In-context Convergence of Transformers

Yu Huang [1] **Yuan Cheng** [2] **Yingbin Liang** [3]

Abstract

Transformers have recently revolutionized many domains in modern machine learning and one salient discovery is their remarkable in-context learning capability, where models can solve an unseen task by utilizing task-specific prompts without further parameters fine-tuning. This also inspired recent theoretical studies aiming to understand the in-context learning mechanism of transformers, which however focused only on *linear* transformers. In this work, we take the first step toward studying the learning dynamics of a one-layer transformer with *softmax* attention trained via gradient descent in order to in-context learn linear function classes. We consider a structured data model, where each token is randomly sampled from a set of feature vectors in either balanced or imbalanced fashion. For data with balanced features, we establish the finite-time convergence guarantee with near-zero prediction error by navigating our analysis over two phases of the training dynamics of the attention map. More notably, for data with imbalanced features, we show that the learning dynamics take a stage-wise convergence process, where the transformer first converges to a near-zero prediction error for the query tokens of dominant features, and then converges later to a near-zero error for query tokens of under-represented features, via one and four training phases. Our proof features new techniques for analyzing the competing strengths of two types of attention weights, the change of which determines different training phases.

1. Introduction

Transformers (Vaswani et al., 2017) have emerged as the foundational architectures in various domains, including natural language processing (Devlin et al., 2018; OpenAI, 2023), computer vision (Dosovitskiy et al., 2020; He et al., 2022), reinforcement learning (Chen et al., 2021; Janner et al., 2021), and so on. Recently, large language models (LLMs) based on transformers have exhibited remarkable in-context learning capabilities, where the model can solve a new task solely through inference based on prompts of the task without further fine-tuning (Brown et al., 2020).

Such striking abilities have inspired a recent line of research to understand the underlying mechanisms of in-context learning from various aspects (Garg et al., 2022; Min et al., 2022; Wei et al., 2023; Von Oswald et al., 2023; Xie et al., 2021). Among these studies, the pioneering work of Garg et al. (2022) empirically studied in-context learning via an interpretable framework, highlighting the capacity of transformers to acquire in-context knowledge of linear and some more complex function classes. Specifically, they showed that an in-context trained model over a function class $\mathcal{F}$ can accurately predict the function value $f(x_{\text{query}})$ of a new query token x_{query} for most $f \in \mathcal{F}$ by using a prompt sequence including in-context input-label pairs along with the query token $(x_1, f(x_1), \ldots, x_N, f(x_N), x_{\text{query}})$.

Built on this theoretically amenable setting, many follow-up works explored theoretical properties of in-context learning of transformers from different perspectives such as expressive power (Akyürek et al., 2022; Giannou et al., 2023), generalization (Li et al., 2023b), internal mechanisms (Von Oswald et al., 2023; Bai et al., 2023), etc. Specially, a few recent studies (Zhang et al., 2023a; Mahankali et al., 2023; Ahn et al., 2023) made interesting progress towards understanding the training dynamics of transformers for in-context learning[1]. However, those studies focused only on 'linear' transformers, and does not capture the crucial role of the 'softmax' mapping, which lies in the core design of transformers to be advantageous over other network architectures. Therefore, the following fundamental problem still remains largely open:

*How do **softmax**-based transformers trained via gradient descent learn in-context?*

This paper takes the first step toward addressing this problem by investigating the learning dynamics of a single-layer

[1]Department of Statistics and Data Science, Wharton School, University of Pennsylvania, Philadelphia, PA, USA [2]National University of Singapore, Singapore [3] Department of Electrical and Computer Engineering, The Ohio State University, Columbus, OH, USA. Correspondence to: Yingbin Liang <liang.889@osu.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024.

[1]More detailed discussions for related work can be found in Appendix A.

transformer with *softmax* attention trained by gradient descent (GD) for in-context learning. We focus on the setting with training prompts generated from linear regression models as in Garg et al. (2022), and with structured input data, where each token is randomly selected from a set of feature vectors $\{v_k\}_{k=1}^K$ with probability $\{p_k\}_{k=1}^K$, respectively. We then train the transformer over the squared loss of prediction error using GD. We study the training dynamics under both balanced and imbalanced feature distributions, and characterize the in-context learning ability for both settings. We highlight our contributions as follows.

