}_{j,y}$, then the following holds: Here we abuse the notation and revert to using those in*
 1720 *Definition C.1.*

- 1721 • *let $g \in \text{Fiber}_{j,y}$, 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*
 1722 *$\phi' = (g', y), g' \notin \text{Fiber}_{j,y}$, or $\phi' = (g, y')$ for any $y' \neq y$.*
- 1723 • *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 $\phi' = (g', y), g' \notin$*
 1724 *$\text{Fiber}_{j,y}$, or $\phi' = (g, y')$ for any $y' \neq y$.*

1725 *Proof.* The proof of this lemma requires slightly more delicate analysis. Suppose let $g \in \text{Fiber}_{j,y}$
 1726 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
 1727 $\psi_{j,r}^{(t)}(y')$ remains small, and therefore the feature $\psi_{j,r}^{(t)}(y')$ cannot cancel out the feature $\psi_{j,r}^{(t)}(g)$. \square

1728 Now we show how the correct features can grow to maximum value possible, as stipulated by the
 1729 upper bound assumption Assumption A.1.

1730 **Lemma D.8** (convergence of positive gradient). *Let $j \in \tau(\mathcal{Y})$ and $\phi \in \mathfrak{F}_j$. For any level $\delta > d\sigma_0^{q-2}$,*
 1731 *the total number of iterations of which the condition $\mathcal{K}_{\text{pos}}(\phi, \delta)$ does not hold must be smaller than*
 1732 *$O(\frac{n_y^3 \log^2 d}{\eta \delta})$.*

1733 *Proof.* The proof is similar to that in the cyclic group case. Instead of analyzing a single pair of g, y ,
 1734 we consider the growth of feature $\psi_{j,r}^{(t)}(g)$ and $\psi_{j,r}^{(t)}(y)$ for all $g \in \text{Fiber}_{j,y}$ and $y \in \mathcal{Y}$. By using the
 1735 same analysis as in Lemma C.10, we can show that the following difference

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

1736 is non-decreasing. Same technique can be applied another difference:

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

1737 which is also non-increasing. By computing their possible growth upper bounds, we can show that
 1738 they can grow for at most $O(\frac{n_y^3 \log^2 d}{\eta \delta})$ iterations, assuming $\mathcal{K}_{\text{pos}}(\phi, \delta)$ does not hold. \square

1739 Similarly, we have the following lemma for the negative gradient.

1740 Using the above corollary, we can now prove the following lemma:

1741 **Lemma D.9** (convergence of negative gradient). *Let $j \in \tau(\mathcal{Y})$ and $\phi \in \mathfrak{F}_j$ and $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$.
 1742 We have that*

- 1743 (a) *For all iterations $t \geq T_{\mathbf{u},1}$, it holds that $\Lambda_{5,j,r}^{(t)}(\mathbf{Z}) \mathbf{1}_{B_{\tilde{\phi}}} \in (-\frac{\rho}{2}, \frac{\rho}{2})$ for any $\tilde{\phi} \in \mathfrak{F}_{\text{conf}}(\phi)$.*
- 1744 (b) *For any level $\delta \in (\frac{\lambda}{d^{1.1}}, (d^{0.01} \sigma_0)^{q-2})$, there exists an iteration $t \geq T_{\mathbf{u},1} + O(\frac{n_y^3 \log^2 d}{\delta})$ such
 1745 that $\mathcal{K}_{\text{neg}}(\phi, \delta)$ holds.*

1746 *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
 1747 $y' \neq y$ or $g \notin \text{Fiber}_{j,y}$, their positive gradient is upper bounded by $O(\sigma_0^{q-2})$ while lower bounded
 1748 by $-\Theta(\rho)$. Any such feature will be pushed to drop below $-\rho$ after cancelling out (i.e., if δ is as
 1749 small as $o(\frac{1}{d})$ or $o(\lambda/d)$). Thus we only need to consider the negative gradient of $\psi_{j,r}^{(t)}(y')$, $y' \neq y$
 1750 and $\psi_{j,r}^{(t)}(g)$ for $g \notin \text{Fiber}_{j,y}$. Then the logic is the same as in Lemma C.10: we consider when
 1751 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
 1752 showing the contradiction that $\mathcal{K}_{\text{neg}}(\phi, \delta)$ does not hold we shall have that for every incorrect pair
 1753 (g, y') , $y' \neq y$, $g \in \text{Fiber}_{j,y}$ or (g', y) , $g' \notin \text{Fiber}_{j,y}$, the feature $\psi_{j,r}^{(t)}(g')$ or $\psi_{j,r}^{(t)}(y')$ will grow below
 1754 $-\psi_{j,r}^{(t)}(y)$ (or $\max_{g \in \text{Fiber}_{j,y}} \psi_{j,r}^{(t)}(g)$) after $O(\frac{n_y^3 \log^2 d}{\eta \delta})$ iterations, when $\delta < \frac{1}{d}$ (if the corresponding
 1755 logit is unsuppressed) or when $\delta < o(\lambda/n_y d)$. \square

1756 By inserting $\delta = O((d^{0.01} \sigma_0)^{q-2} d / \lambda)^{1/(q-1)}$ into Lemma C.11, we have the a similar corollary to
 1757 ??:

1758 Finally, we have the following proposition for the convergence of the feature.

1759 **Proposition D.5.** Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$, then the following holds:

- 1760 • At $t \geq T_{\mathbf{u},2} = T_{\mathbf{u},1} + O(\frac{1}{\eta d^{0.01} \sigma_0^{q-2}})$, we have $\Psi_{\mathbf{u},\min}^{(t)} \geq B - O(d^{0.01} \sigma_0)$.
- 1761 • For any $\phi' \in \mathfrak{F}_{\text{conf}}(\phi)$, we have both $|\Psi_{j,r,\min}^{(t)}(\phi')|$ and $|\Psi_{j,r,\max}^{(t)}(\phi')|$ bounded by
 1762 $O((d^{0.01} \sigma_0)^{q-2} d / \lambda)^{1/(q-1)}$ if $t \geq T_{\mathbf{u},2}$.

1763 D.3.1 Phase II: Shape of the Feature

1764 A similar analysis as in Lemma C.12 gives the following lemma. We do not repeat the proof here.

1765 **Lemma D.10** (symmetry of ϕ in the end). *Let $\mathbf{u} = (j, r, \phi) \in \mathcal{U}^*$ where $\phi = (\text{Fiber}_{j,y}, y) \in \mathfrak{F}_j$.
 1766 Suppose $\Psi_{\mathbf{u}}^{(t)} \geq B - O(d^{0.01} \sigma_0)$ and $\max_{\phi' \in \mathfrak{F}_{\text{conf}}(\phi)} |\Psi_{j,r}^{(t)}(\phi')| \leq O(\delta)$ for all $t \geq T_{\mathbf{u},2}$, then we
 1767 have*

- 1768 • $\psi_{j,r}^{(t)}(g) + \psi_{j,r}^{(t)}(y) \geq B - O(d^{0.01} \sigma_0)$ for all $g \in \text{Fiber}_{j,y}$.
- 1769 • $\mathcal{J}_{\phi}^{(t)} := |n_y \psi_{j,r}^{(t)}(y) - \psi_{j,r}^{(t)}(g)| \leq O(\delta)$ for all $g \in \text{Fiber}_{j,y}$ and $t \geq T_{\mathbf{u},2}$.

1770 plugging in $\delta = O((d^{0.01} \sigma_0)^{q-2} d / \lambda)^{1/(q-1)}$ as in Proposition C.4, we have that $\mathcal{J}_{\phi}^{(t)} \leq$
 1771 $O((d^{0.01} \sigma_0)^{q-2} d / \lambda)^{1/(q-1)}$.

1772 Thus we arrive at the following theorem.

1773 **Theorem D.1** (learning symetry group action). *For group structure Assumption D.1, at iteration $t =$*
 1774 *\mathcal{T}_1 , the following properties of $F^{(t)}$ hold:*

1775 (A) *Optimal loss: The training loss is optimal for $i = 5$:*

$$\text{Loss}_5^1(F^{(t)}) \leq \frac{1}{\text{poly}(d)}$$

1776 (B) *Sparse activations: For $j \in \tau(\mathcal{Y})$, let $\phi = (\text{Fiber}_{j,y}, y) \in \mathfrak{F}_j$, then there exists exactly one*
 1777 *activated neuron $r \in [m]$ such that when $g_1 \in \text{Fiber}_{j,y}$, $y_0 = y$ happens:*

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

1778 (C) *Cancellation of incorrect features: For $j \in \tau(\mathcal{Y})$, let $\phi = (\text{Fiber}_{j,y}, y) \in \mathfrak{F}_j$, and let the*
 1779 *$r \in [m]$ be the activated neuron in (B), then for any $g' \notin \text{Fiber}_{j,y}$, and any $y' \neq y \in \mathcal{Y}$, we*
 1780 *have*

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

1781 *for some $\delta = O((d^{0.01}\sigma_0)^{q-2}d/\lambda)^{1/(q-1)}$*

1782 E Auxiliary Technical Tools

1783 E.1 Probability

1784 First we need a Bernstein inequality for U-statistics

1785 **Lemma E.1** (concentration inequality for pseudo-U-statistics). *Let x_1, \dots, x_n be different symbols,*
 1786 *and let $m \ll n$ be such that $n \equiv 0 \pmod{m}$. Suppose for some function h with $|h| \leq M$ the random*
 1787 *variables $h(x_{i_1}, x_{i_2}, \dots, x_{i_m})$ and $h(x_{i'_1}, x_{i'_2}, \dots, x_{i'_m})$ are independent and identically distributed*
 1788 *as long as $\{x_{i_1}, x_{i_2}, \dots, x_{i_m}\} \cap \{x_{i'_1}, x_{i'_2}, \dots, x_{i'_m}\} = \emptyset$, then the pseudo-U-statistic*

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

1789 *satisfies $\Pr(|U_{n,m} - \mathbb{E}[U_{n,m}]| \geq t) \leq e^{-\frac{nt^2}{mM^2}}$*

1790 *Proof.* The proof is the same as in [74]. □

1791 E.2 Tensor Power Method Bounds

1792 We present two lemmas related to the tensor power method.

1793 **Lemma E.2** (TPM, adapted from [67]). *Consider an increasing sequence $x_t \geq 0$ defined by*
 1794 *$x_{t+1} = x_t + \eta C_t x_t^{q-1}$ for some integer $q \geq 3$ and $C_t = \Theta(1) > 0$, then we have for every $A > x_0$,*
 1795 *for every $\delta > 0$, and every $\eta \in (0, 1)$:*

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

1796 This lemma has a corollary:

1797 **Corollary E.1** (TPM, from [67]). *Let $q \geq 3$ be a constant and $x_0, y_0 = o(1)$ and $A = O(1)$. Let*
 1798 *$\{x_t, y_t\}_{t \geq 0}$ be two positive sequences updated as*

1799 $\bullet \quad x_{t+1} = x_t + \eta C_t x_t^{q-1} \text{ for some } C_t = \Theta(1);$

1800 • $y_{t+1} = y_t + \eta SC_t y_t^{q-1}$ for some $S = \Theta(1)$.

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

$$y_{T_x} \leq \tilde{O}(y_0).$$

1801 We also have a generalized version:

1802 **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 $\{x_t, y_t\}_{t \geq 0}$
1803 be two positive sequences updated as

1804 • $x_{t+1} = x_t + \eta C_t u_t^{q-1}$ for some $C_t = \Theta(1)$;

1805 • $y_{t+1} = y_t + \eta SC_t v_t^{q-1}$ 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

$$y_{T_x} \leq \tilde{O}(y_0).$$

1806 F Learning the Attention Layer: Cyclic Case

1807 In this section, we consider the group operations in the cyclic group. we only update \mathbf{Q} and keep
1808 \mathbf{W} fixed. We only consider the contribution of gradient from $i = 5$, i.e., the prediction of the final
1809 answer. We further assume $\mathbf{Q} = [\mathbf{Q}_{p,q}]_{p,q \in [5]} \in \mathbb{R}^{5d \times 5d}$ with $\mathbf{Q}_{p,q} \in \mathbb{R}^{d \times d}$ has the following
1810 structure:

$$\mathbf{Q}_{p,q} = \mathbf{0}_{d \times d} \text{ except for } (p, q) = (4, 3), (4, 4)$$

1811 F.1 Gradient Computations

1812 **Fact F.1** (Gradients of \mathbf{Q}). For any $p, q \in [5]$, we have

$$\begin{aligned} -\nabla_{\mathbf{Q}_{p,q}} \text{Loss}^L &= \sum_{\ell=1}^L \sum_{i \in [5]} -\nabla_{\mathbf{Q}_{p,q}} \text{Loss}_i^{L,\ell}, \quad \text{where} \\ -\nabla_{\mathbf{Q}_{p,q}} \text{Loss}_i^{L,\ell} &= \mathbb{E} \left[\sum_{\mathbf{k} \in \mathcal{I}^{(L,\ell-1)}} \text{Attn}_{\text{ans}, \ell-1 \rightarrow \mathbf{k}} \Xi_{\ell,i,\mathbf{k}}^L \cdot \left(\mathbf{Z}_{\text{ans}, \ell-1, p} \mathbf{Z}_{\mathbf{k}, q}^\top - G_{\mathbf{Q}}(\mathbf{Z}^{(L,\ell-1)})_{p,q} \right) \right] \\ &= \mathbb{E} \left[\sum_{\mathbf{k} \in \mathcal{I}^{(L,\ell-1)}} \text{Attn}_{\text{ans}, \ell-1 \rightarrow \mathbf{k}} \cdot \left(\Xi_{\ell,i,\mathbf{k}}^L - \sum_{\mathbf{k}' \in \mathcal{I}^{(L,\ell-1)}} \text{Attn}_{\text{ans}, \ell-1 \rightarrow \mathbf{k}'} \Xi_{\ell,i,\mathbf{k}'}^L \right) \mathbf{Z}_{\text{ans}, \ell-1, p} \mathbf{Z}_{\mathbf{k}, q}^\top \right] \end{aligned}$$

1813 Here,

$$\begin{aligned} G_{\mathbf{Q}}(\mathbf{Z}^{L,\ell-1}) &\triangleq \sum_{\mathbf{k} \in \mathcal{I}^{L,\ell-1}} \text{Attn}_{(\text{ans}, \ell-1) \rightarrow (\mathbf{k})} \mathbf{Z}_{\text{ans}, \ell-1} \mathbf{Z}_{\mathbf{k}}^\top \\ \Xi_{\ell,i,\mathbf{k}}^L &\triangleq \sum_{j \in [d]} \mathcal{E}_{i,j}(\mathbf{Z}^{L,\ell-1}) \sum_{r \in [m]} \text{sReLU}'(\Lambda_{i,j,r}(\mathbf{Z}^{L,\ell-1})) \langle \mathbf{W}_{i,j,r}, \mathbf{Z}_{\mathbf{k}} \rangle \end{aligned}$$

1814 In the following, we use $r_{g \cdot y}$ to denote the neuron, where $\langle W_{5,\tau(g \cdot y), r_{g \cdot y}}, e_g \rangle, \langle W_{5,\tau(g \cdot y), r_{g \cdot y}}, e_y \rangle \geq$
1815 $\Omega(\log d)$, i.e., $r_{g \cdot y}$ has been activated to predict $j = \tau(g \cdot y)$ in the stage 1. In the following, we further
1816 calculate the gradient of $\text{Loss}_5^{2,\ell}$ w.r.t. $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ based on Fact F.1.

1817 • for $\ell = 1$, letting $j_1 = \tau(g_1(y_0))$ and $j'_1 = \tau(g_2(y_0))$. Note that only when $j'' \in \{j_1, j'_1\}$, there
1818 exists $r \in [m]$, s.t., $\Lambda_{5,j'',r}^{(t)} > 0$.

1819

– For $[\mathbf{Q}_{4,3}]_{s,s}$ where $s \in \tau(\mathcal{X})$

$$\begin{aligned}
& \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,1} \right]_{s,s} \\
&= \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \left(\Xi_{\ell,i,\text{pred},1}^2 - \sum_{\mathbf{k}' \in \mathcal{I}^{2,\ell-1}} \text{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}'} \Xi_{\ell,i,\mathbf{k}'}^2 \right) \mathbb{1}_{s=\tau(x_0)} \right] \\
&= \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \left((1 - \text{logit}_{5,j_1}) \cdot \text{sReLU}'(\Lambda_{5,j_1,r_{g_1 \cdot y_0}}) (\langle W_{5,j_1,r_{g_1 \cdot y_0},2}, e_{g_1} \rangle - \Lambda_{5,j_1,r_{g_1 \cdot y_0}}) \right. \right. \\
&\quad \left. \left. - \text{logit}_{5,j'_1} \cdot \text{sReLU}'(\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}) (\langle W_{5,j'_1,r_{g_2 \cdot y_0},2}, e_{g_1} \rangle - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}}) \right. \right. \\
&\quad \left. \left. - \sum_{y \in \mathcal{Y} \setminus \{y_0\}} \text{logit}_{5,\tau(g_1(y))} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}) \right. \right. \\
&\quad \left. \left. \cdot (\langle W_{5,\tau(g_2(y)),r_{g_1 \cdot y},2}, e_{g_1} \rangle - \Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}) \right) \mathbb{1}_{\tau(x_0)=s} \right]
\end{aligned}$$

1820

– For $[\mathbf{Q}_{4,4}]_{s,s}$ where $s \in \tau(\mathcal{X})$

$$\begin{aligned}
& \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,1} \right]_{s,s} \\
&= \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \left((1 - \text{logit}_{5,j_1}) \cdot \text{sReLU}'(\Lambda_{5,j_1,r_{g_1 \cdot y_0}}) (\langle W_{5,j_1,r_{g_1 \cdot y_0},5}, e_{y_0} \rangle - \Lambda_{5,j_1,r_{g_1 \cdot y_0}}) \right. \right. \\
&\quad \left. \left. - \text{logit}_{5,j'_1} \cdot \text{sReLU}'(\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}) (\langle W_{5,j'_1,r_{g_2 \cdot y_0},5}, e_{y_0} \rangle - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}}) \right. \right. \\
&\quad \left. \left. - \sum_{y \in \mathcal{Y} \setminus \{y_0\}} \text{logit}_{5,\tau(g_1(y))} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}) \right. \right. \\
&\quad \left. \left. \cdot (\langle W_{5,\tau(g_2(y)),r_{g_1 \cdot y},5}, e_{y_0} \rangle - \Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}) \right) \mathbb{1}_{\tau(x_0)=s} \right]
\end{aligned}$$

1821

Initially, $\Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)}$ and $\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)} \gg \varrho$, and thus both are lying in the linear regime, which

1822

implies $\text{sReLU}' = 1$.