Our Contributions.

- We first establish the convergence guarantee for the setting with balanced features, where $p_k = \Theta(\frac{1}{K})$ for each $k \in [K]$, and characterize the training evolution of the attention map into a two-phase dynamic process. In phase I, for each $k \in [K]$, the parameters of the self-attention module undergo fast growth, aligning the query token featuring v_k with input tokens featuring v_k rapidly disregarding other feature directions. In phase II, the loss of prediction error converges to a near-minimum value.
- We then prove the convergence for the setting with imbalanced features, where one feature dominates, say v_1 with $p_1 = \Theta(1)$, while others are under-represented with $p_k = \Theta(\frac{1}{K})$ for $k > 1$, which serves as a remarkable showcase of the in-context learning capabilities of transformers. We show that the learning dynamics takes a *stage-wise* convergence process. Initially, the transformer quickly attains near-zero prediction error for query tokens of dominant features, and then converges to near-zero prediction error for query tokens of under-represented features, irrespective of their infrequent occurrence.
- Our analysis hinges on a novel proof technique that characterizes the *softmax* attention dynamics via the interplay between two types of bilinear attention weights: 'weight of query token and its target feature' and 'weight of query token and off-target features'. Which weight plays a dominant role in the attention dynamics can change over the learning process, resulting in different training phases. Our analysis tools may be of independent interest and hold the potential to study various other problems involving transformers.

Notations. We let $[K] := \{1, 2, \ldots, K\}$. We use capital letters for matrices (e.g., A), and lowercase letters for vectors and scalars (e.g., a). For a matrix A, we use A_i to represent the i-th column of A and $A_{i:j}$ to indicate a collection of columns spanning from i to j. We use $\mathbf{1}\{\cdot\}$ to denote the indicator function. We use $O(K)$, $\Omega(K)$, and $\Theta(K)$ to omit universal constants concerning the variable K. We use $\mathrm{poly}(K)$ and $\mathrm{polylog}(K)$ to denote large constant-degree polynomials of K and $\log(K)$, respectively. Given $h(x) \leq 0$ and $g(x) > 0$, we denote $h(x) = -\Omega(g(x))$ if there exists some constant $C_1 > 0$ and a_1, s.t. $|h(x)| \geq C_1 g(x)$ for all $x \geq a_1$; $h(x) = -O(g(x))$ if there exist some constant $C_2 > 0$ and a_2, s.t. $|h(x)| \leq C_2 g(x)$ for all $x \geq a_2$; $h(x) = \Theta(g(x))$ if there exists some constant $C_3, C_4 > 0$ and a_3, s.t. $C_3 g(x) \leq |h(x)| \leq C_4 g(x)$ for all $x \geq a_3$.

2. Problem Setup

In this section, we present our problem formulations, including the in-context learning framework, one-layer transformer architecture, and the training settings we consider.

2.1. In-Context Learning Framework

We adopt the well-established in-context learning framework in Garg et al. (2022). The objective is to enable the training of models capable of in-context learning within a specified function class $\mathcal{F}$, where the functions and input data are sampled respectively by the distributions $D_{\mathcal{F}}$ and $D_{\mathcal{X}}$. Specifically, the process is initiated by generating random training prompts as follows. We first sample a random function f from the class according to the distribution $D_{\mathcal{F}}$. We then create a set of random inputs $x_1, \ldots, x_N$ and query x_{query} , all drawn independently by $D_{\mathcal{X}}$. Finally, we compute the value of function f on these inputs to construct the prompt $P = (x_1, y_1, \ldots, x_N, y_N, x_{\text{query}})$, where $y_i = f(x_i)$. The goal for an in-context learner is to use the prompt to form a prediction $\widehat{y}\,(x_{\text{query}})$ for the query such that $\widehat{y}\,(x_{\text{query}}) \approx f\,(x_{\text{query}})$.

Task Distribution. In this work, our focus is on the task of linear functions defined as $\mathcal{F} = \left\{f : \mathcal{X} \to \mathbb{R} \mid f(x) = \langle w, x\rangle \text{ with } w \in \mathbb{R}^d, \mathcal{X} \subset \mathbb{R}^d\right\}$, which is widely adopted in recent studies for in-context learning (Ahn et al., 2023; Zhang et al., 2023a; Mahankali et al., 2023). For each prompt, the task-specific weight w is independently drawn from a task distribution $\mathcal{D}_\Omega$ with zero mean and identity covariance matrix $\mathbf{I}_{d\times d}$.

Data Distribution $\mathcal{D}_{\mathcal{X}}$. We consider a set of distinct features $\{v_k \in \mathbb{R}^d, k = 1, \ldots, K\}$, where all features are orthonormal vectors. Each data point x is sampled from the feature set with the probability p_k for sampling v_k, where $p_k \in (0, 1)$ for $k \in [K]$ and $\sum_{k\in[K]} p_k = 1$. Such a data model has been widely employed in the theoretical studies of deep learning, including ensemble methods (Allen-Zhu & Li, 2020), multi-modal learning (Huang et al., 2022), vision transformers (Li et al., 2023a), etc.

2.2. One-Layer Transformer Architecture

To present the one-layer transformer model we consider, we first introduce the self-attention mechanism (Bahdanau

et al., 2014; Vaswani et al., 2017) for the transformer model.

Definition 2.1 (Self-Attention (SA) Mechanism)**.** A self-attention layer (Bahdanau et al., 2014; Vaswani et al., 2017) in the single-head case with width d_e consists of the following components: a key matrix $W^{\text{Key}} \in \mathbb{R}^{d_e \times d_e}$, a query matrix $W^Q \in \mathbb{R}^{d_e \times d_e}$, and a value matrix $W^V \in \mathbb{R}^{d_e \times d_e}$. Given a prompt P of length N, let $E \in \mathbb{R}^{d_e \times d_N}$ be an embedding matrix of the prompt P, and the self-attention mechanism will output:

$$F_{\text{SA}}\left(E; W^{\text{Key}}, W^Q, W^V\right) = W^V E \cdot \text{softmax}\left(\left(W^{\text{Key}} E\right)^\top W^Q E\right), \quad (1)$$

where the $\text{softmax}(\cdot)$ function is applied column-wisely, i.e., for a vector input z, the i-th entry of $\text{softmax}(z)$ is given by $e^{z_i}/\sum_s e^{z_s}$.

Embeddings. For in-context learning, given a prompt $P = (x_1, y_1, \ldots, x_N, y_N, x_{\text{query}})$, a natural token embedding is to stack $x_i \in \mathbb{R}^d$ and y_i into the first N columns. The final column consists of $x_{\text{query}} \in \mathbb{R}^d$ and 0. Formally,

$$E(P) = \begin{pmatrix} x_1 & x_2 & \cdots & x_N & x_{\text{query}} \\ y_1 & y_2 & \cdots & y_N & 0 \end{pmatrix} \in \mathbb{R}^{(d+1)\times(N+1)}.$$

Therefore, $d_N = N+1$ and $d_e = d+1$ in the above embedding. Let us further denote the first d rows of E as $E^x(P) \in \mathbb{R}^{d\times(N+1)}$ and the last row of E as $E^y(P) \in \mathbb{R}^{1\times(N+1)}$. Then we write $E(P) = \{E^x(P), E^y(P)\}$. We omit the dependency on P for $E(P)$, $E^x(P)$ and $E^x(P)$ when there is no ambiguity.

We next instantiate additional operations and certain parameter settings based on the general SA mechanism (1) for our one-layer transformer model to mitigate unnecessary complications in theoretical analysis while keeping the most critical component of the SA mechanism.

Masking. Let $M(\cdot)$ denote the masking operation, which masks (removes) the last column of the entry matrix. In other words, for a given matrix $A \in \mathbb{R}^{(d+1)\times(N+1)}$, $M(A)$ yields $A_{1:N} \in \mathbb{R}^{(d+1)\times N}$. We will first mask the embedding matrix E before its multiplication with the key matrix W^{Key} and the value matrix W^V, which results in $W^{\text{Key}}M(E)$ and $W^V M(E)$, in order to prevent the query token from attending to itself. This approach has been commonly taken in previous works (Tian et al., 2023; Mahankali et al., 2023; Von Oswald et al., 2023; Kitaev et al., 2020).