1823

- for $\ell = 2$, letting $j_2 = \tau(g_2(y_1))$ and $j'_2 = \tau(g_2(y_0))$. Initially, due to the uniform attention value $\frac{1}{4}$, and the result from stage 1, only $\Lambda_{5,\tau(g_{k'}(y_k)),r_{g_{k'} \cdot y_k}}^{(t)}$ for $k \in \{0,1\}, k' \in \{1,2\}$ are activated

1824

and in the smoothed regime. They are still very close to zero and $\text{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) =$

1825

$$\frac{\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}}{\varrho^{q-1}} = 1/\text{poly}(d).$$

1826

– For $[\mathbf{Q}_{4,3}]_{s,s}$ where $s \in \tau(\mathcal{X})$

1827

$$\begin{aligned}
& \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} \\
&= \mathbb{E} \left[\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left((1 - \text{logit}_{5,j_2}^{(t)}) \cdot \text{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) (\langle W_{5,j_2,r_{g_2 \cdot y_1},2}, e_{g_2} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \right. \right. \\
&\quad \left. \left. - \text{logit}_{5,j'_2}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}) (\langle W_{5,j'_2,r_{g_2 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}) \right. \right. \\
&\quad \left. \left. - \text{logit}_{5,\tau(g_1(y_0))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)}) (\langle W_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)}) \right. \right. \\
&\quad \left. \left. - \text{logit}_{5,\tau(g_1(y_1))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}) (\langle W_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1},2}, e_{g_2} \rangle - \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}) \right. \right. \\
&\quad \left. \left. - \sum_{y \in \mathcal{Y} \setminus \{y_0,y_1\}} \text{logit}_{5,\tau(g_2(y))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) (\langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},2}, e_{g_2} \rangle - \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) \right) \mathbb{1}_{\tau(x_1)=s} \right]
\end{aligned}$$

1828 – For $[\mathbf{Q}_{4,4}]_{s,s}$ where $s \in \tau(\mathcal{X})$

$$\begin{aligned}
& \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} \right]_{s,s} \\
&= \mathbb{E} \left[\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left((1 - \text{logit}_{5,j_2}^{(t)}) \cdot \text{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \left(\langle W_{5,j_2,r_{g_2 \cdot y_1},5}, e_{y_1} \rangle - \Lambda_{5,j,r_{g_2 \cdot y_1}}^{(t)} \right) \right. \right. \\
&\quad - \text{logit}_{5,j_2'}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_2',r_{g_2 \cdot y_0},5}, e_{y_1} \rangle - \Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)} \right) \\
&\quad - \text{logit}_{5,\tau(g_1(y_0))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)}) \left(\langle W_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0},5}, e_{y_1} \rangle - \Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)} \right) \\
&\quad \left. - \text{logit}_{5,\tau(g_1(y_1))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}) \left(\langle W_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1},5}, e_{y_1} \rangle - \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} \right) \right) \\
&\quad - \sum_{y \in \mathcal{Y} \setminus \{y_0,y_1\}} \text{logit}_{5,\tau(g_2(y))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) \left(\langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},5}, e_{y_1} \rangle - \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \right) \Big) \mathbb{1}_{\tau(x_1)=s} \right]
\end{aligned}$$

1829 In the following analysis, we assume $B = \lambda \log d$ for some large constant $\lambda > 1$. We denote the
1830 positive and negative gradients for $[\mathbf{Q}_{4,p}]_{s,s}$ from loss $\text{Loss}_{\ell,5}^2$ as $\mathcal{N}_{s,p,\ell,\text{pos}}$ and $\mathcal{N}_{s,p,\ell,\text{neg}}$ respectively.
1831 Specifically, we have

1832 • for $\ell = 1$,

1833 – For $[\mathbf{Q}_{4,3}]_{s,s}$ where $s \in \tau(\mathcal{X})$

$$\begin{aligned}
& \mathcal{N}_{s,3,1,\text{pos}} \\
&= \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \left((1 - \text{logit}_{5,j_1}^{(t)}) \cdot \text{sReLU}'(\Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_1,r_{g_1 \cdot y_0},2}, e_{g_1} \rangle - \Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} \right) \right. \right. \\
&\quad \left. \left. - \text{logit}_{5,j_1'}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j_1',r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_1',r_{g_2 \cdot y_0},2}, e_{g_1} \rangle - \Lambda_{5,j_1',r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_0)=s} \right] \\
& \mathcal{N}_{s,3,1,\text{neg}} \\
&= \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \left(- \sum_{y \in \mathcal{Y} \setminus \{y_0\}} \text{logit}_{5,\tau(g_1(y))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)}) \right. \right. \\
&\quad \left. \left. \left(\langle W_{5,\tau(g_2(y)),r_{g_1 \cdot y},2}, e_{g_1} \rangle - \Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_0)=s} \right]
\end{aligned}$$

1834 – For $[\mathbf{Q}_{4,4}]_{s,s}$ where $s \in \tau(\mathcal{X})$

$$\begin{aligned}
& \mathcal{N}_{s,4,1,\text{pos}} \\
&= \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \left((1 - \text{logit}_{5,j_1}^{(t)}) \cdot \text{sReLU}'(\Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_1,r_{g_1 \cdot y_0},5}, e_{y_0} \rangle - \Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} \right) \right. \right. \\
&\quad \left. \left. - \sum_{y \in \mathcal{Y} \setminus \{y_0\}} \text{logit}_{5,\tau(g_1(y))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)}) \left(\langle W_{5,\tau(g_2(y)),r_{g_1 \cdot y},5}, e_{y_0} \rangle - \Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_0)=s} \right] \\
& \mathcal{N}_{s,4,1,\text{neg}} \\
&= \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \left(- \text{logit}_{5,j_1'}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j_1',r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_1',r_{g_2 \cdot y_0},5}, e_{y_0} \rangle - \Lambda_{5,j_1',r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_0)=s} \right]
\end{aligned}$$

1835 • for $\ell = 2$,

1836 – For $[\mathbf{Q}_{4,3}]_{s,s}$ where $s \in \tau(\mathcal{X})$

$$\begin{aligned}
& \mathcal{N}_{s,3,2,\text{pos}} \\
&= \mathbb{E} \left[\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left((1 - \text{logit}_{5,j_2}^{(t)}) \cdot \text{sReLU}'(\Lambda_{5,j,r_{g_2 \cdot y_1}}^{(t)}) \left(\langle W_{5,j,r_{g_2 \cdot y_1},2}, e_{g_2} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \right. \right.
\end{aligned}$$

$$\begin{aligned}
& - \text{logit}_{5,\tau(g_1(y_0))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)}) \left(\langle W_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)} \right) \\
& - \text{logit}_{5,\tau(g_1(y_1))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}) \left(\langle W_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1},2}, e_{g_2} \rangle - \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} \right) \Big) \mathbb{1}_{\tau(x_1)=s} \Big] \\
\mathcal{N}_{s,3,2,\text{neg}} &= \mathbb{E} \left[\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left(- \text{logit}_{5,j'_2}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_2,r_{g_2 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} \right) \right. \right. \\
& \left. \left. - \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \text{logit}_{5,\tau(g_2(y))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) \left(\langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},2}, e_{g_2} \rangle - \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
1837 \quad & - \text{For } [\mathbf{Q}_{4,4}]_{s,s} \text{ where } s \in \tau(\mathcal{X}) \\
\mathcal{N}_{s,4,2,\text{pos}} &= \mathbb{E} \left[\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left((1 - \text{logit}_{5,j_2}^{(t)}) \cdot \text{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \left(\langle W_{5,j_2,r_{g_2 \cdot y_1},5}, e_{y_1} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \right. \right. \\
& - \text{logit}_{5,j'_2}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_2,r_{g_2 \cdot y_0},5}, e_{y_1} \rangle - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} \right) \\
& - \text{logit}_{5,\tau(g_1(y_0))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)}) \left(\langle W_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0},5}, e_{y_1} \rangle - \Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)} \right) \\
& \left. \left. - \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \text{logit}_{5,\tau(g_2(y))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) \left(\langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},5}, e_{y_1} \rangle - \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
\mathcal{N}_{s,4,2,\text{neg}} &= \mathbb{E} \left[\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left(- \text{logit}_{5,\tau(g_1(y_1))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}) \right. \right. \\
& \left. \left. \cdot \left(\langle W_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1},5}, e_{y_1} \rangle - \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right]
\end{aligned}$$

1838 F.2 Stage 2.1: Growth of Gap

1839 **Induction F.1.** Given $s \in \tau(\mathcal{X})$, let $T_{2,1,s}$ denote the first time that $[\mathbf{Q}_{4,3}]_{s,s}$ reaches $\Omega(\frac{1}{\log d})$. For
1840 all iterations $t \leq T_{2,1,s}$, we have the following holds

- 1841 • $[\mathbf{Q}_{4,3}]_{s,s}$ monotonically increases
- 1842 • $|[\mathbf{Q}_{4,4}]_{s,s}| \leq [\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,3}]_{s,s} - [\mathbf{Q}_{4,4}]_{s,s} = \Theta([\mathbf{Q}_{4,3}]_{s,s})$;
- 1843 • for $p \in \{3, 4\}$, for $j \in \tau(\mathcal{X}) \neq s$, $|[\mathbf{Q}_{4,p}]_{s,j}| \leq O(\frac{[\mathbf{Q}_{4,p}]_{s,j}}{d})$; otherwise $[\mathbf{Q}_{4,p}]_{s,j} = 0$

1844 F.2.1 Attention and Logit Preliminaries

1845 **Lemma F.1.** If Induction F.1 holds for all iterations $< t$, then we have

- 1846 1. for $\ell = 1$,
 - 1847 (a) $\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \in [\frac{1}{3}, \frac{1}{3} + O(\frac{1}{\log d})]$;
 - 1848 (b) $\text{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)}, \text{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \leq \text{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)}$;
 - 1849 (c) $|\text{Attn}_{\text{ans},0 \rightarrow \mathbf{k}}^{(t)} - \text{Attn}_{\text{ans},0 \rightarrow \mathbf{k}'}^{(t)}| \leq O(\frac{1}{\log d})$ for $\mathbf{k} \neq \mathbf{k}' \in \mathcal{I}^{(2,0)}$
- 1850 2. for $\ell = 2$,

- 1851 (a) $\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \in [\frac{1}{4}, \frac{1}{4} + O(\frac{1}{\log d})]$;
 1852 (b) $\text{Attn}_{\text{ans},1 \rightarrow \mathbf{k}}^{(t)} \leq \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}$ for $\mathbf{k} \neq (\text{pred}, 2)$;
 1853 (c) $|\text{Attn}_{\text{ans},1 \rightarrow \mathbf{k}}^{(t)} - \text{Attn}_{\text{ans},1 \rightarrow \mathbf{k}'}^{(t)}| \leq O(\frac{1}{\log d})$ for $\mathbf{k} \in \{(\text{ans}, 0), (\text{pred}, 1)\}, \mathbf{k}' \in$
 1854 $\{(\text{ans}, 1), (\text{pred}, 2)\}$

1855 **Lemma F.2.** *If Induction F.1 holds for all iterations $< t$, then we have*

- 1856 1. for $\ell = 1$, $\text{logit}_{5,j}^{(t)} = \Omega(1)$ for $j \in \{j_1, j'_1\}$, $1 - \text{logit}_{5,j_1}^{(t)} - \text{logit}_{5,j'_1}^{(t)} = \frac{1}{\text{poly}d}$;
 1857 2. for $\ell = 2$, $\text{logit}_{5,j}^{(t)} = O(\frac{1}{d})$ for all j .

1858 *Proof.* • For $\ell = 1$, we have

$$\begin{aligned}
 & \text{logit}_{j_1}^{(t)} \\
 &= \frac{1}{1 + e^{2(\text{Attn}_{\text{ans},0 \rightarrow \text{pred},2} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},1})^F} + O(d) \cdot e^{-(\text{Attn}_{\text{ans},0 \rightarrow \text{ans},1} + \text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2})^F \pm \frac{1}{\text{poly}d}}} \\
 &= \frac{1}{1 + e^{-O(\frac{1}{\log d})^F} + O(d) \cdot e^{-(\frac{1}{3} + O(\frac{1}{\log d}))^F}} \geq \Omega(1) \\
 & \text{logit}_{j'_1}^{(t)} \\
 &= \frac{1}{1 + e^{2(\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2})^F} + O(d) \cdot e^{-(\text{Attn}_{\text{ans},0 \rightarrow \text{ans},1} + \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},1})^F \pm \frac{1}{\text{poly}d}}} \\
 &= \frac{1}{1 + e^{O(\frac{1}{\log d})^F} + O(d) \cdot e^{-(\frac{1}{3} - O(\frac{1}{\log d}))^F}} \geq \Omega(1) \\
 & 1 - \text{logit}_{j_1}^{(t)} - \text{logit}_{j'_1}^{(t)} \\
 &= \frac{O(1)}{\left(e^{(1-2\text{Attn}_{\text{ans},0 \rightarrow \text{pred},2})^F \pm \frac{1}{\text{poly}d}} + e^{(1-2\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1})^F \pm \frac{1}{\text{poly}d}}\right) \cdot d^{-1} + O(1)} \\
 &\leq \frac{O(1)}{e^{(\frac{1}{3} + O(\frac{1}{\log d}))^F} \cdot d^{-1} + O(1)} \leq \frac{1}{\text{poly}d}
 \end{aligned}$$

- 1861 • For $\ell = 2$, we only need to focus on j_2 and j'_2 ,

$$\begin{aligned}
 & \text{logit}_{j_2}^{(t)} = \\
 & \frac{1}{1 + e^{2(\text{Attn}_{\text{ans},1 \rightarrow \text{ans},0} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1})^F} + O(d) \cdot e^{-(\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1} + \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \text{Attn}_{\text{ans},1 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},0})^F \pm \frac{1}{\text{poly}d}}} \\
 &= \frac{1}{1 + e^{\pm O(\frac{1}{\log d})^F} + O(d) \cdot e^{\pm O(\frac{1}{\log d})^F}} = O\left(\frac{1}{d}\right) \\
 & \text{logit}_{j'_2}^{(t)} = \\
 & \frac{1}{1 + e^{2(\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},0})^F} + O(d) \cdot e^{-(\text{Attn}_{\text{ans},1 \rightarrow \text{ans},0} + \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \text{Attn}_{\text{ans},1 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1})^F \pm \frac{1}{\text{poly}d}}} \\
 &= \frac{1}{1 + e^{\pm O(\frac{1}{\log d})^F} + O(d) \cdot e^{\pm O(\frac{1}{\log d})^F}} = O\left(\frac{1}{d}\right)
 \end{aligned}$$

1863 \square

1864 **F.2.2 Gradient Lemma**

1865 Notice that during these phase, for $\ell = 1$, $\Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)}$ for $y \neq y_0$ cannot be activated since
 1866 $2\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} - 1 < 0$; and similarly for $\ell = 2$, $\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}$ for $y \notin \{y_0, y_1\}$ has not been
 1867 activated.

1868 **Lemma F.3.** *If Induction F.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$, we have*

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \Theta\left(\frac{\log d}{d}\right).$$

1869 *Proof.* Clearly, $\mathcal{N}_{s,3,1,\text{neg}} = 0$. By Lemma F.2, we can bound negative gradient $\mathcal{N}_{s,3,2,\text{neg}}$ as follows

$$\begin{aligned} & |\mathcal{N}_{s,3,2,\text{neg}}| \\ &= \mathbb{E} \left[\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left(\text{logit}_{5,j'_2}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_2,r_{g_2 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\ &\leq O\left(\frac{\log d}{d}\right) \cdot \frac{1}{d} = \tilde{O}\left(\frac{1}{d^2}\right) \end{aligned}$$

1870 On the other hand, we have

$$\begin{aligned} & \mathcal{N}_{s,3,1,\text{pos}} \\ &= \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \cdot \left((1 - \text{logit}_{5,j'_1}^{(t)}) \cdot \text{sReLU}'(\Lambda_{5,j'_1,r_{g_1 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_1,r_{g_1 \cdot y_0},2}, e_{g_1} \rangle - \Lambda_{5,j'_1,r_{g_1 \cdot y_0}}^{(t)} \right) \right. \right. \\ &\quad \left. \left. - \text{logit}_{5,j'_1}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_1,r_{g_2 \cdot y_0},2}, e_{g_1} \rangle - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_0)=s} \right] \\ &= \Theta(1) \cdot \Theta(\log d) \cdot \frac{1}{d} = \Theta\left(\frac{\log d}{d}\right). \end{aligned}$$

1871 Furthermore, since $\text{sReLU}'(\Lambda_{5,j'_2,r_{g_2 \cdot y_1}}^{(t)}) \leq 1$, we have $\mathcal{N}_{s,3,2,\text{pos}} \leq O(\mathcal{N}_{s,3,1,\text{pos}})$. Thus we
 1872 complete the proof. \square

1873 **Lemma F.4** (negative gradient). *If Induction F.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, we*
 1874 *have*

$$\sum_{\ell=1}^2 \left| \min \left\{ \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s}, 0 \right\} \right| \leq O\left(\frac{1}{\log d}\right) \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s}.$$

Proof.

$$\begin{aligned} & \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,1} \right]_{s,s} \\ &= \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot \left((1 - \text{logit}_{5,j'_1}^{(t)}) \cdot \text{sReLU}'(\Lambda_{5,j'_1,r_{g_1 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_1,r_{g_1 \cdot y_0},5}, e_{y_0} \rangle - \Lambda_{5,j'_1,r_{g_1 \cdot y_0}}^{(t)} \right) \right. \right. \\ &\quad \left. \left. - \text{logit}_{5,j'_1}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_1,r_{g_2 \cdot y_0},5}, e_{y_0} \rangle - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_0)=s} \right] \\ &= \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot \left((1 - \text{logit}_{5,j'_1}^{(t)}) \cdot \left(2\text{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)} F \pm \frac{1}{\text{poly}d} \right) \right. \right. \\ &\quad \left. \left. - (1 - \text{logit}_{5,j'_1}^{(t)} - \frac{1}{\text{poly}d}) \cdot \left(2\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} F \pm \frac{1}{\text{poly}d} \right) \right) \mathbb{1}_{\tau(x_0)=s} \right] \end{aligned}$$

1875 Therefore, by Lemma F.1, we have

$$|\min \left\{ \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,1} \right]_{s,s}, 0 \right\}|$$

$$\begin{aligned}
&\leq \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot \left((1 - \mathbf{logit}_{5,j_1}^{(t)}) \cdot O\left(\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)}\right) F \right) \mathbb{1}_{\tau(x_0)=s} \right] \\
&\leq \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot \left((1 - \mathbf{logit}_{5,j_1}^{(t)}) \cdot O\left(\frac{1}{\log d}\right) F \right) \mathbb{1}_{\tau(x_1)=s} \right] \leq O\left(\frac{\mathcal{N}_{s,3,1,\text{pos}}}{\log d}\right)
\end{aligned}$$

1876 Moreover,

$$\begin{aligned}
&|\min \left\{ \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s}, 0 \right\}| \leq |\mathcal{N}_{s,4,2,\text{neg}}| \\
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left(\mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}) \right. \right. \\
&\quad \left. \left. \cdot \left(\langle W_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1},5}, e_{y_1} \rangle - \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
&\stackrel{(i)}{\leq} \tilde{O}\left(\frac{1}{\text{poly}d}\right) \ll O\left(\frac{\mathcal{N}_{s,3,1,\text{pos}}}{\log d}\right),
\end{aligned}$$

1877 where the inequality (i) holds since $\mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} \leq O(\frac{1}{d})$ and $\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} = O(\frac{\log d}{\text{poly}d})$.