Reparameterization. We consolidate the query and key matrices into one matrix denoted as $W^{KQ} \in \mathbb{R}^{(d+1)\times(d+1)}$, often taken in recent theoretical frameworks (Zhang et al., 2023a; Jelassi et al., 2022; Tian et al., 2023). Furthermore, we consider W^V and W^{KQ} in the following specific forms:

$$W^V = \begin{pmatrix} 0_{d\times d} & 0_d \\ 0_d^\top & \nu \end{pmatrix}, \quad W^{KQ} = \begin{pmatrix} Q & 0_d \\ 0_d^\top & 0 \end{pmatrix}, \quad (2)$$

where $\nu \in \mathbb{R}$ and $Q \in \mathbb{R}^{d\times d}$. The above structures of W^V and W^{KQ} are inspired by the recent study (Zhang et al., 2023a), which showed that such structured matrices achieve the global optimum in the *linear* SA model. Furthermore, we set $\nu = 1$ (where ν is the only parameter in W^V) and do not update it during the training. The reason is twofold: 1) this aligns with the common practice in theoretical studies of deep learning, where the last linear layer is often kept fixed to focus on the analysis of hidden layers. Our objective remains highly nonconvex and challenging even with a fixed ν; and 2) the form of the global optimum outlined in recent work (Zhang et al., 2023a) suggests that for *linear* SA, the optimal solution for ν serves as a scaling factor to normalize the output of linear attention. In our case, the output of *softmax* attention is already inherently normalized.

Remark 2.2 (Nealy no loss of optimality). *Despite the specific form of $\{W^V, W^{KQ}\}$, the minimum of the loss function $L^* = \Theta(e^{-\text{poly}(K)})$ (as shown in Theorem 3.2) implies that such a specific form at most incurs an error of $\Theta(e^{-\text{poly}(K)})$ that vanishes exponentially with K, compared to the minimum loss over the general parameter space $\{W^V, W^{\text{Key}}, W^Q\}$. Therefore, for our nonlinear* softmax *SA, such specific parameterization does not lose optimality.*

With the aforementioned masking operations and reparameterization, the overall transformer model consisting of a single SA layer can be recast in the parameterization of $\theta = \{1, Q\}$ as follows:

$$F_{\text{SA}}(E; \theta) = M(E^y) \cdot \text{softmax}\left(M(E^x)^\top Q E^x\right). \quad (3)$$

Such a reparameterization separates the label E^y from the softmax operator while maintaining simultaneous processing of both input E^x and label E^y information. The prediction for the token x_{query} will be the last entry of F_{SA},

$$\widehat{y}_{\text{query}} = \widehat{y}_{\text{query}}(E; \theta) = [F_{\text{SA}}(E; \theta)]_{(N+1)}.$$

Henceforth, we may omit the reference to E and θ, and use $\widehat{y}_{\text{query}}$ if it is not ambiguous.

2.3. Training Settings

Loss Function. To train the transformer model F_{SA} over linear regression tasks, we minimize the following squared loss of the prediction error, which has also been taken by Zhang et al. (2023a); Ahn et al. (2023):

$$L(\theta) = \frac{1}{2}\mathbb{E}\left[\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right)^2\right] \quad (4)$$

where the expectation is taken with respect to the prompt P including input and query tokens $\{x_i\}_{i=1}^N \cup \{x_{\text{query}}\}$ and the weight vector w.

Training Algorithm. The above learning objective in eq.(4) is minimized via GD with the learning rate η. At $t = 0$, we initialize $Q^{(0)}$ as zero matrix $\mathbf{0}_{d\times d}$. The parameter is updated as follows:

$$\theta^{(t+1)} = \theta^{(t)} - \eta\nabla_\theta L(\theta^{(t)}).$$

3. Main Results

In this section, we characterize the convergence of in-context learning by GD for the settings with balanced and imbalanced features, respectively.

To measure the degree to which the query token x_{query} attends to the specific input token and to a certain class of features, we define the notions of the attention scores.

Definition 3.1 (Attention Score)**.** Given a prompt $P = (x_1, y_1, \cdots, x_N, y_N, x_{\text{query}})$ and its corresponding embedding E, where $\{x_i \in \mathbb{R}^d\}_{i=1}^N, x_{\text{query}}$ is drawn independently from $\mathcal{D}_\mathcal{X}$, then at time t, for F_{SA} with parameter $\theta^{(t)}$, we define the attention score as follows.