1878 Combining the above two inequalities and the fact that $\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \geq \Omega(\mathcal{N}_{s,3,1,\text{pos}})$
1879 from the previous lemma, we complete the proof. \square

1880 **Lemma F.5** (Growth of gap). *If Induction F.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, we have*

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} - \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \geq \Omega\left(\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s}\right).$$

1881 *Proof.* For $\ell = 2$

$$\begin{aligned}
&\left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} \\
&\leq \mathcal{N}_{s,4,2,\text{pos}} - \mathcal{N}_{s,3,2,\text{pos}} - \mathcal{N}_{s,3,2,\text{neg}} \\
&\leq \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left(-\mathbf{logit}_{5,j_2'}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_2',r_{g_2 \cdot y_0},5}, e_{y_1} \rangle - \Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
&+ \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left(\mathbf{logit}_{5,j_2'}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_2',r_{g_2 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)} \right) \right. \right. \\
&\quad \left. \left. - \sum_{y \in \mathcal{Y} \setminus \{y_0, 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)}) \left(\langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},2}, e_{g_2} \rangle - \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left(-\mathbf{logit}_{5,j_2'}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_2',r_{g_2 \cdot y_0},5}, e_{y_1} \rangle - \Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
&+ \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left(\mathbf{logit}_{5,j_2'}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_2',r_{g_2 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
&\leq O\left(\frac{\log d}{d^2}\right).
\end{aligned}$$

1882 This implies the gradient difference between $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ contributed by $\ell = 2$ is negligible. For
1883 $\ell = 1$, since $\mathcal{N}_{s,3,1,\text{neg}}$ has not been activated, we have

$$\begin{aligned}
&\left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,1} \right]_{s,s} + \left[\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,1} \right]_{s,s} = \mathcal{N}_{s,3,1,\text{pos}} - \mathcal{N}_{s,4,1,\text{pos}} - \mathcal{N}_{s,4,1,\text{neg}} \\
&\geq \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \cdot \left(-\mathbf{logit}_{5,j_1'}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j_1',r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_1',r_{g_2 \cdot y_0},2}, e_{g_1} \rangle - \Lambda_{5,j_1',r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_0)=s} \right]
\end{aligned}$$

$$\begin{aligned}
& + \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \cdot \left(\mathbf{logit}_{5,j_1'}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j_1',r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_1',r_{g_2 \cdot y_0}}, 5, e_{y_0} \rangle - \Lambda_{5,j_1',r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_0)=s} \right] \\
& = \Omega \left(\left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} \right) \geq \Omega \left(\frac{\log d}{d} \right).
\end{aligned}$$

1884 Putting it all together, we finish the proof. \square

1885 **Lemma F.6.** *If Induction F.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} \geq \Omega(\frac{\rho}{\log d})$,*
1886 *we have*

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \geq \Omega \left(\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \right).$$

1887 *Proof.* Notice that by Lemma F.4, when $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} \geq \Omega(\frac{\rho}{\log d})$, we have $[\mathbf{Q}_{4,4}^{(t)}]_{s,s} \geq -O(\frac{\rho}{\log^2 d})$.
1888 Thus

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \geq \Omega \left(\frac{\rho}{\log d} \right)$$

1889 which implies $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}$ already lies in the linear regime. Then we have

$$\begin{aligned}
& \mathcal{N}_{s,4,2,\text{pos}} = \\
& \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left((1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \left(\langle W_{5,j_2,r_{g_2 \cdot y_1}}, 5, e_{y_1} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \right. \right. \\
& \quad \left. \left. - O\left(\frac{1}{d}\right) \cdot \mathbf{sReLU}'(\Lambda_{5,j_2',r_{g_2 \cdot y_1}}^{(t)}) \left(\langle W_{5,j_2',r_{g_2 \cdot y_1}}, 5, e_{y_1} \rangle - \Lambda_{5,j_2',r_{g_2 \cdot y_1}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \geq \Omega \left(\frac{\log d}{d} \right)
\end{aligned}$$

1890 Similarly, we have $|\mathcal{N}_{s,4,2,\text{neg}}| \leq O(\frac{\log d}{d^2})$ and by Lemma F.4, we have $|\mathcal{N}_{s,4,1,\text{neg}}| \leq O(\frac{1}{\log d})$.

1891 $\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s}$. Putting it all together, we complete the proof. \square

1892 **Lemma F.7.** *If Induction F.2 holds for all iterations $< t$, given $s \neq j \in \tau(\mathcal{X})$, for $p \in \{3, 4\}$, we*
1893 *have*

$$\left| \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,j} \right| \leq O\left(\frac{\log d}{d^2}\right) = O\left(\frac{1}{d}\right) \cdot \left| \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \right|.$$

1894 F.2.3 At the End of Stage 2.1

1895 Putting gradient lemmas together, we can directly prove that Induction F.1 holds for all iterations t
1896 until the end of stage 2.1, where we can conclude the following:

1897 **Lemma F.8** (End of Stage 2.1). *Given $s \in \tau(\mathcal{X})$, Induction F.1 holds for all iterations $t < T_{2,1,s} =$*
1898 *$O(\frac{d}{\eta \log^2 d})$, then at the end of stage 2.1, we have*

- 1899 • $[\mathbf{Q}_{4,p}^{(t)}]_{s,s} = \Omega(\frac{1}{\log d})$ for $p \in \{3, 4\}$;
- 1900 • $|[\mathbf{Q}_{4,3}^{(t)}]_{s,s} - [\mathbf{Q}_{4,4}^{(t)}]_{j,j}| \geq \Omega(\frac{1}{\log d})$;
- 1901 • $|[\mathbf{Q}_{4,p}^{(t)}]_{s,j}| \leq O(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,s}}{d})$ for $j \in \tau(\mathcal{X}) \neq s$ for $p \in \{3, 4\}$; otherwise, $[\mathbf{Q}_{4,p}^{(t)}]_{j,s} = 0$;

1902 F.3 Stage 2.2: Growth of $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$

1903 In stage 2.2, we will continue to grow $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ until they reach a certain threshold. The analysis
1904 here is similar to the analysis in stage 2.1, but we shift our focus to the gradients contributed by $\ell = 2$
1905 since the loss for $\ell = 1$ has already been activated well and start to contribute negligible to the growth
1906 of $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$.

1907 **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 (small
 1908 enough to ensure $2\lambda(c_1+c_2) < 1-c_5\lambda$ and $1-(1-2c_4)\lambda > 0$). For all iterations $T_{2,1,s} < t < T_{2,2,s}$,
 1909 we have the following holds

- 1910 • $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}, [\mathbf{Q}_{4,4}^{(t)}]_{s,s} \leq O(1)$ monotonically increases;
- 1911 • $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} - [\mathbf{Q}_{4,4}^{(t)}]_{s,s} \leq c_0$ monotonically increases;
- 1912 • for $(p, q) \in \{(4, 3), (4, 4)\}$, $||[\mathbf{Q}_{p,q}^{(t)}]_{s,j}|| \leq O(\frac{[\mathbf{Q}_{p,q}^{(t)}]_{s,s}}{d})$ for $j \in \tau(\mathcal{X}) \neq s$; other $[\mathbf{Q}_{p,q}^{(t)}]_{s,j} =$
 1913 0

1914 F.3.1 Attention and Logit Preliminaries

1915 **Lemma F.9.** If Induction F.2 holds for all iterations $< t$, then we have

- 1916 1. for $\ell = 1$,
 - 1917 • $\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \in [\frac{1}{3} + \Omega(\frac{1}{\log d}), \frac{1}{3} + c_1]$
 - 1918 • $\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)} \in [\frac{1}{3} - c_2, \frac{1}{3} - \Omega(\frac{1}{\log d})]$
 - 1919 • $\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)} + \Omega(\frac{1}{\log d}) \leq \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)}$
 - 1920 • $\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \in [\Omega(\frac{1}{\log d}), c_1 + c_2]$
- 1921 2. for $\ell = 2$,
 - 1922 • $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \in [\frac{1}{4} + \Omega(\frac{1}{\log d}), \frac{1}{4} + c_3]$
 - 1923 • $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)}, \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \in [\frac{1}{4} - c_4, \frac{1}{4} - \Omega(\frac{1}{\log d})]$, $|\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} -$
 1924 $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)}| \leq O(\frac{1}{d})$
 - 1925 • $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)}, \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \leq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \Omega(\frac{1}{\log d})$
 - 1926 • $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \in [\Omega(\frac{1}{\log d}), c_5]$

1927 **Lemma F.10.** If Induction F.2 holds for all iterations $< t$, then we have

- 1928 1. for $\ell = 1$, $\mathbf{logit}_{5,j'_1}^{(t)} \geq \Omega(\frac{1}{d^{2\lambda(c_1+c_2)}})$.
- 1929 2. for $\ell = 2$, $1 - \mathbf{logit}_{5,j'_2}^{(t)} = \Omega(1)$, $\mathbf{logit}_{5,j'_2}^{(t)} = O(\frac{1}{d^{1-c_5\lambda}})$.

1930 *Proof.* • For $\ell = 1$, we have

$$\begin{aligned}
 & \mathbf{logit}_{j'_1}^{(t)} \\
 &= \frac{1}{1 + e^{2(\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} - \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2})^F} + O(d) \cdot e^{-(\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},1} + \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1})^F \pm \frac{1}{\text{poly}d}}} \\
 &= \frac{1}{1 + e^{2(c_1+c_2)F} + O(d) \cdot e^{-(\frac{1}{3}-2c_1)^F}} \geq \Omega(\frac{1}{d^{2\lambda(c_1+c_2)}})
 \end{aligned}$$

1931 • For $\ell = 2$, we have

$$\begin{aligned}
 & \mathbf{logit}_{j'_2}^{(t)} = \\
 & \frac{1}{1 + e^{2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1})^F} + O(d) \cdot e^{-(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} + \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0})^F \pm \frac{1}{\text{poly}d}}} \\
 & \leq \frac{1}{1 + O(d) \cdot e^{-(1-2c_4)^F}} = O(\frac{1}{d^{1-(1-2c_4)\lambda}})
 \end{aligned}$$

1932

$$\begin{aligned} \mathbf{logit}_{j'_2}^{(t)} &= \frac{1}{1 + e^{2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0})^F + O(d) \cdot e^{-(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} + \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1})^F \pm \frac{1}{\text{poly}d}}} \\ &\leq \frac{1}{1 + e^{\Omega(\frac{1}{\log d})^F} + O(d) \cdot e^{-(c_5 + O(\frac{1}{d}))^F}} \leq O\left(\frac{1}{d^{1-c_5\lambda}}\right) \end{aligned}$$

1933

□

1934 **F.3.2 Gradient Lemma**

1935 Notice that during these sub-stage, for $\ell = 2$, $\Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)}$ for $y \in \{y_0, y_1\}$ cannot be activated
 1936 anymore, due to slightly dominant attention role of $\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}$.

1937 **Lemma F.11.** *If Induction F.2 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$, we have*

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \Theta\left(\frac{\log d}{d}\right).$$

1938 *Proof.* By Lemma F.10, since $1 - \mathbf{logit}_{5,j_2}^{(t)} = \Omega(1)$ and $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}$ is already in the linear regime,
 1939 we have

$$\mathcal{N}_{s,3,1,\text{pos}} \leq O(\mathcal{N}_{s,3,2,\text{pos}}).$$

1940 For $\mathcal{N}_{s,3,2,\text{pos}}$, we have

$$\begin{aligned} \mathcal{N}_{s,3,2,\text{pos}} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left((1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \left(\langle W_{5,j_2,r_{g_2 \cdot y_1},2}, e_{g_2} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \right) \mathbf{1}_{\tau(x_1)=s} \right] \\ &= \Theta(1) \cdot \Theta(\log d) \cdot \frac{1}{d} = \Theta\left(\frac{\log d}{d}\right). \end{aligned}$$

1941 Furthermore, we can bound negative gradient $\mathcal{N}_{s,3,2,\text{neg}}$ as follows

$$\begin{aligned} |\mathcal{N}_{s,3,2,\text{neg}}| &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left(\mathbf{logit}_{5,j'_2}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_2,r_{g_2 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbf{1}_{\tau(x_1)=s} \right] \\ &\leq O\left(\frac{\log d}{d^{1-c_5\lambda}}\right) \cdot \frac{1}{d} = \tilde{O}\left(\frac{1}{d^{2-c_5\lambda}}\right), \end{aligned}$$

1942 which is negligible compared to $\mathcal{N}_{s,3,2,\text{pos}}$. Thus we complete the proof. □

1943 **Lemma F.12.** *If Induction F.2 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, we have*

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \Theta\left(\frac{\log d}{d}\right).$$

1944 *Proof.* For $\ell = 2$,

$$\begin{aligned} \mathcal{N}_{s,4,2,\text{pos}} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left((1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \left(\langle W_{5,j_2,r_{g_2 \cdot y_1},5}, e_{y_1} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \right. \right. \\ &\quad \left. \left. - O\left(\frac{1}{d}\right) \cdot \mathbf{sReLU}'(\Lambda_{5,j'_2,r_{g_2 \cdot y_1}}^{(t)}) \left(\langle W_{5,j'_2,r_{g_2 \cdot y_0},5}, e_{y_1} \rangle - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbf{1}_{\tau(x_1)=s} \right] \geq \Omega\left(\frac{\log d}{d}\right) \end{aligned}$$

1945 In this sub-stage $\mathcal{N}_{s,4,2,\text{neg}} = 0$. Then turn to $\ell = 1$, when $[\mathbf{Q}_{4,3}]_{s,s} \leq O(\frac{1}{\log \log d})$, following the
 1946 analysis in Lemma F.4, we have

$$\begin{aligned}
 & \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,1} \right]_{s,s} \\
 &= \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \left((1 - \text{logit}_{5,j_1}^{(t)}) \cdot \text{sReLU}'(\Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_1,r_{g_1 \cdot y_0},5}, e_{y_0} \rangle - \Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} \right) \right. \right. \\
 & \quad \left. \left. - \text{logit}_{5,j'_1}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_1,r_{g_2 \cdot y_0},5}, e_{y_0} \rangle - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_0)=s} \right] \\
 &= \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \left((1 - \text{logit}_{5,j_1}^{(t)}) \cdot \left(2\text{Attn}_{\text{ans},0 \rightarrow \text{pred},2} F \pm \frac{1}{\text{poly}d} \right) \right. \right. \\
 & \quad \left. \left. - (1 - \text{logit}_{5,j_1}^{(t)} - \frac{1}{\text{poly}d}) \cdot \left(2\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} F \pm \frac{1}{\text{poly}d} \right) \right) \mathbb{1}_{\tau(x_0)=s} \right] \\
 &\geq -O\left(\frac{1}{\log \log d}\right) \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \left((1 - \text{logit}_{5,j_1}^{(t)}) \cdot \right. \right. \\
 & \quad \left. \left. \left(2(\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2}) F \pm \frac{1}{\text{poly}d} \right) \right) \mathbb{1}_{\tau(x_1)=0} \right] \\
 &\geq -O\left(\frac{1}{\log \log d}\right) \cdot \mathcal{N}_{s,4,1,\text{pos}}
 \end{aligned}$$

1947 Since $\mathcal{N}_{s,4,1,\text{pos}} \leq O(\mathcal{N}_{s,4,2,\text{pos}})$, thus the contribution of $\left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,1} \right]_{s,s}$ is negligible compared
 1948 to $\mathcal{N}_{s,4,2,\text{pos}}$. Once $[\mathbf{Q}_{4,3}]_{s,s}$ reaches $\Omega(\frac{1}{\log \log d})$, we have

$$\begin{aligned}
 & \text{logit}_{j'_1}^{(t)} \\
 &= \frac{1}{1 + e^{2(\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2}) F + O(d)} \cdot e^{-(\text{Attn}_{\text{ans},0 \rightarrow \text{ans},1} + \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},1}) F \pm \frac{1}{\text{poly}d}}} \\
 &\leq \frac{1}{1 + e^{O(\frac{\log d}{\log \log d})}} = o(1) \ll 1 - \text{logit}_{5,j_2}^{(t)},
 \end{aligned}$$

1949 which implies $\mathcal{N}_{s,4,1,\text{neg}}$ can be dominated by $\mathcal{N}_{s,4,2,\text{pos}}$. Putting it all together, we have

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,\ell} \right]_{s,s} = \Theta(\mathcal{N}_{s,4,2,\text{pos}}) = \Theta\left(\frac{\log d}{d}\right).$$

1950 □

1951 **Lemma F.13** (Growth of gap). *If Induction F.2 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, we have*

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}^{(t)}_{4,3}} \text{Loss}_5^{2,\ell} \right]_{s,s} - \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,\ell} \right]_{s,s} \geq 0$$

1952 *Proof.* For $\ell = 1$, since $\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \geq \text{Attn}_{\text{ans},0 \rightarrow \text{ans},0}$ we have

$$\begin{aligned}
 & \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} \right]_{s,s} = \mathcal{N}_{s,3,1,\text{pos}} - \mathcal{N}_{s,4,1,\text{pos}} - \mathcal{N}_{s,4,1,\text{neg}} \\
 &\geq \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \left(-\text{logit}_{5,j'_1}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_1,r_{g_2 \cdot y_0},2}, e_{g_1} \rangle - \Lambda_{5,j'_1,r_{g_1 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_0)=s} \right] \\
 &+ \mathbb{E} \left[\text{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \left(\text{logit}_{5,j'_1}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_1,r_{g_2 \cdot y_0},5}, e_{y_0} \rangle - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_0)=s} \right]
 \end{aligned}$$

$$\geq \Omega\left(\frac{\log d}{d^{1+2\lambda(c_1+c_2)}}\right).$$

1953 where the last inequality due to Lemma F.10 that $\mathbf{logit}_{5,j'_1}^{(t)} \geq \Omega\left(\frac{1}{d^{2\lambda(c_1+c_2)}}\right)$.