1. Given $i \in [N]$, the attention score for the i-th token x_i is
$$\mathbf{attn}_i(\theta^{(t)}; E) := \Big[\text{softmax}(M(E^x)^\top Q^{(t)} E^x)\Big]_i = \frac{e^{E_i^{x\top} Q^{(t)} E_{N+1}^x}}{\sum_{j\in[N]} e^{E_j^{x\top} Q^{(t)} E_{N+1}^x}}.$$
2. For $k \in [K]$, denote $\mathcal{V}_k(P) \subset [N]$ as the index set for input tokens, such that $x_i = v_k$ for $i \in \mathcal{V}_k(P)$. Then the attention score for the k-th feature is given by
$$\mathbf{Attn}_k(\theta^{(t)}; E) := \sum_{i\in\mathcal{V}_k(P)} \mathbf{attn}_i(\theta^{(t)}; E).$$

For simplicity, we represent $\mathbf{attn}_i(\theta^{(t)}; E)$ and $\mathbf{Attn}_k(\theta^{(t)}; E)$ as $\mathbf{attn}_i^{(t)}$ and $\mathbf{Attn}_k^{(t)}$, respectively, and denote $\mathcal{V}_k(P)$ as $\mathcal{V}_k$. We also rewrite the prediction output at time t as follows:

$$\widehat{y}_{\text{query}}^{(t)} = \sum_{i\in[N]} \mathbf{attn}_i^{(t)}\, y_i = \sum_{k\in[K]} \mathbf{Attn}_k^{(t)} \langle w, v_k\rangle. \tag{5}$$

3.1. In-Context Learning with Balanced Features

In this subsection, we study in-context learning with *balanced* features, where the probabilities of sampling all K features are in the same order, i.e., $p_k = \Theta(\frac{1}{K})$ for each $k \in [K]$. In such a setting, each feature appears equally likely in the prompt, ensuring their equal recognition. The following theorem characterizes the convergence of GD.

Theorem 3.2 (In-context Learning with Balanced Features)**.** *Suppose $p_k = \Theta(\frac{1}{K})$ for $k \in [K]$. For any $0 < \epsilon < 1$, suppose $N \geq \text{poly}(K)$ and $\text{polylog}(K) \gg \log(\frac{1}{\epsilon})$. We apply GD to train the loss function given in eq.(4). Then with at most $T^* = O(\frac{\log(K)K^2}{\eta} + \frac{K\log\left(K\epsilon^{-\frac{1}{2}}\right)}{\epsilon\eta})$ iterations, we have*

1. *The loss converges: $L(\theta^{(T^*)}) - L^* \leq \epsilon$, where $L^* = \Theta(e^{-\text{poly}(K)})$ is the global minimum of eq.(4).*
2. *Attention score concentrates: if $x_{query} = v_k$, then with probability at least $1 - e^{-\Omega(\text{poly}(K))}$[2],the one-layer transformer nearly "pays all attention" to input tokens featuring v_k, i.e., $(1 - \mathbf{Attn}_k^{(T^*)})^2 \leq O(\epsilon)$.*

Theorem 3.2 shows that training a one-layer transformer with softmax attention can converge to the minimum of the objective loss in the reparameterization space via GD, with polynomial time efficiency with respect to K and $\frac{1}{\epsilon}$. The learning dynamics for such a case with balanced features exhibit a **two-phase behavior**. (i) The first term of T^* captures the duration of phase I, where the network actively aligns the query token (suppose $x_{\text{query}} = v_k$) with those tokens featuring v_k itself, thus substantially increasing $\mathbf{Attn}_k^{(t)}$ to a constant level. (ii) The second term captures the duration of phase II, where the loss converges to the near-zero prediction error.