1954 For $\ell = 2$

$$\begin{aligned} & \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} \right]_{s,s} + \left[\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s} \\ & \leq \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left(-\mathbf{logit}_{5,j'_2}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_2,r_{g_2 \cdot y_0},5}, e_{y_1} \rangle - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\ & + \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left(\mathbf{logit}_{5,j'_2}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_2,r_{g_2 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\ & \leq O\left(\frac{\log d}{d^2}\right). \end{aligned}$$

1955 This implies the gradient difference between $\mathbf{Q}_{4,3}$ and $\mathbf{Q}_{4,4}$ contributed by $\ell = 2$ is still negligible.
1956 Putting it all together, we finish the proof. \square

1957 **Lemma F.14.** *If Induction F.2 holds for all iterations $< t$, given $s \neq j \in \tau(\mathcal{X})$, for $p \in \{3, 4\}$, we*
1958 *have*

$$\left| \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,j} \right| \leq O\left(\frac{\log d}{d^2}\right) = O\left(\frac{1}{d}\right) \cdot \left| \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \right|.$$

1959 F.3.3 At the End of Stage 2.2

1960 Putting gradient lemmas together, we can directly prove that Induction F.2 holds for all iterations t
1961 until the end of stage 2.2, where we can conclude the following:

1962 **Lemma F.15** (End of Stage 2.2). *Given $s \in \tau(\mathcal{X})$, Induction F.2 holds for all iterations $t < T_{2,2,s} =$*
1963 *$O\left(\frac{d}{\eta \log d}\right)$, then at the end of stage 2.2, we have*

- 1964 • $[\mathbf{Q}_{4,p}^{(t)}]_{s,s} = \Omega(1)$ for $p \in \{3, 4\}$;
- 1965 • $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} - [\mathbf{Q}_{4,4}^{(t)}]_{s,s} \in [\Omega\left(\frac{1}{\log d}\right), O(1)]$;
- 1966 • $|[\mathbf{Q}_{4,p}^{(t)}]_{s,j}| \leq O\left(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,s}}{d}\right)$ for $j \in \tau(\mathcal{X}) \neq s$, and other $[\mathbf{Q}_{4,p}]_{s,j} = 0$;

1967 F.4 Stage 2.3: Decrease of Gap and Convergence

1968 **Induction F.3.** *Given $\epsilon \geq \tilde{\Omega}(\sigma_0)$. For $s \in \tau(\mathcal{X})$, let $T_{2,3,s}$ denote the first time that*
1969 *$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \geq 1 - \epsilon$. For all iterations $T_{2,2,s} < t < T_{2,3,s}$, we have*
1970 *the following holds:*

- 1971 • $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$ and $[\mathbf{Q}_{4,4}^{(t)}]_{s,s}$ monotonically increases $\leq \tilde{O}(1)$;
- 1972 • $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \geq 0$;
- 1973 • $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \leq 0.5 + \tilde{c}$ for some small constant $\tilde{c} > 0$;
- 1974 • for $(p, q) \in \{(4, 3), (4, 4)\}$, $|[\mathbf{Q}_{p,q}^{(t)}]_{s,j}| \leq O\left(\frac{[\mathbf{Q}_{p,q}^{(t)}]_{s,s}}{d}\right)$ for $j \in \tau(\mathcal{X}) \neq s$; other $[\mathbf{Q}_{p,q}^{(t)}]_{s,j} =$
1975 0 .

1976 F.4.1 Attention and Logit Preliminaries

1977 **Lemma F.16.** *If Induction F.3 holds for all iterations $< t$, then we have*

$$1. \text{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(t)} \geq \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)};$$

$$2. \text{ for } \ell = 1,$$

$$\bullet \text{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} \geq \Omega(1), \text{ and } \text{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(t)} > \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(t)};$$

$$3. \text{ for } \ell = 2,$$

$$\bullet \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \geq \Omega(1), \text{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)}, \text{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \leq \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)};$$

$$\bullet |\text{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} - \text{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)}| \leq \tilde{O}(\frac{1}{d}).$$

Lemma F.17. *If Induction F.3 holds for all iterations $< t$, then we have*

$$1. \text{logit}_{5,j'_1}^{(t)} \leq O\left(\frac{1}{d^{2\lambda}(\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2})}\right),$$

$$2. 1 - \text{logit}_{j_1}^{(t)} \leq O\left(\frac{1}{d^{2\lambda}(\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2}) - \min\{1 - \lambda(1 - 2\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1}), 0\}}\right),$$

$$3. 1 - \text{logit}_{j_2}^{(t)} \geq \min\left\{\frac{1}{2}, \Omega\left(\frac{1}{d^{\lambda}(\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1} + \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \text{Attn}_{\text{ans},1 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},0}) - 1}\right)\right\}.$$

Proof. • For $\ell = 1$, we have

$$\begin{aligned} \text{logit}_{j'_1}^{(t)} &= \frac{1}{1 + e^{2(\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2})F} + O(d) \cdot e^{-(\text{Attn}_{\text{ans},0 \rightarrow \text{ans},1} + \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},1})F \pm \frac{1}{\text{poly}d}}} \\ &\leq O\left(\frac{1}{d^{2\lambda}(\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2})}\right) \end{aligned}$$

Similarly, we have

$$\begin{aligned} 1 - \text{logit}_{j_1}^{(t)} &= \frac{1 + O(d) \cdot e^{-(\text{Attn}_{\text{ans},0 \rightarrow \text{ans},1} + \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},1})F \pm \frac{1}{\text{poly}d}}}{1 + e^{2(\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2})F} + O(d) \cdot e^{-(\text{Attn}_{\text{ans},0 \rightarrow \text{ans},1} + \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},1})F \pm \frac{1}{\text{poly}d}}} \\ &\leq O\left(\frac{1}{d^{2\lambda}(\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2}) - \min\{1 - \lambda(1 - 2\text{Attn}_{\text{ans},0 \rightarrow \text{pred},1}), 0\}}\right) \end{aligned}$$

• For $\ell = 2$, we have

$$\begin{aligned} 1 - \text{logit}_{j_2}^{(t)} &= \frac{e^{2(\text{Attn}_{\text{ans},1 \rightarrow \text{ans},0} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1})F} + O(d) \cdot e^{-(\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1} + \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \text{Attn}_{\text{ans},1 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},0})F \pm \frac{1}{\text{poly}d}}}{1 + e^{2(\text{Attn}_{\text{ans},1 \rightarrow \text{ans},0} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1})F} + O(d) \cdot e^{-(\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1} + \text{Attn}_{\text{ans},0 \rightarrow \text{pred},2} - \text{Attn}_{\text{ans},1 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},0})F \pm \frac{1}{\text{poly}d}}} \\ &\geq \min\left\{\frac{1}{2}, \Omega\left(\frac{1}{d^{\lambda}(\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1} + \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \text{Attn}_{\text{ans},1 \rightarrow \text{pred},1} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},0}) - 1}\right)\right\} \end{aligned}$$

□

F.4.2 Gradient Lemma

Lemma F.18. *If Induction F.3 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,4}]_{s,s}$, we have*

$$\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \geq \Omega\left(\frac{\epsilon \log d}{d^{(1-2\epsilon)\lambda}}\right).$$

1994 *Proof.* When $\lambda(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}) \leq 1$,
 1995 we have $\mathbf{logit}_{j'_1}^{(t)} \leq O(\frac{1}{d^{c_8}}) \cdot (1 - \mathbf{logit}_{j_2}^{(t)})$ for some small constant $c_8 > 0$. During this phase,
 1996 $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \geq \Omega(1)$, and $\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}, \Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)}$ are still in the linear
 1997 regime. Therefore, we have the following negative gradient for $[\mathbf{Q}_{4,4}]_{s,s}$, which only comes from
 1998 $\ell = 1$:

$$\begin{aligned} & \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \mathbf{logit}_{5,j'_1}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_1,r_{g_2 \cdot y_0},5}, e_{y_0} \rangle - \Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)} \right) \mathbb{1}_{\tau(x_0)=s} \right] \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \mathbf{logit}_{5,j'_1}^{(t)} \cdot 2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_0)=s} \right] \\ &= O(1) \cdot \mathbb{E} \left[\mathbf{logit}_{5,j'_1}^{(t)} \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_0)=s} \right] \end{aligned}$$

1999 On the other hand, we have

$$\begin{aligned} & \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{2,5}^{(2)} \right]_{s,s} \\ & \geq \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \left(\langle W_{5,j_2,r_{g_2 \cdot y_1},5}, e_{y_1} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \mathbb{1}_{\tau(x_1)=s} \right] \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}) \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_1)=s} \right] \\ & \geq \Omega(1) \cdot \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_1)=s} \right] \end{aligned}$$

2000 Since $\mathbf{logit}_{j'_1}^{(t)} \leq O(\frac{1}{d^{c_8}}) \cdot (1 - \mathbf{logit}_{j_2}^{(t)})$, we have that the negative gradient for $[\mathbf{Q}_{4,4}]_{s,s}$ is dominated
 2001 by $\left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{2,5}^{(2)} \right]_{s,s}$.

2002 When $\lambda(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}) > 1$, we have
 2003 $\mathbf{logit}_{j'_1}^{(t)} \leq O(\frac{1}{d}) \cdot (1 - \mathbf{logit}_{j_2}^{(t)})$. If $\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)}$ for $\ell = 1$ is still in the linear regime, we have

$$\Lambda_{5,j'_1,r_{g_2 \cdot y_0}}^{(t)} = (1 - 2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}) \cdot \left(F \pm \frac{1}{\text{poly}d} \right) \geq \varrho$$

2004 which implies

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \geq 1 - 2\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \geq 1 - 2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \geq \Omega\left(\frac{\varrho}{F}\right).$$

2005 We have the same upper bound for the negative gradient, and for $\left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{2,5}^{(2)} \right]_{s,s}$, we have

$$\begin{aligned} & \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{2,5}^{(2)} \right]_{s,s} \\ & \geq \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}) \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_1)=s} \right] \\ & \geq \Omega\left(\frac{1}{\text{poly}d}\right) \cdot \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_1)=s} \right] \\ & \geq \Omega\left(\frac{d}{\text{poly}d}\right) \cdot \mathbb{E} \left[\mathbf{logit}_{5,j'_1}^{(t)} \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_1)=s} \right] \end{aligned}$$

2006 which implies that the negative gradient for $[\mathbf{Q}_{4,4}]_{s,s}$ is also dominated by $\left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{2,5}^{(2)} \right]_{s,s}$.

2007 If $\Lambda_{5,j'_1}^{(t)}$ for $\ell = 1$ falls into the smoothed regime, we have

$$\begin{aligned}
& \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0} \cdot \mathbf{logit}_{5,j'_1}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j'_1}^{(t)}, r_{g_2 \cdot y_0}) \left(\langle W_{5,j'_1, r_{g_2 \cdot y_0}, 5}, e_{y_0} \rangle - \Lambda_{5,j'_1, r_{g_2 \cdot y_0}}^{(t)} \right) \mathbb{1}_{\tau(x_0)=s} \right] \\
& \leq O(1) \cdot \mathbb{E} \left[\mathbf{logit}_{5,j'_1}^{(t)} \cdot \left(\frac{(1 - 2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1})(F \pm \frac{1}{\text{poly}d})}{\varrho} \right)^{q-1} \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_0)=s} \right] \\
& = \mathbb{E} \left[O \left(\left(\frac{(1 - 2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1})(F \pm \frac{1}{\text{poly}d})}{\varrho} \right)^{q-1} \right) \mathbf{logit}_{5,j'_1}^{(t)} \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_0)=s} \right] \\
& \leq \mathbb{E} \left[O \left(\frac{(1 - 2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1})(F \pm \frac{1}{\text{poly}d})}{\varrho} \right) \mathbf{logit}_{5,j'_1}^{(t)} \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_0)=s} \right] \\
& \leq \mathbb{E} \left[O \left((1 - 2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}) \right) \mathbf{logit}_{5,j'_1}^{(t)} \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_0)=s} \right]
\end{aligned}$$

2008

$$\begin{aligned}
& \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{2,5}^{(2)} \right]_{s,s} \\
& \geq \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}) \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
& \geq \mathbb{E} \left[\Omega(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}) \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
& \geq \mathbb{E} \left[\Omega \left(d \cdot (1 - 2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}) \right) \cdot \mathbf{logit}_{5,j'_1}^{(t)} \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_1)=s} \right]
\end{aligned}$$

2009 Similarly, we can conclude that the negative gradient for $[\mathbf{Q}_{4,4}]_{s,s}$ is also dominated by $\left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{2,5}^{(2)} \right]_{s,s}$. Hence, to give a lower bound for $\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{\ell,5}^{(2)} \right]_{s,s}$, we only need to consider $\ell = 2$ and have

$$\begin{aligned}
& \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{2,5}^{(2)} \right]_{s,s} \\
& \geq \mathbb{E} \left[\Omega(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}) \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
& = \mathbb{E} \left[\Omega(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \left(F \pm \frac{1}{\text{poly}d} \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
& \geq \Omega \left(\frac{\epsilon \log d}{d} \cdot \frac{1}{d^{\lambda(1-2\epsilon)-1}} \right) = \Omega \left(\frac{\epsilon \log d}{d^{(1-2\epsilon)\lambda}} \right)
\end{aligned}$$

2012

□

2013 **Lemma F.19.** If Induction F.3 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,p}]_{s,j}$, $p \in \{3, 4\}$,
2014 $j \neq s \in \tau(\mathcal{X})$, $\ell \in [2]$, we have

$$\left| \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,j} \right| \leq O\left(\frac{1}{d}\right) \left| \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,\ell} \right]_{s,s} \right|.$$

2015 F.4.3 Non-negative Gap

2016 Let \tilde{T} denote the first time that $\Delta < \alpha$, where $\alpha = \frac{1}{\epsilon \frac{\alpha-2}{\text{poly}d}}$. Therefore, we have for $\ell = 2$,

2017 • For $[\mathbf{Q}_{4,3}]_{s,s}$,

$$\left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_{2,5}^{(2)} \right]_{s,s}$$

$$= \mathbb{E} \left[\underbrace{\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot (1 - \text{logit}_{5,j_2}^{(t)}) \cdot \text{sReLU}'(\Lambda_{5,j,r_{g_2 \cdot y_1}}^{(t)}) \left(\langle W_{5,j,r_{g_2 \cdot y_1},2}, e_{g_2} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right)}_{J_1} \mathbb{1}_{\tau(x_1)=s} \right. \\ \left. - \underbrace{\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \text{logit}_{5,j_2'}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_2',r_{g_2 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)} \right)}_{J_2} \mathbb{1}_{\tau(x_1)=s} \right]$$

2018 • For $[\mathbf{Q}_{4,4}]_{s,s}$ where $s \in \tau(\mathcal{X})$

$$\left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{2,5}^{(2)} \right]_{s,s} \\ = \mathbb{E} \left[\underbrace{\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot (1 - \text{logit}_{5,j_2}^{(t)}) \cdot \text{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \left(\langle W_{5,j_2,r_{g_2 \cdot y_1},5}, e_{y_1} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right)}_{J_3} \mathbb{1}_{\tau(x_1)=s} \right. \\ \left. - \underbrace{\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \text{logit}_{5,j_2'}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_2',r_{g_2 \cdot y_0},5}, e_{y_1} \rangle - \Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)} \right)}_{J_4} \mathbb{1}_{\tau(x_1)=s} \right]$$

$$J_4 - J_2 = O((\alpha \log d)^{q-1}) \cdot \mathbb{E} \left[(1 - \text{logit}_{5,j_2}^{(t)}) \cdot \mathbb{1}_{\tau(x_1)=s} \right] \\ J_1 - J_3 = O(\alpha \epsilon \log d) \mathbb{E} \left[(1 - \text{logit}_{5,j_2}^{(t)}) \cdot \mathbb{1}_{\tau(x_1)=s} \right]$$

2019 which implies $\left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} \right]_{s,s} \geq \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} \right]_{s,s}$. Also it is straightforward to see that
 2020 $\left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,1} \right]_{s,s} \geq \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,1} \right]_{s,s}$. Hence, we complete the proof.

2021 **F.4.4 Upper Bound for Q**

2022 Denote the first time that $[\mathbf{Q}_{4,3}]_{s,s}$ reaches $\Omega(\log^{1+c} d)$ for some small constant $c > 0$ as T^* . Then
 2023 we have

$$\text{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(T^*)} \leq O\left(\frac{1}{e^{\Omega(\log^{1+c} d)}}\right) \\ \text{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(T^*)} + \text{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(T^*)} \leq O\left(\frac{1}{e^{\Omega(\log^{1+c} d)}}\right)$$

2024 Moreover, we can simply bound the logits as follows:

$$1 - \text{logit}_{j_1}^{(t)} \leq O\left(\frac{1}{d^{\lambda(1-2\text{Attn}_{\text{ans},0 \rightarrow \text{pred},2}^{(T^*)})-1}}\right) \leq O\left(\frac{1}{d^{\lambda/2-1}}\right) \\ 1 - \text{logit}_{j_2}^{(t)} \leq O\left(\frac{1}{d^{\lambda(1-2\text{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(T^*)}-2\text{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(T^*)})-1}}\right) \leq O\left(\frac{1}{d^{\lambda/2-1}}\right)$$

2025 Thus, we have

$$\left| \sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_{\ell,5}^{(2)} \right]_{s,s} \right| \leq O\left(\frac{\log d}{e^{\Omega(\log^{1+c} d)} d^{\lambda/2}}\right)$$

2026 which implies

$$\mathbf{Q}_{4,3}^{(T)} \leq \mathbf{Q}_{4,3}^{(T^*)} + O\left(\frac{\eta T \log d}{e^{\Omega(\log^{1+c} d)} d^{\lambda/2}}\right) \leq \mathbf{Q}_{4,3}^{(T^*)} + O\left(\frac{\text{poly} d \cdot \log d}{e^{\Omega(\log^{1+c} d)} d^{\lambda/2}}\right) \leq O(\log^{1+c} d).$$

2027 **F.4.5 Attention Upper Bound**

2028 Let \tilde{T} denote the first time that $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}$ reaches $0.5 + \tilde{c}$. where $\tilde{c} > 0$ is some small
 2029 constant s.t., $2\tilde{c}\lambda > 1$. At \tilde{T} , we have $\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1}^{(\tilde{T})} \geq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})}$; $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})}$ and
 2030 $\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0}^{(\tilde{T})}$ is still at the constant level. For $\ell = 1$, j'_1 is no longer activate, while $\tau(y)$ for
 2031 $y \in \mathcal{Y} \setminus \{y_0\}$ has been activated to the linear regime. We have

2032 • For $\ell = 1$, for $y \in \mathcal{Y} \setminus \{y_0\}$, we have

$$\sum_{y \in \mathcal{Y} \setminus \{y_0\}} \mathbf{logit}_{5,\tau(g_1(y))}^{(t)} \geq \frac{\log d \cdot e^{2\tilde{c}\lambda \log d}}{\log d \cdot e^{2\tilde{c}\lambda \log d} + O(d)} (1 - \mathbf{logit}_{5,j_1}^{(t)}) = (1 - o(1)) \cdot (1 - \mathbf{logit}_{5,j_1}^{(t)})$$

2033 Thus,

$$\begin{aligned} & \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,1} \right]_{s,s} \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \left((1 - \mathbf{logit}_{5,j_1}^{(t)}) \cdot \left(\langle W_{5,j_1,r_{g_1 \cdot y_0},2}, e_{g_1} \rangle - \Lambda_{5,j_1,r_{g_1 \cdot y_0}}^{(t)} \right) \right. \right. \\ & \quad \left. \left. - \sum_{y \in \mathcal{Y} \setminus \{y_0\}} \mathbf{logit}_{5,\tau(g_1(y))}^{(t)} \cdot \left(\langle W_{5,\tau(g_1(y)),r_{g_1 \cdot y},2}, e_{g_1} \rangle - \Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)} \right) \right) \mathbf{1}_{\tau(x_0)=s} \right] \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},1} \cdot \left((1 - \mathbf{logit}_{5,j_1}^{(t)}) \cdot 2\mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2} \cdot (\lambda F \pm \frac{1}{\text{poly}d}) \right. \right. \\ & \quad \left. \left. - \sum_{y \in \mathcal{Y} \setminus \{y_0\}} \mathbf{logit}_{5,\tau(g_1(y))}^{(t)} \cdot 2(\mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},0} + \mathbf{Attn}_{\text{ans},0 \rightarrow \text{pred},2}) \cdot (\lambda F \pm \frac{1}{\text{poly}d}) \right) \mathbf{1}_{\tau(x_0)=s} \right] \\ &< 0 \end{aligned}$$

2034 • For $\ell = 2$, similarly, we have

$$\begin{aligned} \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5,\tau(g_2(y))}^{(t)} &\geq \frac{\log d \cdot e^{2\tilde{c}\lambda \log d}}{\log d \cdot e^{2\tilde{c}\lambda \log d} + O(d)} (1 - \mathbf{logit}_{5,j_2}^{(t)} - \mathbf{logit}_{5,j'_2}^{(t)}) \\ &= (1 - o(1)) \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)} - \mathbf{logit}_{5,j'_2}^{(t)}) \end{aligned}$$

2035

$$\begin{aligned} & \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_{2,5}^{(2)} \right]_{s,s} \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left((1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(\langle W_{5,j,r_{g_2 \cdot y_1},2}, e_{g_2} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \right. \right. \\ & \quad \left. \left. - \mathbf{logit}_{5,j'_2}^{(t)} \cdot \left(\langle W_{5,j'_2,r_{g_2 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} \right) \right) \right. \\ & \quad \left. - \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \cdot \left(\langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},2}, e_{g_2} \rangle - \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \right) \right) \mathbf{1}_{\tau(x_1)=s} \right] \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left((1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot 2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}) (\lambda F \pm \frac{1}{\text{poly}d}) \right. \right. \\ & \quad \left. \left. - \mathbf{logit}_{5,j'_2}^{(t)} \cdot 2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}) (\lambda F \pm \frac{1}{\text{poly}d}) \right) \right. \\ & \quad \left. - \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \cdot 2(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}) (\lambda F \pm \frac{1}{\text{poly}d}) \right) \\ & \quad \cdot \mathbf{1}_{\tau(x_1)=s} \Big] < 0 \end{aligned}$$

2036 Combing the two inequalities, we have $\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_{\ell,5}^{(2)} \right]_{s,s} < 0$. It is also clear that
 2037 $\sum_{\ell=1}^2 \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{\ell,5}^{(2)} \right]_{s,s} \geq 0$. Therefore, $\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}$ cannot further grow once it reaches
 2038 $0.5 + \tilde{c}$.

2039 F.4.6 Decreasing Gap at the End of Convergence

2040 Let \tilde{T} denote the first time that $1 - \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \leq 3\epsilon$, if $\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} -$
 2041 $\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \leq O\left(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d}\right)$, then we can let $T^* = \tilde{T}$ and stop the training. Otherwise,
 2042 we have $\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \geq \Omega\left(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d}\right)$. Following the similar argument
 2043 as in Lemma F.18, we have the gradient contribution from $\ell = 1$ is dominated by the gradient
 2044 contribution from $\ell = 2$. Thus, we focus on $\ell = 2$ and have the following logit relations: $\text{logit}_{j_2'}^{(t)} \geq$
 2045 $\Omega\left(\frac{1}{d}\right) \cdot (1 - \text{logit}_{j_2}^{(t)})$. Hence

$$\begin{aligned}
 & \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_{2,5}^{(2)} \right]_{s,s} \\
 &= \mathbb{E} \left[\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left((1 - \text{logit}_{5,j_2}^{(t)}) \cdot \left(\langle W_{5,j,r_{g_2 \cdot y_1},2}, e_{g_2} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \right. \right. \\
 & \quad \left. \left. - \text{logit}_{5,j_2'}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_2',r_{g_2 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)} \right) \right) \right. \\
 & \quad \left. - \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \text{logit}_{5,\tau(g_2(y))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) \left(\langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},2}, e_{g_2} \rangle - \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \Big] \\
 &\leq \mathbb{E} \left[(1 - \text{logit}_{5,j_2}^{(t)}) \cdot O(\epsilon \log d) \mathbb{1}_{\tau(x_1)=s} \right] \\
 & \\
 & \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{2,5}^{(2)} \right]_{s,s} \\
 &= \mathbb{E} \left[\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left((1 - \text{logit}_{5,j_2}^{(t)}) \cdot \left(\langle W_{5,j,r_{g_2 \cdot y_1},5}, e_{y_1} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \right. \right. \\
 & \quad \left. \left. - \text{logit}_{5,j_2'}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j_2',r_{g_2 \cdot y_0},5}, e_{y_1} \rangle - \Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)} \right) \right) \right. \\
 & \quad \left. - \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \text{logit}_{5,\tau(g_2(y))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) \left(\langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},5}, e_{y_1} \rangle - \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \Big] \\
 &\geq \mathbb{E} \left[\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left(-\text{logit}_{5,j_2'}^{(t)} \cdot \right. \right. \\
 & \quad \left. \left. \min\{\Omega(\log d), \Omega\left(\left((\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}) \log d\right)^{q-1} \log d\right)\}\right) \mathbb{1}_{\tau(x_1)=s} \right] \\
 &\geq \mathbb{E} \left[(1 - \text{logit}_{5,j_2}^{(t)}) \cdot \Omega\left(\frac{1}{d}\right) \cdot \right. \\
 & \quad \left. \min\{\Omega(\log d), \Omega\left(\left((\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}) \log d\right)^{q-1} \log d\right)\} \mathbb{1}_{\tau(x_1)=s} \right]
 \end{aligned}$$

Since $\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \geq \Omega\left(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d}\right)$, we have

$$((\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}) \log d)^{q-1} \geq d^{0.01}\epsilon,$$

2046 which implies for $\epsilon \ll \frac{1}{d}$, we have $\left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{2,5}^{(2)}\right]_{s,s} \gg \left| \left[-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_{2,5}^{(2)}\right]_{s,s} \right|$. Therefore, if the
 2047 gap does not decrease to the level of $O(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d})$, $\mathbf{Q}_{4,4}$ will start to dominately increase while $\mathbf{Q}_{4,3}$
 2048 will not change too much. On the other hand, if the gap of attention holds, then $[\mathbf{Q}_{4,3}]_{s,s} \geq [\mathbf{Q}_{4,4}]_{s,s}$,
 2049 we have

$$\begin{aligned} 1 - \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} &= \frac{O(1)}{O(1) + e^{[\mathbf{Q}_{4,4}]_{s,s}^{(t)}} + e^{[\mathbf{Q}_{4,3}]_{s,s}^{(t)}}} \\ &\geq \frac{O(1)}{O(1) + 2e^{[\mathbf{Q}_{4,3}]_{s,s}^{(t)}}} \geq \frac{1}{2} \cdot 3\epsilon > \epsilon. \end{aligned}$$

2050 This implies, we can find some time between \tilde{T} and $T_{2,3,s}$, s.t., the gap will decrease to $O(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d})$.
 2051 We denote this time as T^* and stop the training.

2052 F.4.7 At the End of the Training

2053 Putting everything together, we have that at the end of the training, we have

2054 **Lemma F.20.** *At $T^* = \tilde{O}(\frac{d^{(1-2\epsilon)\lambda}}{\eta\epsilon})$, if $\epsilon = o(\frac{1}{d^{1.01}})$, we have*

- 2055 • *Attention convergence: $\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(T_{2,3,s})} + \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(T_{2,3,s})} \leq \epsilon$, and $\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(T_{2,3,s})} -$*
 2056 *$\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(T_{2,3,s})} \leq O(\frac{(d^{1.01}\epsilon)^{\frac{1}{q-1}}}{\log d})$;*
- 2057 • *$[\mathbf{Q}_{4,p}^{(T_{2,3,s})}]_{j,s} \geq \Omega(\log d)$ for $p \in \{3, 4\}$ if $j = s \in \tau(\mathcal{X})$, otherwise, $[\mathbf{Q}_{4,p}^{(T_{2,3,s})}]_{j,s} \leq \tilde{O}(\frac{1}{d})$;*
- 2058 • *Loss convergence: $\sum_{\ell=1}^2 \text{Loss}_{\ell,5}^{(T_{2,3,s})} \leq \frac{1}{\text{poly}d}$.*

2059 F.5 Proof of Main Theorem

2060 **Theorem F.1** (Direct short-to-poly length generalization (Restatement)). *Under Assumptions 3.1,*
 2061 *3.2, 3.4, and 4.1, for every $L \leq \text{poly}d$, the transformer model $F^{(T_1+T_2)}$ obtained by Algorithm 1 with*
 2062 *learning rate $\eta = \frac{1}{\text{poly}(d)}$, and stage 1 and 2 iteration $T_1 = \tilde{O}(\frac{1}{\eta(\sigma_0)^{q-2}})$, $T_2 = \tilde{O}(\frac{\text{poly}(d)}{\eta\sigma_0})$ satisfies*

$$\text{Acc}_L(F^{(T_1+T_2)}) \geq 1 - \frac{1}{\text{poly}(d)},$$

2063 *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}^ℓ , $\ell \leq L$.*

2064 *Proof.* By Lemma F.20, at the end of Stage 2 training, we have $[\mathbf{Q}_{4,p}^{(T_{2,3,s})}]_{s,s} \geq \Omega(\log d)$ for all
 2065 $p \in \{3, 4\}$ and $s \in \tau(\mathcal{X})$.

2066 Therefore, for any $L \leq \text{poly}(d)$ in phase \mathcal{T}^L , we obtain

$$\epsilon_{\text{attn}}^{L,\ell} \leq \frac{O(1) \cdot L}{O(1) \cdot L + e^{[\mathbf{Q}_{4,3}^{(T_{2,3,s})}]_{s,s}} + e^{[\mathbf{Q}_{4,4}^{(T_{2,3,s})}]_{s,s}}} = o(1).$$

2067 Moreover, we have

$$\Delta^{L,\ell} \leq \Delta^{2,1} = \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(T_{2,3,s})} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(T_{2,3,s})} \leq o(1).$$

2068 These together guarantee that

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

2069 which implies

$$\text{Acc}_L(F^{(T^*)}) \geq 1 - \frac{1}{\text{poly}(d)}.$$

2070 □

2071 G Learning the Attention Layer: Symmetry Case for short-length

2072 Throughout the following discussion, for samples drawn from the LEGO distribution, we omit the
 2073 subscript $Z \sim \mathcal{D}$ in the expectation notation.

2074 G.1 Gradient Computations

2075 **Notations:** We first define the following notations. For any $i \in [d]$, $j \in [d]$, and $r \in [m]$, we define
 2076 the following quantities:

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

2077 **Fact G.1** (Gradients of \mathbf{Q}). For any $p, q \in [5]$, we have

$$\begin{aligned}& -\nabla_{\mathbf{Q}_{p,q}} \text{Loss}_i^{L,\ell} \\ &= \mathbb{E} \left[\sum_{\mathbf{k} \in \mathcal{I}^{L,\ell-1}} \text{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}} \cdot \left(\Xi_{\ell,i,\mathbf{k}}^L - \sum_{\mathbf{k}' \in \mathcal{I}^{L,\ell-1}} \text{Attn}_{\text{ans},\ell-1 \rightarrow \mathbf{k}'} \Xi_{\ell,i,\mathbf{k}'}^L \right) \mathbf{Z}_{\text{ans},\ell-1,p} \mathbf{Z}_{\mathbf{k},q}^\top \right],\end{aligned}$$

2078 where

$$\Xi_{\ell,i,\mathbf{k}}^L \triangleq \sum_{j \in [d]} \mathcal{E}_{i,j}(\mathbf{Z}^{L,\ell-1}) \sum_{r \in [m]} \text{sReLU}'(\Lambda_{i,j,r}(\mathbf{Z}^{L,\ell-1})) \langle \mathbf{W}_{i,j,r}, \mathbf{Z}_{\mathbf{k}} \rangle$$

2079 In the following discussion, letting $j_2 = \tau(g_2(y_1))$ and $j'_2 = \tau(g_2(y_0))$.

2080 **Lemma G.1.** For $s \in \tau(\mathcal{X})$ we have

$$\begin{aligned}-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{2,2} &= \mathcal{N}_{s,3,2,i} + \mathcal{N}_{s,3,2,ii} + \mathcal{N}_{s,3,2,iii} + \mathcal{N}_{s,3,2,iv} + \mathcal{N}_{s,3,2,v} + \mathcal{N}_{s,3,2,vi} \\ -\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{2,2} &= \mathcal{N}_{s,4,2,i} + \mathcal{N}_{s,4,2,ii} + \mathcal{N}_{s,4,2,iii} + \mathcal{N}_{s,4,2,iv} + \mathcal{N}_{s,4,2,v} + \mathcal{N}_{s,4,2,vi}\end{aligned}$$

2081 where

$$\begin{aligned}\mathcal{N}_{s,3,2,i} &= \mathbb{E} \left[\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot (1 - \text{logit}_{5,j_2}^{(t)}) \cdot \text{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \right. \\ &\quad \left. \left(\langle W_{5,j_2,r_{g_2 \cdot y_1},2}, e_{g_2} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\ \mathcal{N}_{s,3,2,ii} &= -\mathbb{E} \left[\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \text{logit}_{5,j'_2}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}) \right. \\ &\quad \left. \left(\langle W_{5,j'_2,r_{g_2 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s, y_0 \neq y_1} \right] \\ \mathcal{N}_{s,3,2,iii} &= \mathbb{E} \left[\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \text{logit}_{5,\tau(g_1(y_1))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}) \right. \\ &\quad \left. \left(\langle W_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1},2}, e_{g_2} \rangle - \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s, g_1(y_1) \neq g_2(y_1)} \right] \\ \mathcal{N}_{s,3,2,iv} &= \mathbb{E} \left[-\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \text{logit}_{5,\tau(g_1(y_0))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)}) \right.\end{aligned}$$

$$\begin{aligned}
& \left(\langle W_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s, g_1(y_0) \neq g_2(y_1)} \Big] \\
\mathcal{N}_{s,3,2,v} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left(- \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \cdot \right. \right. \\
& \quad \left. \left. \mathbf{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) \left(\langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},2}, e_{g_2} \rangle - \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \right) \mathbb{1}_{\tau(x_1)=s} \right]
\end{aligned}$$

$$\begin{aligned}
\mathcal{N}_{s,3,2,vi} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left(- \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5,\tau(g_1(y))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)}) \right. \right. \\
& \quad \left. \left. \left(\langle W_{5,\tau(g_1(y)),r_{g_1 \cdot y},2}, e_{g_2} \rangle - \Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \right) \mathbb{1}_{\tau(x_1)=s, g_1(y) \neq g_2(y_1)} \right]
\end{aligned}$$

2082

$$\begin{aligned}
\mathcal{N}_{s,4,2,i} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \right. \\
& \quad \left. \left(\langle W_{5,j_2,r_{g_2 \cdot y_1},5}, e_{y_1} \rangle - \Lambda_{5,j,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s} \right]
\end{aligned}$$

$$\begin{aligned}
\mathcal{N}_{s,4,2,ii} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left(- \mathbf{logit}_{5,\tau(g_2(y_0))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)}) \right. \right. \\
& \quad \left. \left. \cdot \left(\langle W_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0},5}, e_{y_1} \rangle - \Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \right) \mathbb{1}_{\tau(x_1)=s, y_0 \neq y_1} \right]
\end{aligned}$$

$$\begin{aligned}
\mathcal{N}_{s,4,2,iii} &= \mathbb{E} \left[- \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}) \right. \\
& \quad \left. \left(\langle W_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1},5}, e_{y_1} \rangle - \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s} \right]
\end{aligned}$$

$$\begin{aligned}
\mathcal{N}_{s,4,2,iv} &= \mathbb{E} \left[- \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \mathbf{logit}_{5,\tau(g_1(y_0))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)}) \right. \\
& \quad \left. \left(\langle W_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0},5}, e_{y_1} \rangle - \Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s, g_1(y_0) \neq g_2(y_1)} \right]
\end{aligned}$$

$$\begin{aligned}
\mathcal{N}_{s,4,2,v} &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left(- \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \right. \right. \\
& \quad \left. \left. \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) \left(\langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},5}, e_{y_1} \rangle - \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \right) \mathbb{1}_{\tau(x_1)=s} \right]
\end{aligned}$$

$$\mathcal{N}_{s,4,2,vi}$$

$$= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left(- \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5, \tau(g_1(y))}^{(t)} \mathbf{sReLU}'(\Lambda_{5, \tau(g_1(y)), r_{g_1 \cdot y}}^{(t)}) \right. \right. \\ \left. \left. \cdot \left(\langle W_{5, \tau(g_1(y)), r_{g_1 \cdot y}, 5}, e_{y_1} \rangle - \Lambda_{5, \tau(g_1(y)), r_{g_1 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \right) \mathbf{1}_{\tau(x_1)=s, g_1(y) \neq g_2(y_1)} \right]$$

2083 *Proof.* The proof follows from direct calculations based on Fact G.1. \square

2084 Let $\text{Loss}_{5,s}^{L,2} = -\mathbb{E} \left[\log p_F(\mathbf{Z}_{\text{ans},2,5} \mid \mathbf{Z}^{(L,1)}) \mid \tau(x_1) = s \right]$ for $s \in \tau(\mathcal{X})$. Due to the symmetry of
 2085 $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$ across $s \in \tau(\mathcal{X})$, we may, without loss of generality, focus on a particular
 2086 $s \in \tau(\mathcal{X})$ and analyze the corresponding loss $\text{Loss}_{5,s}^{L,2}$ in what follows. We further define some
 2087 events:

$$\begin{aligned} \mathcal{E}_1 &= \{y_0 \neq y_1\}, \\ \mathcal{E}_2 &= \{g_1(y_0) \neq g_2(y_1)\}, \\ \mathcal{E}_3 &= \{y \notin \{y_0, y_1\}, g_1(y) \neq g_2(y_1)\}. \end{aligned}$$

2088 G.2 Stage 1.2.1: Inital Growth of $\mathbf{Q}_{4,3}$

2089 The analysis of this stage is similar to the analysis in Appendix F.2, thus we will not repeat the details
 2090 and only present the main lemmas.