In-context Learning Ability. For the obtained model with $\theta^{(T^*)}$, let us evaluate a test prompt associated with a linear task w, which *might not be drawn from* the support of $\mathcal{D}_\Omega$ (i.e., w may not be present in the training process), but has its data drawn by $\mathcal{D}_\mathcal{X}$. Suppose the query token is $x_{\text{query}} = v_k$. Following from the attention score concentration principle in Theorem 3.2, eq.(5) yields that with high probability the query prediction $\widehat{y}_{\text{query}}^{(T^*)}$ is given by

$$\mathbf{Attn}_k^{(T^*)} \langle w, v_k\rangle + \sum_{m\neq k} \mathbf{Attn}_m^{(T^*)} \langle w, v_m\rangle \approx \langle w, v_k\rangle.$$

This implies that the in-context learned model can still well approximate the test prompt even if the task model w does not lie in the support of the training task distribution $\mathcal{D}_\Omega$ and was *unseen* during training. This showcases the remarkable in-context learning capability of trained transformers. We also highlight that the in-context learning mechanism characterized by our theorems has been verified by the empirical findings in many recent works trained with transformers at the GPT-2 (Radford et al., 2019) scale. For instance, in Yadlowsky et al. (2023), they showed that the in-context learning ability of transformers may be closely tied to the coverage of their pre-training data mixtures. This indeed aligns with our attention concentration principle, which demonstrates that the transformer can perform in-context learning by correctly capturing and identifying different types of features in the training data.

3.2. In-Context Learning with Imbalanced Features

In real-world datasets, skewed distributions are common, where a few classes or features dominate in data while others are under-represented. It is typically difficult to train models

[2]The randomness originates from the first N input tokens in the test prompt.

to perform well on features that have limited representation in those datasets (Cui et al., 2019; Chou et al., 2020). In this subsection, we investigate the setting with imbalanced features, where the dominant feature v_1 is sampled with the probability $p_1 = \Theta(1)$, and all other features are sampled with $p_k = \Theta(\frac{1}{K})$ for $2 \le k \le K$. We will show that somewhat remarkably, in-context learning is less sensitive to imbalanced features and can achieve a near-zero error even when the query token takes an under-represented feature.

To investigate the performance for the imbalanced case, we focus on the following prediction error for each feature v_k:

$$\mathcal{L}_k(\theta) = \frac{1}{2}\mathbb{E}\left[(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle)^2 \,|\, x_{\text{query}} = v_k\right]. \quad (6)$$

The following theorem identifies the convergence of GD.

Theorem 3.3 (In-context Learning with Imbalanced Features)**.** *Suppose $p_1 = \Theta(1)$ and $p_k = \Theta(\frac{1}{K})$ for $2 \le k \le K$. For any $0 < \epsilon < 1$, suppose $N \ge \text{poly}(K)$, and $\text{polylog}(K) \gg \log(\frac{1}{\epsilon})$. We apply GD to train the loss function given in eq.(4). Then the following results hold.*

1. *The prediction error for the **dominant** feature converges: for v_1, with at most $T_1 = O(\frac{\log(\epsilon^{-\frac{1}{2}})}{\eta\epsilon})$ GD iterations, $\mathcal{L}_1(\theta^{(T_1)}) \le \mathcal{L}_1^* + \epsilon$, where $\mathcal{L}_1^* = \Theta(e^{-\text{poly}(K)})$ is the global minimum of eq.(6) for $k = 1$;*
2. *The prediction error for the **under-represented** features converges: for v_k with $2 \le k \le K$, with at most $T_k = O(\frac{\log(K)K^2}{\eta} + \frac{K\log\left(K\epsilon^{-\frac{1}{2}}\right)}{\epsilon\eta})$ GD iterations, $\mathcal{L}_k(\theta^{(T_k)}) \le \mathcal{L}_k^* + \epsilon$, where $\mathcal{L}_k^* = \Theta(e^{-\text{poly}(K)})$ is the global minimum of eq.(6);*
3. *Attention score concentrates: for each $k \in [K]$, if the query token is v_k, then after T_k iterations, with probability at least $1 - e^{-\Omega(\text{poly}(K))}$, the one-layer transformer nearly "pays all attention" to input tokens featuring v_k: $(1 - \mathbf{Attn}_k^{(T_k)})^2 \le O(\epsilon)$.*

Theorem 3.3 shows that the GD dynamics of the in-context training exhibit *'stage-wise'* convergence. The trained transformer rapidly (within T_1) converges to a model that achieves a near-zero prediction error $\mathcal{L}_1$ for the dominant feature; and then takes a much longer time (up to $T_k \gg T_1$) to converge to a model that attains a near-zero prediction error $\mathcal{L}_k$ for the under-represented features. Our analysis captures the later learning dynamics associated with the under-represented features into a four-phase behavior as further described in the subsequent section. Despite the longer convergence time it takes, in-context learning still achieves the same accurate prediction for under-represented features as that for the dominant feature.