2091 **Induction G.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 $\frac{1}{B} = \Theta(\frac{1}{\log d})$.
 2092 For all iterations $t \leq T_{2,1,s}$, we have the following holds

- 2093 • $[\mathbf{Q}_{4,3}^{(t)}]_{s,s}$ monotonically increases;
- 2094 • $[\mathbf{Q}_{4,4}^{(t)}]_{s,s}$ monotonically decreases, $|[\mathbf{Q}_{4,4}^{(t)}]_{s,s}| \leq [\mathbf{Q}_{4,3}^{(t)}]_{s,s}$ and $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} - [\mathbf{Q}_{4,4}^{(t)}]_{s,s} =$
 2095 $\Theta([\mathbf{Q}_{4,3}^{(t)}]_{s,s})$;
- 2096 • for $p \in \{3, 4\}$, for $j \in \tau(\mathcal{X}) \neq s$, $|[\mathbf{Q}_{4,p}^{(t)}]_{s,j}| \leq O(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,j}}{d})$; otherwise $[\mathbf{Q}_{4,p}^{(t)}]_{s,j} = 0$

2097 G.2.1 Attention and Logit Preliminaries

2098 **Lemma G.2.** If Induction G.1 holds for all iterations $< t$, then we have

- 2099 1. $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \in [\frac{1}{4}, \frac{1}{4} + O(\frac{1}{\log d})]$;
- 2100 2. $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - O(\frac{1}{\log d}) \leq \mathbf{Attn}_{\text{ans},1 \rightarrow \mathbf{k}}^{(t)} \leq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}$ for $\mathbf{k} \in$
 2101 $\{(\text{pred}, 1), (\text{ans}, 0)\}$;
- 2102 3. $|\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)}| \leq \tilde{O}(\frac{1}{d})$;
- 2103 4. $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - O(\frac{1}{\log d}) \leq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \leq \mathbf{Attn}_{\text{ans},1 \rightarrow \mathbf{k}}^{(t)}$ for $\mathbf{k} \in$
 2104 $\{(\text{pred}, 1), (\text{ans}, 0)\}$.

2105 **Lemma G.3.** If Induction G.1 holds for all iterations $< t$, then we have, $\mathbf{logit}_{5,j}^{(t)} = O(\frac{1}{d})$ for all
 2106 $j \in \tau(\mathcal{Y})$.

2107 The proof of Lemma G.2 Lemma G.3 are similar to the proof of Lemma F.1 and Lemma F.2, and we
 2108 omit it here.

2109 G.2.2 Gradient Lemma

2110 **Lemma G.4.** *If Induction G.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,3}]_{s,s}$, we have*

$$\left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} \geq \Omega\left(\frac{1}{d} \cdot ([\mathbf{Q}_{4,3}]_{s,s} \cdot B)^{q-1}\right).$$

2111 **Lemma G.5.** *If Induction G.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,4}]_{s,s}$, we have*

$$-\Omega\left(\left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s}\right) \leq \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} \leq 0.$$

2112 **Lemma G.6.** *If Induction G.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,p}]_{s,j}$, $p \in \{3, 4\}$,
2113 $j \neq s \in \tau(\mathcal{X})$, we have*

$$\left| \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,j} \right| \leq O\left(\frac{1}{d}\right) \left| \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} \right|.$$

2114 G.2.3 At the End of Stage 1.2.1

2115 At the end of stage 1.2.1, we have

2116 **Lemma G.7.** *At $T_{2,1,s} = \Theta(\frac{d}{\eta\sigma_0^{q-2}}) + O(\frac{d}{\eta \log d})$ we have*

- 2117 • $[\mathbf{Q}_{4,3}^{(T_{2,1,s})}]_{s,s} \geq \Omega(\frac{1}{\log d})$, $[\mathbf{Q}_{4,4}^{(T_{2,1,s})}]_{s,s} \leq 0$;
- 2118 • $[\mathbf{Q}_{4,3}^{(T_{2,1,s})}]_{s,s} - [\mathbf{Q}_{4,4}^{(T_{2,1,s})}]_{s,s} \geq \Omega(\frac{1}{\log d})$;
- 2119 • *other $[\mathbf{Q}_{4,p}^{(T_{2,1,s})}]_{s,j}$ for $p \in \{3, 4\}$, $j \in \tau(\mathcal{X}) \neq s$ are at most $\tilde{O}(\frac{1}{d})$.*

2120 The consequence of Lemma G.7 is that j_2 is activated to the linear regime.

2121 G.3 Stage 1.2.2: Convergence with Small Wrong Attention

2122 Denote $\text{Loss}_{2,5,s}^{(2)} = -\mathbb{E}\left[\log p_F(\mathbf{Z}_{\text{ans},2,5} | \mathbf{Z}^{(2,1)}) | \tau(x_1) = s\right]$ for $s \in \tau(\mathcal{X})$.

2123 **Induction G.2.** *Given $s \in \tau(\mathcal{X})$, let $T_{2,2,s}$ denote the first time that $\text{Loss}_{2,5,s}^{(2)}$ decreases below
2124 $e^{-\frac{1}{2}B+1.52 \log d}$. For all iterations $t \leq T_{2,2,s}$, we have the following holds*

- 2125 • $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} + [\mathbf{Q}_{4,4}^{(t)}]_{s,s}$ monotonically increases;
- 2126 • $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} \geq [\mathbf{Q}_{4,4}^{(t)}]_{s,s}$;
- 2127 • *for $p \in \{3, 4\}$, for $j \in \tau(\mathcal{X}) \neq s$, $[\mathbf{Q}_{4,p}^{(t)}]_{s,j} \leq O(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,j}}{d})$; otherwise $[\mathbf{Q}_{4,p}^{(t)}]_{s,j} = 0$*

2128 G.3.1 Gradient Lemma

2129 **Lemma G.8.** *If Induction G.2 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $\tau(x_1) = s$, if
2130 $\text{Attn}_{\text{ans},1 \rightarrow \text{pred},2} < \frac{1}{2}$, we have*

$$\begin{aligned} & \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_{2,5}^{(2)} \right]_{s,s} + \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_{2,5}^{(2)} \right]_{s,s} \\ & \geq \Omega(1) \cdot \mathbb{E} \left[(1 - \text{logit}_{5,j_2}^{(t)}) \cdot \left(1 - \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) B \mathbb{1}_{\tau(x_1)=s} \right] \end{aligned}$$

2131 *Proof.* By Lemma G.1, we have

$$\mathcal{N}_{s,3,2,i} + \mathcal{N}_{s,4,2,i}$$

$$\begin{aligned}
&= \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \right. \\
&\quad \left(\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \langle W_{5,j_2,r_{g_2 \cdot y_1},2}, e_{g_2} \rangle + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \langle W_{5,j_2,r_{g_2 \cdot y_1},5}, e_{y_1} \rangle \right) \right. \\
&\quad \left. \left. - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
&= \mathbb{E} \left[(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \right. \\
&\quad \left(- \underbrace{\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot \langle W_{5,j_2,r_{g_2 \cdot y_1},2}, e_{g_1} \rangle + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot \langle W_{5,j_2,r_{g_2 \cdot y_1},5}, e_{y_0} \rangle \right)}_{J_{i,1}} \right. \\
&\quad \left. \left. + \underbrace{\left(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \right) \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}}_{J_{i,2}} \right) \mathbb{1}_{\tau(x_1)=s} \right]
\end{aligned}$$

2132 $J_{i,1} \geq 0$ since $\langle W_{5,j,r_{g_2 \cdot y_1},2}, e_{g_1} \rangle, \langle W_{5,j_2,r_{g_2 \cdot y_1},5}, e_{y_0} \rangle \leq 0$ on the event $\{g_1(y_1) \neq g_2(y_1)\} \cap \mathcal{E}_1$.
2133 Notice that on the event $\{g_1(y_1) \neq g_2(y_1)\}^c \cup \mathcal{E}_1^c$, the correct \mathbf{logit}_{5,j_2} is very close to 1 and the
2134 probability of this event is also small, so we can ignore it.

$$\begin{aligned}
&\mathcal{N}_{s,3,2,ii} + \mathcal{N}_{s,4,2,ii} \\
&= \mathbb{E} \left[-\mathbf{logit}_{5,j'_2}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}) \right. \\
&\quad \left(\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \langle W_{5,j'_2,r_{g_2 \cdot y_0},2}, e_{g_2} \rangle + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \langle W_{5,j'_2,r_{g_2 \cdot y_0},5}, e_{y_1} \rangle \right) \right. \\
&\quad \left. \left. - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
&= \mathbb{E} \left[\mathbf{logit}_{5,j'_2}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}) \right. \\
&\quad \left(\underbrace{\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot \langle W_{5,j'_2,r_{g_2 \cdot y_0},2}, e_{g_1} \rangle + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot \langle W_{5,j'_2,r_{g_2 \cdot y_0},5}, e_{y_0} \rangle \right)}_{J_{ii,1}} \right. \\
&\quad \left. \left. - \underbrace{\left(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \right) \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}}_{J_{ii,2}} \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_1} \right]
\end{aligned}$$

2135 $J_{ii,1} \approx 0$, since $\langle W_{5,j'_2,r_{g_2 \cdot y_0},2}, e_{g_1} \rangle + \langle W_{5,j'_2,r_{g_2 \cdot y_0},5}, e_{y_0} \rangle \approx \epsilon_{\text{cancel}}$ and $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} =$
2136 $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}$

$$\begin{aligned}
&\mathcal{N}_{s,3,2,iii} + \mathcal{N}_{s,4,2,iii} \\
&= \mathbb{E} \left[-\mathbf{logit}_{5,\tau(g_1(y_1))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}) \right. \\
&\quad \left(\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \langle W_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1},2}, e_{g_2} \rangle + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \langle W_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1},5}, e_{y_1} \rangle \right) \right. \\
&\quad \left. \left. - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)} \right) \mathbb{1}_{\tau(x_1)=s} \right]
\end{aligned}$$

$$\begin{aligned}
&= \mathbb{E} \left[\mathbf{logit}_{5,g_1(y_1)}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}) \right. \\
&\quad \left(\underbrace{\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot \langle W_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1},2}, e_{g_1} \rangle + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot \langle W_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1},5}, e_{y_0} \rangle \right)}_{J_{iii,1}} \right) \\
&\quad \left. - \underbrace{\left(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \right) \Lambda_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1}}^{(t)}}_{J_{iii,2}} \right) \mathbb{1}_{\tau(x_1)=s} \Big]
\end{aligned}$$

$$\begin{aligned}
&J_{iii,1} \geq \Omega(1) \cdot \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} B, \quad \text{since} \quad |\langle W_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1},5}, e_{y_0} \rangle| \leq \\
&O\left(\frac{1}{\log d}\right) \langle W_{5,\tau(g_1(y_1)),r_{g_1 \cdot y_1},2}, e_{g_1} \rangle \\
&\mathcal{N}_{s,3,2,iv} + \mathcal{N}_{s,4,2,iv} \\
&= \mathbb{E} \left[-\mathbf{logit}_{5,\tau(g_1(y_0))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)}) \right. \\
&\quad \left(\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \cdot \langle W_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0},2}, e_{g_2} \rangle + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \cdot \langle W_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0},5}, e_{y_1} \rangle \right) \right. \\
&\quad \left. - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \right) \Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)} \right) \mathbb{1}_{\tau(x_1)=s} \Big] \\
&= \mathbb{E} \left[\mathbf{logit}_{5,g_1(y_0)}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)}) \right. \\
&\quad \left(\underbrace{\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \cdot \langle W_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0},2}, e_{g_1} \rangle + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} \cdot \langle W_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0},5}, e_{y_0} \rangle \right)}_{J_{iv,1}} \right) \\
&\quad \left. - \underbrace{\left(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \right) \Lambda_{5,\tau(g_1(y_0)),r_{g_1 \cdot y_0}}^{(t)}}_{J_{iv,2}} \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_2} \Big]
\end{aligned}$$

2139 Similarly, $J_{iv,1} \geq \Omega(1) \cdot \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} B$.

$$\begin{aligned}
&\mathcal{N}_{s,3,2,v} + \mathcal{N}_{s,4,2,v} \\
&= \mathbb{E} \left[- \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) \right. \\
&\quad \left(\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},2}, e_{g_2} \rangle + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},5}, e_{y_1} \rangle \right) \right. \\
&\quad \left. - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \right) \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \right) \mathbb{1}_{\tau(x_1)=s} \Big] \\
&= \mathbb{E} \left[\sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) \right. \\
&\quad \left(\underbrace{\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},2}, e_{g_1} \rangle + \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},1} \langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},5}, e_{y_0} \rangle \right)}_{J_{v,1}} \right) \\
&\quad \left. - \underbrace{\left(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \right) \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}}_{J_{v,2}} \right) \mathbb{1}_{\tau(x_1)=s} \Big]
\end{aligned}$$

For $y \notin \{y_0, y_1\}$, on the event $\{g_1(y) \neq g_2(y)\}$

$$\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \approx (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}) \cdot 2B$$

$$\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} \approx (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1})B$$

When $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} \leq \frac{1}{2}$, we have $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} = \frac{1}{2}(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}) + (\frac{1}{2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2})$, which implies

$$\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \approx \left(\frac{1}{2}(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}) - \left(\frac{1}{2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}\right)\right)B.$$

2140 Thus, for $y \notin \{y_0, y_1\}$, we can conclude $\mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \leq O(\frac{1}{d})(1 - \mathbf{logit}_{5,j_2}^{(t)})$ if
 2141 $(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1})$ is small or $\mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \leq O(\frac{1}{d^c})(1 - \mathbf{logit}_{5,j_2}^{(t)})$
 2142 for some small constant, if $(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1})$ is large. Hence, while
 2143 $\langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},5}, e_{y_0} \rangle \leq -\Omega(B)$, the logit will be small. Moreover, on the event $\{g_1(y) = g_2(y)\}$,
 2144 $\langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},5}, e_{y_0} \rangle + \langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},5}, e_{g_1} \rangle = \epsilon_{\text{cancel}} \approx 0$

$$\begin{aligned} & \mathcal{N}_{s,3,2,vi} + \mathcal{N}_{s,4,2,vi} \\ &= \mathbb{E} \left[- \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5,\tau(g_1(y))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)}) \right. \\ & \quad \left(\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \langle W_{5,\tau(g_1(y)),r_{g_1 \cdot y},2}, e_{g_2} \rangle + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \langle W_{5,\tau(g_1(y)),r_{g_1 \cdot y},5}, e_{y_1} \rangle \right) \right. \\ & \quad \left. \left. - \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) \Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)} \right) \mathbb{1}_{\tau(x_1)=s} \right] \\ &= \mathbb{E} \left[- \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5,\tau(g_1(y))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)}) \right. \\ & \quad \left(\underbrace{\left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1} \langle W_{5,\tau(g_1(y)),r_{g_1 \cdot y},2}, e_{g_1} \rangle + \mathbf{Attn}_{\text{ans},0 \rightarrow \text{ans},1} \langle W_{5,\tau(g_1(y)),r_{g_1 \cdot y},5}, e_{y_0} \rangle \right)}_{J_{vi,1}} \right. \\ & \quad \left. \left. - \underbrace{\left(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} \right) \Lambda_{5,\tau(g_1(y)),r_{g_1 \cdot y}}^{(t)}}_{J_{vi,2}} \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{\mathcal{E}_3} \right] \end{aligned}$$

$$\begin{aligned} 2145 \quad J_{vi,1} &\geq \Omega(1) \cdot \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0} B, \quad \text{since} \quad |\langle W_{5,\tau(g_1(y)),r_{g_1 \cdot y},5}, e_{y_0} \rangle| \leq \\ 2146 \quad &O\left(\frac{1}{\log d}\right) \langle W_{5,\tau(g_1(y)),r_{g_1 \cdot y},2}, e_{g_1} \rangle \end{aligned}$$

2147 Clearly, $J_{\kappa,2} \kappa \in \{ii, iii, iv, v, vi\}$ can be dominated by $J_{i,2}$. Combining with the above analysis on
 2148 $J_{\kappa,1}$, we can conclude that the summation of gradient is dominated by i term and thus conclude the
 2149 lemma.

2150 □

2151 **Lemma G.9.** *If Induction G.2 holds for all iterations $< t$, we have*

$$[\mathbf{Q}_{4,3}^{(t)}]_{s,s} \geq [\mathbf{Q}_{4,4}^{(t)}]_{s,s}.$$

2152 *Proof.* When $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} = [\mathbf{Q}_{4,4}^{(t)}]_{s,s}$, we have $\mathcal{N}_{s,3,2,i} > 0$ dominate $[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2}]_{s,s}$, while
 2153 $\mathcal{N}_{s,4,2,i} < 0$ dominate $[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,2}]_{s,s}$ since $\mathcal{N}_{s,4,2,ii} = 0$. □

2154 **Lemma G.10.** *If Induction G.2 holds for all iterations $< t$, we have*

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \leq \frac{1.001 \log d}{B} = c_2.$$

2155 *wher $c_2 > 0$ is some sufficiently small constant since $B / \log d$ is sufficiently large.*

2156 *Proof.* Let \tilde{T} denote the first time $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \geq \frac{1.0005 \log d}{B}$, then

$$\mathbf{logit}_{5,j_2}^{(\tilde{T})} = \frac{e^{\frac{1.0005 \log d}{B} \cdot B}}{e^{\frac{1.0005 \log d}{B} \cdot B} + O(d)} (1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})}) = (1 - O(\frac{1}{d^{0.0005}}))(1 - \mathbf{logit}_{5,j_2}^{(\tilde{T})})$$

2157 thus, we have

$$\begin{aligned} [-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2}]_{s,s} &\leq \mathcal{N}_{s,3,2,i} + \mathcal{N}_{s,3,2,ii} \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left((1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(\langle W_{5,j_2,r_{g_2 \cdot y_1},2}, e_{g_2} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \right. \right. \\ &\quad \left. \left. - (1 - O(\frac{1}{d^{0.0005}}))(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(\langle W_{5,j_2',r_{g_2 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left((1 - O(\frac{1}{d^{0.0005}}))(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(\Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)} - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \right. \right. \\ &\quad \left. \left. + O(\frac{1}{d^{0.0005}})(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(\langle W_{5,j_2,r_{g_2 \cdot y_1},2}, e_{g_2} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \end{aligned}$$

2158 on the other hand, we have

$$\begin{aligned} [-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{(2)}]_{s,s} &\geq \mathcal{N}_{s,4,2,i} + \mathcal{N}_{s,4,2,ii} \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left((1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(\langle W_{5,j_2,r_{g_2 \cdot y_1},5}, e_{y_1} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \right. \right. \\ &\quad \left. \left. - (1 - O(\frac{1}{d^{0.0005}}))(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \left(\langle W_{5,j_2',r_{g_2 \cdot y_0},5}, e_{y_1} \rangle - \Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)} \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\ &= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left((1 - O(\frac{1}{d^{0.0005}}))(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \right. \right. \\ &\quad \left. \left(- \langle W_{5,j_2',r_{g_2 \cdot y_0},5}, e_{y_1} \rangle + \Lambda_{5,j_2',r_{g_2 \cdot y_0}}^{(t)} - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \right) \right. \\ &\quad \left. \left. - O(\frac{1}{d^{0.0005}})(1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \Lambda_{5,j_2,r_{g_2 \cdot y_0}}^{(t)} \right) \mathbb{1}_{\tau(x_1)=s} \right] \end{aligned}$$

2159 Since at time \tilde{T} , we have $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},1}^{(\tilde{T})} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})} \leq \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(\tilde{T})} -$
 2160 $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(\tilde{T})} \leq \frac{1.001 \log d}{B}$, we have $\Lambda_{5,j_2,r_{g_2 \cdot y_0}}^{(t)} - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \ll -\langle W_{5,j_2',r_{g_2 \cdot y_0},5}, e_{y_1} \rangle$. There-
 2161 fore,

$$[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_5^{2,2}]_{s,s} \leq [-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_5^{2,2}]_{s,s}$$

2162 which means the attention gap is not increasing. \square

2163 G.3.2 At the End of Training

2164 **Lemma G.11** (At the end of stage 1.2). *Induction G.2 holds for all iterations $T_{2,1,s} < t \leq T_{2,2,s} =$
 2165 $O(\frac{\text{poly}(d)}{\eta})$. At the end of training, we have*

2166 • *Attention concentration:* $\epsilon_{\text{attn}}^2 \leq c_1$ for some small constant $0 < c_1 < 0.01$;

2167 • *Loss convergence:* $\text{Loss}_{5,s}^{2,2} \leq e^{-B+4.008 \log d} = \frac{1}{\text{poly}d}$.

2168 *Proof.* Lemma G.8 guarantees the growth of $[\mathbf{Q}_{4,3}]_{s,s}$ until $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}$ reaches $\frac{1}{2}$. Notice that
 2169 if $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}$ reaches $\frac{1}{2} - \frac{1}{\log d}$, the training is still not finished, we have

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \geq (\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}) \left(1 - O\left(\frac{1}{\log d}\right)\right) 2B$$

2170 Furthermore, we have

$$\max_{y \neq y_1} \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \leq (\Delta^2 + \frac{\epsilon_{\text{attn}}^2}{2}) \cdot \left(1 - O\left(\frac{1}{\log d}\right)\right) 2B$$

2171 which implies

$$\begin{aligned} \text{Loss}_{5,s}^{2,2} &\leq e^{\left(\frac{1.001 \log d}{B} - \left(\frac{1}{2} - \frac{3}{\log d} - \frac{1.001 \log d}{B}\right)\right) \cdot 2B} \\ &\leq e^{-B+4.005 \log d} \end{aligned}$$

2172 which is a contradiction and we have the existence of stopping time of training. Moreover, at the time
 2173 of stopping, we have

$$\begin{aligned} \epsilon_{\text{attn}}^2 + \Delta^2 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} &\leq -\frac{1}{2} + \frac{2.004 \log d}{B}, \\ \Rightarrow \frac{3}{2} \epsilon_{\text{attn}}^2 + \frac{1}{2} \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) &\leq \frac{2.004 \log d}{B}, \\ \Rightarrow \epsilon_{\text{attn}}^2 &\leq \frac{1.336 \log d}{B}. \end{aligned}$$

2174 Letting $c_1 = \frac{1.336 \log d}{B}$, and $T_1 = T_{2,2,s}$ then we complete the proof of the stage 1.2. \square

2175 H Recursive Learning the Attention Layer: Symmetric Case

2176 At $L_k = 2^k$ with $k \geq 1$, the attention error $\epsilon_{\text{attn}}^{L_{k-1}} \leq c_1$ is already controlled at a small constant
 2177 level from the previous stage. Therefore, when transitioning from 2^{k-1} to 2^k , the incorrect attention
 2178 cannot increase significantly, and we have $\epsilon_{\text{attn}}^{L_k} \leq 2\epsilon_{\text{attn}}^{L_{k-1}}$, which remains small.

2179 Moreover, the attention gap $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}$ continues to decrease. As a result,
 2180 we can directly enter Stage 1.2.2 of convergence as in \mathcal{T}^2 .

2181 H.1 Preliminaries

2182 We first present preliminaries on the recursive learning attention layer, including its gradient compu-
 2183 tations and some useful probability lemmas.

2184 H.1.1 Gradient Computations

2185 **Fact H.1** (Gradients of \mathbf{Q}). Given $F^{(T_{k-1})}$ with $\epsilon_{\text{attn}}^{L_{k-1}} < c_1$ and $\Delta^{L_{k-1}} < c_2$ for $k \geq 2$, for
 2186 $(p, q) \in \{(4, 3), (4, 4)\}$, we have

$$-\nabla_{\mathbf{Q}_{p,q}} \text{Loss}_{F^{(T_{k-1})},5}^{L_k,2} = -\nabla_{\mathbf{Q}_{p,q}} \text{Loss}_5^{L_k,2}.$$

2187 **Lemma H.1.** Given $F^{(T_{k-1})}$ with $\epsilon_{\text{attn}}^{L_{k-1}} < c_1$ and $\Delta^{L_{k-1}} < c_2$ for $k \geq 2$, for $s \in \tau(\mathcal{X})$ we have

$$\begin{aligned} [-\nabla_{\mathbf{Q}_{4,3}} \text{Loss}_{F^{(T_{k-1})},5}^{L_k,2}]_{s,s} &= \mathcal{N}_{s,3,L_k,i} + \mathcal{N}_{s,3,L_k,ii} + \mathcal{N}_{s,3,L_k,iii} \\ [-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_{F^{(T_{k-1})},5}^{L_k,2}]_{s,s} &= \mathcal{N}_{s,4,L_k,i} + \mathcal{N}_{s,4,L_k,ii} + \mathcal{N}_{s,4,L_k,iii} \end{aligned}$$

2188 *where*

$$\begin{aligned}
& \mathcal{N}_{s,3,L,i} \\
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \right. \\
&\quad \left. \left(\langle W_{5,j_2,r_{g_2 \cdot y_1},2}, e_{g_2} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
& \mathcal{N}_{s,3,L,ii} \\
&= -\mathbb{E} \left[\mathbb{1}_{y_0 \neq y_1} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \mathbf{logit}_{5,j'_2}^{(t)} \cdot \right. \\
&\quad \left. \mathbf{sReLU}'(\Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)}) \left(\langle W_{5,j'_2,r_{g_2 \cdot y_0},2}, e_{g_2} \rangle - \Lambda_{5,j'_2,r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
& \mathcal{N}_{s,3,L,iii} \\
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left(- \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \cdot \right. \right. \\
&\quad \left. \left. \mathbf{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) \left(\langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},2}, e_{g_2} \rangle - \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \right) \mathbb{1}_{\tau(x_1)=s} \right]
\end{aligned}$$

2189

$$\begin{aligned}
& \mathcal{N}_{s,4,L,i} \\
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot (1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \right. \\
&\quad \left. \mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \left(\langle W_{5,j_2,r_{g_2 \cdot y_1},5}, e_{y_1} \rangle - \Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
& \mathcal{N}_{s,4,L,ii} \\
&= \mathbb{E} \left[\mathbb{1}_{y_0 \neq y_1} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left(- \mathbf{logit}_{5,\tau(g_2(y_0))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)}) \right. \right. \\
&\quad \left. \left. \cdot \left(\langle W_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0},5}, e_{y_1} \rangle - \Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\
& \mathcal{N}_{s,4,L,iii} \\
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \cdot \left(- \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \right. \right. \\
&\quad \left. \left. \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) \left(\langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},5}, e_{y_1} \rangle - \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \right) \mathbb{1}_{\tau(x_1)=s} \right]
\end{aligned}$$

2190 *Proof.*

- For $[\mathbf{Q}_{4,3}]_{s,s}$ where $s \in \tau(\mathcal{X})$

$$\begin{aligned}
& \left[- \nabla_{\mathbf{Q}_{4,3}} \text{Loss}_5^{L,2} \right]_{s,s} \\
&= \mathbb{E} \left[\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \cdot \left((1 - \mathbf{logit}_{5,j_2}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)}) \right. \right. \\
&\quad \left. \left. \left(\langle W_{5,j_L,r_{g_2 \cdot y_1},2}, e_{g_2} \rangle - \Lambda_{5,j_L,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \right) \right]
\end{aligned}$$

$$\begin{aligned}
& - \mathbb{1}_{y_0 \neq y_1} \mathbf{logit}_{5, \tau(g_2(y_0))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5, \tau(g_2(y_0)), r_{g_2 \cdot y_0}}^{(t)}) \\
& \quad \left(\langle W_{5, \tau(g_2(y_0)), r_{g_2 \cdot y_0}, 2, e_{g_2}} \rangle - \Lambda_{5, \tau(g_2(y_0)), r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \\
& - \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5, \tau(g_2(y))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5, \tau(g_2(y)), r_{g_2 \cdot y}}^{(t)}) \\
& \quad \left(\langle W_{5, \tau(g_2(y)), r_{g_2 \cdot y}, 2, e_{g_2}} \rangle - \Lambda_{5, \tau(g_2(y)), r_{g_2 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s} \Big]
\end{aligned}$$

2191

• For $[\mathbf{Q}_{4,4}]_{s,s}$ where $s \in \tau(\mathcal{X})$

$$\begin{aligned}
& \left[-\nabla_{\mathbf{Q}_{4,4}} \text{Loss}_5^{L,2} \right]_{s,s} \\
& = \mathbb{E} \left[\mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{ans}, 1}^{(t)} \cdot \left((1 - \mathbf{logit}_{5, j_L}^{(t)}) \cdot \mathbf{sReLU}'(\Lambda_{5, j_L, r_{g_2 \cdot y_1}}^{(t)}) \right. \right. \\
& \quad \left. \left(\langle W_{5, j_L, r_{g_2 \cdot y_1}, 5, e_{y_1}} \rangle - \Lambda_{5, j_L, r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \right. \\
& - \mathbb{1}_{y_0 \neq y_1} \mathbf{logit}_{5, \tau(g_2(y_0))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5, \tau(g_2(y_0)), r_{g_2 \cdot y_0}}^{(t)}) \\
& \quad \left. \left(\langle W_{5, \tau(g_2(y_0)), r_{g_2 \cdot y_0}, 5, e_{y_1}} \rangle - \Lambda_{5, \tau(g_2(y_0)), r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \right. \\
& - \sum_{y \in \mathcal{Y} \setminus \{y_0, y_1\}} \mathbf{logit}_{5, \tau(g_2(y))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5, \tau(g_2(y)), r_{g_2 \cdot y}}^{(t)}) \\
& \quad \left. \left. \left(\langle W_{5, \tau(g_2(y)), r_{g_2 \cdot y}, 5, e_{y_1}} \rangle - \Lambda_{5, \tau(g_2(y)), r_{g_2 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \right) \mathbb{1}_{\tau(x_1)=s} \right]
\end{aligned}$$

2192 Combing with Fact H.1, we complete the proof. \square

2193 H.1.2 Probabilistic Lemmas

2194 We first define some events:

$$\begin{aligned}
\mathcal{E}_{L,1}(\mathbf{Q}) &= \left\{ \sum_{\ell \in [L]} \mathbf{Attn}_{\text{ans}, 1 \rightarrow \text{pred}, \ell} \mathbb{1}_{g_\ell \cdot y = g_2 \cdot y} \geq \frac{1}{2} \right\} \\
\mathcal{E}_{L,2} &= \left\{ \max_{y \in \mathcal{Y}} \left(\sum_{\ell \in [L], \ell \neq 2} \mathbb{1}_{g_\ell \cdot y = g_2 \cdot y} \right) < K_L \right\} \\
\mathcal{E}_{L,3} &= \left\{ \max_{y \neq y_1} \left(\sum_{\ell \in [L], \ell \neq 2} \mathbb{1}_{g_\ell \cdot y = g_2 \cdot y} \right) \geq \Omega(L) \right\} \\
\mathcal{E}_{L,4,\ell} &= \left\{ \max_{y \neq y_1} \left(\sum_{\ell \in [L], \ell \neq 2} \mathbb{1}_{g_\ell \cdot y = g_2 \cdot y} \right) = \ell \right\}
\end{aligned}$$

2195 where

$$U_L = \begin{cases} \Theta \left(\frac{\log n_y}{\log \left(\frac{4n_y}{L} \cdot \log n_y \right)} \right) & L \leq n_y \log n_y, \\ \Theta \left(\frac{L}{n_y} \right) & L \geq n_y \log n_y. \end{cases}$$

2196 **Lemma H.2** (Maximum load of balls and bins [75]). Suppose we sequentially throw m balls into n
2197 bins, where each ball is thrown uniformly at random into one of the bins. Let X_i denote the load of
2198 bin i , $1 \leq i \leq n$. Then, with probability at least $1 - n^{-\Omega(1)}$ the maximum load of any bin satisfies

$$\max_{i \in [n]} X_i < \begin{cases} \Theta \left(\frac{\log n}{\log \left(\frac{4n}{m} \cdot \log n \right)} \right) & m \leq n \log n, \\ \frac{m}{n} + \Theta \left(\sqrt{\frac{m}{n}} \cdot \log n \right) & m \geq n \log n. \end{cases}$$

2199 H.2 Reducing the Wrong Attention

2200 Let $\text{Loss}_{5,s}^{L,2} = -\mathbb{E}\left[\log p_F(\mathbf{Z}_{\text{ans},2,5} \mid \mathbf{Z}^{(L,1)}) \mid \tau(x_1) = s\right]$ for $s \in \tau(\mathcal{X})$. Due to the symmetry
 2201 of $[\mathbf{Q}_{4,3}]_{s,s}$ and $[\mathbf{Q}_{4,4}]_{s,s}$, we may, without loss of generality, focus on a particular $s \in \tau(\mathcal{X})$ and
 2202 analyze the corresponding loss $\text{Loss}_{5,s}^{L,2}$ in what follows.