4. Overview of Training Phases

In this section, we explain our key ideas for analyzing the in-context learning capabilities of transformers. We will focus on characterizing the training process of the setting with imbalanced features for under-represented features in Section 4.2, which comprehensively exhibits four phases. Other scenarios take only one or two of those phases, which we will briefly describe in Appendix C. The complete proofs of all the results are provided in the appendix.

4.1. Bilinear Attention Weights

We will first provide the general training dynamics for the *bilinear attention weights* (defined in Definition 4.1 below), which is useful for analyzing all learning phases. These quantities are the key elements in the attention scores $\mathbf{attn}_i^{(t)}$ for $1 \le i \le N$, which play an important role in determining the prediction $\widehat{y}_{\text{query}}^{(t)}$. Hence, our analysis mainly tracks the training dynamics of those bilinear attention weights.

Definition 4.1. (Bilinear Attention Weights) Given $k, n \in [K]$, where $k \ne n$, for $t \ge 0$, we define the bilinear attention weights as follows:

$$A_k^{(t)} := v_k^\top Q^{(t)} v_k, \quad B_{k,n}^{(t)} := v_n^\top Q^{(t)} v_k.$$

By our initialization, we have $A_k^{(0)} = B_{k,n}^{(0)} = 0$.

To further interpret these weights, suppose the query token corresponds to the feature v_k. Then $e^{A_k^{(t)}}$ serves as the (unnormalized) weight for the input token featuring v_k, while $e^{B_{k,n}^{(t)}}$ captures the weight for the input token featuring a different vector v_n with $n \ne k$. Having a larger $A_k^{(t)}$ compared to other $B_{k,n}^{(t)}$ indicates a better capture of the target feature v_k. As shown in eq.(5), this condition implies a higher 'attention' towards input tokens featuring v_k, resulting in $\widehat{y}_{\text{query}}^{(t)} \approx \sum_{i \in \mathcal{V}_k} \mathbf{attn}_i^{(t)} y_i \approx \langle w, v_k \rangle$, where the prediction well approximates the ground truth.

The following lemma provides the GD updates of the bilinear attention weights $A_k^{(t)}$ and $B_{k,n}^{(t)}$.

Lemma 4.2. *Let $t \ge 0$. For $k, n \in [K]$, where $k \ne n$, $A_k^{(t)}$ and $B_{k,n}^{(t)}$ satisfy:*

$$A_k^{(t+1)} = A_k^{(t)} + \eta\alpha_k^{(t)}, \qquad B_{k,n}^{(t+1)} = B_{k,n}^{(t)} + \eta\beta_{k,n}^{(t)},$$
$$\alpha_k^{(t)} = \mathbb{E}\left[\mathbf{1}\{x_{query} = v_k\}\,\mathbf{Attn}_k^{(t)} \cdot \left(\textstyle\sum_{m\ne k} \mathbf{Attn}_m^{(t)^2} + (1 - \mathbf{Attn}_k^{(t)})^2\right)\right],$$
$$\beta_{k,n}^{(t)} = \mathbb{E}\left[\mathbf{1}\{x_{query} = v_k\}\,\mathbf{Attn}_n^{(t)} \cdot \left(\textstyle\sum_{m\ne k} \mathbf{Attn}_m^{(t)^2} - \mathbf{Attn}_n^{(t)} - \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right)\right].$$

Lemma 4.2 shows that $A_k^{(t)}$ is monotonically increasing at any time since $\alpha_k^{(t)} \geq 0$, whereas the monotonicity does not always hold for $B_{k,n}^{(t)}$. Therefore, we need to analyze whether $B_{k,n}^{(t)}$ decreases and determine its rate of change compared to $A_k^{(t)}$. Such a comparison between $B_{k,n}^{(t)}$ and $A_k^{(t)}$ determines which bilinear weight plays a dominant role in the attention dynamics, and the change of the leading weight over the learning process results in different training phases.