2203 **Induction H.1.** Given $s \in \tau(\mathcal{X})$, at $L_k = 2^k$ with $k \geq 2$, let T_k denote the first time that $\text{Loss}_{5,s}^{L_k,2}$
 2204 decreases below $e^{-B+4.008 \log d}$. For all iterations $t \leq T_k$, we have the following holds

- 2205 • $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} + [\mathbf{Q}_{4,4}^{(t)}]_{s,s}$ monotonically increases;
- 2206 • $[\mathbf{Q}_{4,3}^{(t)}]_{s,s} \geq [\mathbf{Q}_{4,4}^{(t)}]_{s,s}$;
- 2207 • for $p \in \{3, 4\}$, for $j \in \tau(\mathcal{X}) \neq s$, $[\mathbf{Q}_{4,p}^{(t)}]_{s,j} \leq O(\frac{[\mathbf{Q}_{4,p}^{(t)}]_{s,j}}{d})$; otherwise $[\mathbf{Q}_{4,p}^{(t)}]_{s,j} = 0$

2208 **Lemma H.3.** If Induction H.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $\tau(x_1) = s$, if
 2209 $\frac{1}{2} - \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \geq \frac{0.001 \log d}{B}$, we have

$$\begin{aligned} & \left[-\nabla_{\mathbf{Q}_{4,3}^{(t)}} \text{Loss}_{F^{T_k-1},5}^{L_k,2} \right]_{s,s} + \left[-\nabla_{\mathbf{Q}_{4,4}^{(t)}} \text{Loss}_{F^{T_k-1},5}^{L_k,2} \right]_{s,s} \\ & \geq \Omega(1) \cdot \mathbb{E} \left[(1 - \text{logit}_{5,y_2}^{(t)}) \cdot (1 - \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}) B \mathbb{1}_{\tau(x_1)=s} \right] \end{aligned}$$

2210 *Proof.* By Lemma H.1, we have

$$\begin{aligned} & \mathcal{N}_{s,3,L_k,i} + \mathcal{N}_{s,4,L_k,i} \\ & = \mathbb{E} \left[(1 - \text{logit}_{5,y_2}^{(t)}) \cdot \text{sReLU}'(\Lambda_{5,y_2,r_{g_2 \cdot y_1}}^{(t)}) \right. \\ & \quad \cdot \left(\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \langle W_{5,y_2,r_{g_2 \cdot y_1},5}, e_{y_1} \rangle + \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \langle W_{5,y_2,r_{g_2 \cdot y_1},2}, e_{g_2} \rangle \right. \\ & \quad \left. \left. - (\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} + \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,y_2,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s} \right] \\ & = \mathbb{E} \left[(1 - \text{logit}_{5,j_2}^{(t)}) \cdot \text{sReLU}'(\Lambda_{5,y_2,r_{g_2 \cdot y_1}}^{(t)}) \right. \\ & \quad \cdot \left(-\sum_{\ell \neq 2} \text{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \langle W_{5,y_2,r_{g_2 \cdot y_1},\ell}, e_{g_\ell} \rangle - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \langle W_{5,y_2,r_{g_2 \cdot y_1},5}, e_{y_0} \rangle \right. \\ & \quad \left. \left. + (1 - \text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,y_2,r_{g_2 \cdot y_1}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s} \right]. \end{aligned}$$

$$\begin{aligned} & \mathcal{N}_{s,3,L_k,ii} + \mathcal{N}_{s,4,L_k,ii} \\ & = \mathbb{E} \left[-\text{logit}_{5,\tau(g_2(y_0))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)}) \right. \\ & \quad \cdot \left(\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \langle W_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0},5}, e_{y_0} \rangle + \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} \langle W_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0},2}, e_{g_2} \rangle \right. \\ & \quad \left. \left. - (\text{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} + \text{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{g_2(y_0) \neq g_2(y_1)} \right] \\ & = \mathbb{E} \left[\text{logit}_{5,\tau(g_2(y_0))}^{(t)} \cdot \text{sReLU}'(\Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)}) \right. \end{aligned}$$

$$\cdot \left(\sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \langle W_{5,j'_2,r_{g_2 \cdot y_0},2}, e_{g_\ell} \rangle + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \langle W_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0},5}, e_{y_0} \rangle \right. \\ \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s} \mathbb{1}_{y_0 \neq y_1} \Bigg].$$

$$\mathcal{N}_{s,3,L_k,iii} + \mathcal{N}_{s,4,L_k,iii} \\ = \mathbb{E} \left[\sum_{y \notin \{y_0, y_1\}} \mathbf{logit}_{5,\tau(g_2(y))}^{(t)} \cdot \mathbf{sReLU}'(\Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}) \right. \\ \cdot \left(\sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},2}, e_{g_\ell} \rangle + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \langle W_{5,\tau(g_2(y)),r_{g_2 \cdot y},5}, e_{y_0} \rangle \right. \\ \left. - (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)}) \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)} \pm \tilde{O}(\sigma_0) \right) \mathbb{1}_{\tau(x_1)=s} \Bigg].$$

2211 The event $\mathcal{E}_{L_k,1}$ contributes a negative gradient, as $\Lambda_{5,y_2,r_{g_2 \cdot y_1}}^{(t)}$ reaches its maximum in this regime,
 2212 causing the corresponding positive gradient to vanish. Note that $Z^{(L_k,1)} \in \mathcal{E}_{L_k,1}$ implies that
 2213 $\sum_{\ell \in [L_k]} \mathbb{1}_{g_\ell \cdot y = g_2 \cdot y} \geq \Omega(L_k)$, and hence $\mathbb{P}(Z^{(L_k,1)} \in \mathcal{E}_{L_k,1}) \leq n_y^{-\Omega(L_k)}$. Moreover, observe that
 2214 $(1 - \mathbf{logit}_{5,y_2}) \mathbb{1}_{\mathcal{E}_{L_k,1}^c \cap \mathcal{E}_{L_k,4,\ell}} \geq (1 - \mathbf{logit}_{5,y_2}) \mathbb{1}_{\mathcal{E}_{L_k,1} \cap \mathcal{E}_{L_k,4,\ell}}$ for every $\ell \in [L_k]$. Therefore, the
 2215 negative gradient contributed by $\mathcal{E}_{L_k,1}$ is negligible, and it suffices to focus on the event $\mathcal{E}_{L_k,1}^c$.
 2216 When $L = O(1)$, with high probability we have $|\sum_{\ell \neq 2} \mathbb{1}_{g_\ell \cdot y_1 = g_2 \cdot y_1}| \leq 1$. Then:

$$- \sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \langle W_{5,j_2,r_{g_2 \cdot y_1},2}, e_{g_\ell} \rangle - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \langle W_{5,j_2,r_{g_2 \cdot y_1},5}, e_{y_0} \rangle \\ \geq -O\left(\frac{B}{\log d}\right) (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}).$$

2217 If $L = \omega(1)$, then by the maximum load of balls into bins in the sparse regime (Lemma H.2), we
 2218 can further restrict our attention to the event $\mathcal{E}_{L_k,1}^c \cap \mathcal{E}_{L_k,2}$. Moreover, by the bound on U_{L_K} in
 2219 Lemma H.2, we have

$$\frac{\sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \mathbb{1}_{g_\ell \cdot y_1 = g_2 \cdot y_1}}{\sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)}} \leq \max \left\{ \Theta\left(\frac{1}{L_k}\right), \Theta\left(\frac{1}{n_y}\right) \right\}.$$

2220 Therefore, we have

$$- \sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \langle W_{5,j_2,r_{g_2 \cdot y_1},2}, e_{g_\ell} \rangle - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \langle W_{5,j_2,r_{g_2 \cdot y_1},5}, e_{y_0} \rangle \\ \geq -O\left(\frac{B}{L_k}\right) \cdot (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}).$$

2221 Thus, $\mathcal{N}_{s,3,L,i} + \mathcal{N}_{s,4,L,i}$ is dominated by the term $(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} -$
 2222 $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \Lambda_{5,y_2,r_{g_2 \cdot y_1}}^{(t)}$.

2223 For $y \notin \{y_0, y_1\}$, let $\tilde{y} = \arg \max_{y \notin \{y_0, y_1\}} \Lambda_{5,\tau(g_2(y)),r_{g_2 \cdot y}}^{(t)}$, if this maximum is $\geq \Lambda_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0}}^{(t)}$,
 2224 we have:

$$\sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \langle W_{5,\tau(g_2(\tilde{y})),r_{g_2 \cdot \tilde{y}},2}, e_{g_\ell} \rangle + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \langle W_{5,\tau(g_2(\tilde{y})),r_{g_2 \cdot \tilde{y}},5}, e_{y_0} \rangle \\ \geq \sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \langle W_{5,\tau(g_2(\tilde{y})),r_{g_2 \cdot \tilde{y}},2}, e_{g_\ell} \rangle \mathbb{1}_{g_\ell(\tilde{y}) \neq g_2(\tilde{y})}$$

$$\geq -O\left(\frac{B}{\log d}\right) (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}),$$

2225 since $\langle W_{5,\tau(g_2(\tilde{y})),r_{g_2 \cdot \tilde{y}},2}, e_{g_\ell} \rangle \leq O(\frac{B}{\log d})$ is small when $g_\ell(\tilde{y}) \neq g_2(\tilde{y})$. Moreover, $\Lambda_{5,y_2,r_{g_2 \cdot y_1}}^{(t)} -$
 2226 $\Lambda_{5,\tau(g_2(\tilde{y})),r_{g_2 \cdot \tilde{y}}}^{(t)} \geq \Omega(1) \cdot \Lambda_{5,y_2,r_{g_2 \cdot y_1}}^{(t)}$, so the dominant contribution in $\mathcal{N}_{s,3,L,ii} + \mathcal{N}_{s,4,L,ii}$ and
 2227 $\mathcal{N}_{s,3,L,iii} + \mathcal{N}_{s,4,L,iii}$ comes from $\tau(g_2(\tilde{y}))$, which can be bounded by the i term. If instead dominated
 2228 by $\tau(g_2(y_0))$, we similarly have:

$$\begin{aligned} & \sum_{\ell \neq 2} \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},\ell}^{(t)} \langle W_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0},2}, e_{g_\ell} \rangle + \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \langle W_{5,\tau(g_2(y_0)),r_{g_2 \cdot y_0},5}, e_{y_0} \rangle \\ & \geq -O\left(\frac{B}{\log d}\right) (1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)}). \end{aligned}$$

2229 In summary, the lower bound on the total gradient is primarily determined by the con-
 2230 tribution from $\mathcal{N}_{s,3,L,i} + \mathcal{N}_{s,4,L,i}$, which is in turn dominated by $(1 - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1} -$
 2231 $\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}) \Lambda_{5,y_2,r_{g_2 \cdot y_1}}^{(t)}$. This concludes the proof.

2232 □

2233 **Lemma H.4.** *If Induction H.1 holds for all iterations $< t$, we have*

$$[\mathbf{Q}_{4,3}^{(t)}]_{s,s} \geq [\mathbf{Q}_{4,4}^{(t)}]_{s,s}.$$

2234 **Lemma H.5.** *If Induction H.1 holds for all iterations $< t$, we have*

$$\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},0}^{(t)} \leq \frac{1.001 \log d}{B}.$$

2235 **Lemma H.6.** *If Induction H.1 holds for all iterations $< t$, given $s \in \tau(\mathcal{X})$, for $[\mathbf{Q}_{4,p}]_{s,j}$, $p \in \{3,4\}$,
 2236 $j \neq s \in \tau(\mathcal{X})$, we have*

$$\left| \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,j} \right| \leq O\left(\frac{1}{d}\right) \left| \left[-\nabla_{\mathbf{Q}_{4,p}^{(t)}} \text{Loss}_5^{2,2} \right]_{s,s} \right|.$$

2237 **Lemma H.7** (At the end of stage). *Induction H.1 holds for all iterations $T_{k-1} < t \leq T_k =$
 2238 $O(\frac{\log L_k \cdot \text{poly}(d)}{\eta})$. At the end of stage k , we have*

- 2239 • *Attention concentration:* $\epsilon_{\text{attn}}^{L_k} \leq c_1$ for some small constant $0 < c_1 < 0.01$;
- 2240 • *Loss convergence:* $\text{Loss}_{F(T_{k-1}),5}^{L_k,2} \leq e^{-B+4.008 \log d} = \frac{1}{\text{poly} d}$.

2241 *Proof.* Lemma H.3 guarantees the growth of either $[\mathbf{Q}_{4,3}]_{s,s}$ or $[\mathbf{Q}_{4,4}]_{s,s}$ until the attention weight
 2242 satisfies

$$\frac{1}{2} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} < \frac{0.001 \log d}{B}.$$

2243 At this point, we have

$$\Lambda_{5,j_2,r_{g_2 \cdot y_1}}^{(t)} \geq \left(\frac{1}{2} - \frac{0.001 \log d}{B} - \frac{\epsilon_{\text{attn}}^{L_k}}{L_k} \right) \cdot \left(1 - O\left(\frac{1}{\log d}\right) \right) \cdot 2B.$$

- 2244 • If $L_k \ll O(\frac{\log d}{\log \log d})$, then the event $\mathcal{E}_{L_k,3}$ holds with probability $\geq \omega(\frac{1}{d})$, which is not
 2245 negligible. In this case, we need to consider the following bound on the maximum of the
 2246 wrong logits

$$\begin{aligned} & \max_{y \neq y_1} \Lambda_{5,j_2,r_{g_2 \cdot y}}^{(t)} \\ & \leq \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} + \frac{(L_k - 1)}{L_k} \epsilon_{\text{attn}}^{L_k} \right) \cdot \left(1 - O\left(\frac{1}{\log d}\right) \right) \cdot 2B \end{aligned}$$

$$\leq \left(\frac{2L_k - 1}{L_k} \Delta^{L_k} + \frac{2(L_k - 1)}{L_k} \frac{0.001 \log d}{B} \right) \cdot \left(1 - O\left(\frac{1}{\log d}\right) \right) \cdot 2B$$

2247

Hence, the loss satisfies

$$\text{Loss}_{5,s}^{L_k,2} \leq \exp \left(\left(-1 + \frac{4.006 \log d}{B} \right) \cdot \left(1 - O\left(\frac{1}{\log d}\right) \right) \cdot B \right),$$

2248

which implies that training has effectively converged, and a stopping time T_k exists. $T_k = O(\frac{\log L_k \cdot \text{poly}(d)}{\eta})$ can be directly derived from Lemma H.3. Furthermore, at the stopping

2249

time, we have the following gurantee for the wrong attention weights:

2250

$$\begin{aligned} \epsilon_{\text{attn}}^{L_k} + \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} &\leq -\frac{1}{2} + \frac{2.004 \log d}{B}, \\ \Rightarrow \frac{3}{2} \epsilon_{\text{attn}}^{L_k} + \frac{1}{2} \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) &\leq \frac{2.004 \log d}{B}, \\ \Rightarrow \epsilon_{\text{attn}}^{L_k} &\leq \frac{1.336 \log d}{B}. \end{aligned}$$

2251

- If $L_k \geq \Omega(\frac{\log d}{\log \log d})$, then the loss on the event $\mathcal{E}_{L_k,3}$ is negligible. In this case, we can focus on the event $\mathcal{E}_{L_k,2}$, and have

2252

$$\begin{aligned} \max_{y \neq y_1} \Lambda_{5,j_2,r_{g_2 \cdot y}}^{(t)} &\leq \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} + O\left(\frac{1}{n_y}\right) \epsilon_{\text{attn}}^{L_k} \right) \cdot \left(1 - O\left(\frac{1}{\log d}\right) \right) \cdot 2B \\ &\leq \left(\frac{1.001 \log d}{B} + \frac{0.001 \log d}{B} \right) \cdot \left(1 - O\left(\frac{1}{\log d}\right) \right) \cdot 2B. \end{aligned}$$

2253

Hence, the loss satisfies

$$\text{Loss}_{5,s}^{L_k,2} \leq \exp \left(\left(-1 + \frac{2.004 \log d}{B} \right) \cdot \left(1 - O\left(\frac{1}{\log d}\right) \right) \cdot B \right),$$

2254

which also implies the existence of a stopping time for stage k . Furthermore, at the stopping time, we have the inequality

2255

$$\begin{aligned} O\left(\frac{1}{n_y}\right) \epsilon_{\text{attn}}^{L_k} + \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} &\leq -\frac{1}{2} + \frac{2.004 \log d}{B}, \\ \Rightarrow \left(\frac{1}{2} + O\left(\frac{1}{n_y}\right) \right) \epsilon_{\text{attn}}^{L_k} + \frac{1}{2} \left(\mathbf{Attn}_{\text{ans},1 \rightarrow \text{pred},2}^{(t)} - \mathbf{Attn}_{\text{ans},1 \rightarrow \text{ans},1}^{(t)} \right) &\leq \frac{2.004 \log d}{B}, \\ \Rightarrow \epsilon_{\text{attn}}^{L_k} &\leq \frac{4.008 \log d}{B}. \end{aligned}$$

2256

Letting $c_1 = \frac{4.008 \log d}{B}$, we have $\epsilon_{\text{attn}}^{L_k} \leq c_1$ for some small constant $c_1 > 0$ since $B/\log d$ is a large enough constant.

2257

2258

□

2259

H.3 Proof of Main Theorem

Theorem H.1 (Recursive self-training (Restatement)). *Assume the distribution \mathcal{D}^L induced from $\mathbf{LEGO}(\mathcal{X}, \mathcal{G}, \mathcal{Y})$ satisfies Assumption 3.1, 3.2 and 4.2, and assume the transformer network satisfies Assumption 3.3 and 3.4. 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 $T_k = O(\frac{\text{poly}(d)}{\eta})$, is able to solve the task $\mathcal{T}^{L_{k+1}}$, $L_{k+1} = 2^{k+1}$ with accuracy:*

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

2260

Proof. By Lemma H.7, at the time T_k , we have $\epsilon_{\text{attn}}^{L_k} \leq c_1$, combining with Induction H.1, non-

2261

diagonal entry of $\mathbf{Q}_{p,q}$ remains close to 0, thus moving to the next stage, we have $\epsilon_{\text{attn}}^{L_{k+1},\ell} \leq 4\epsilon_{\text{attn}}^{L_k} \leq$

2262

$4c_1 < 0.04$ for all $\ell < L_{k+1}$. Moreover, $\Delta^{L_{k+1},\ell} \leq \Delta^{L_k,1}$. Hence, we have

$$\mathbb{E}_{Z^{(L_{k+1})} \sim \mathcal{D}^{L_{k+1}}} \left[\mathbb{E}_{\widehat{Z}_{\text{ans},\ell+1} \sim \mathcal{P}_F^{(T_k)}(\cdot | Z^{(L_{k+1},\ell)})} [\mathbb{1}_{\{\widehat{Z}_{\text{ans},\ell+1} \neq Z_{\text{ans},\ell+1}\}}] \right]$$

$$\begin{aligned}
&\leq O(1) \cdot \mathbb{E}_{Z^{(L_{k+1})} \sim \mathcal{D}^{L_{k+1}}} [1 - \mathbf{logit}_{5, \tau(\mathbf{Z}_{\text{ans}, \ell+1, 5})}] \\
&\leq O(1) \cdot e^{\left(-(\frac{1}{2} - 2c_1) + \Delta^{L_{k+1}, \ell} + 4c_1\right) \cdot 2B} \\
&\leq O(1) \cdot e^{(-1 + 2 \times 7c_1)B} = O\left(\frac{1}{\text{poly}d}\right).
\end{aligned}$$

2263 Thus $\text{Acc}_{L_{k+1}}(F^{T_k}) \geq 1 - O\left(\frac{1}{\text{poly}d}\right)$.

□