4.2. Learning Process for Under-represented Features

We consider the setting with imbalanced features and focus on the under-represented features.

Given a prompt $P = (x_1, y_1, \cdots, x_N, y_N, x_{\text{query}})$, denote P_{input} to be the collection of input tokes, i.e., $\{x_i\}_{i=1}^N$. Recall that $|\mathcal{V}_k|$ is the number of input tokens featuring v_k. Based on our data generation setup, we can show that for imbalanced data, with high probability, P_{input} belongs to

$$\mathcal{E}^*_{\text{imbal}} := \left\{ P_{\text{input}} : |\mathcal{V}_1| = \Theta(N), |\mathcal{V}_k| = \Theta\left(\frac{N}{K}\right) \text{ for } k \geq 2 \right\}.$$

In the following, we focus on the event that $P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}$ unless otherwise specified. We next characterize the learning process for under-represented features v_k with $k > 1$ by four phases. An illustration of these four phases is provided in Figure 1.

4.2.1. Phase I: Decrease of Dominant Feature.

Consider the query token featuring v_k for some $k > 1$. At $t = 0$, $A_k^{(0)} = B_{k,n}^{(0)} = 0$, and hence $\mathbf{attn}_i^{(0)} = \frac{1}{N}$ for $i \in [N]$ which implies that the transformer equally attends each input token. However, due to the imbalanced occurrence of features in $\mathcal{E}^*_{\text{imbal}}$, the number of tokens featuring v_1 is much larger than others. Hence, $\mathbf{Attn}_1^{(0)} = \frac{|\mathcal{V}_1|}{N} \geq \Omega(1)$ while $\mathbf{Attn}_m^{(0)} = \Theta(\frac{1}{K})$ for $m > 1$. Therefore, by Lemma 4.2, we obtain $\beta_{k,1}^{(0)} \leq -\Omega(\frac{1}{K})$, whereas $\alpha_k^{(0)}, |\beta_{k,n}^{(0)}| \approx \Theta(\frac{1}{K^2})$ for $n \neq k, 1$. Therefore, $B_{k,1}^{(t)}$ enjoys a much larger decreasing rate initially. It can be shown that the decrease of $B_{k,1}^{(t)}$ will dominate for a certain time period that defines phase I. The following lemma summarizes our main result in this phase.

Lemma 4.3 (Informal). *Under the same conditions as Theorem 3.3, given $k > 1$, there exists $T_{1,k} = O(\frac{\log(K)K^{1.98}}{\eta})$, such that for all $0 \leq t \leq T_{1,k}$*

$$\beta_{k,1}^{(t)} \leq -\Omega\left(\frac{1}{K^{1.98}}\right), \quad \alpha_k^{(t)} = \Theta\left(\frac{1}{K^2}\right),$$
$$|\beta_{k,n}^{(t)}| \leq O\left(\frac{\alpha_k^{(t)} + |\beta_{k,1}^{(t)}|}{K}\right) \quad \text{for all } n \neq k, 1.$$

At time $t = T_{1,k} + 1$, $B_{k,1}^{(T_{1,k}+1)} \leq -0.49\log(K)$, while $A_k^{(T_{1,k}+1)}$ and $B_{k,n}^{(T_{1,k}+1)}$ for $n \neq k, 1$ remain close to zero.

During phase I, $B_{k,1}^{(t)}$ significantly decreases, leading to a reduction in $\mathbf{Attn}_1^{(t)}$, whereas other $\mathbf{Attn}_n^{(t)}$ with $n > 1$ remain at the level of $\Theta(\frac{1}{K})$. By the end of this phase, $(\mathbf{Attn}_1^{(t)})^2$ drops to $O(\frac{1}{K^{0.98}})$, resulting in a decrease in $|\beta_{k,1}^{(t)}|$ as it approaches $\alpha_k^{(t)}$. Phase II then begins.

4.2.2. Phase II: Switching of Leading Influence

Soon after entering this phase, the dominance role of $B_{k,1}^{(t)}$ diminishes as $|\beta_{k,1}^{(t)}|$ reaches the same order of magnitude as $\alpha_k^{(t)}$. The following result captures the shift of the leading influence, where the growth of $A_k^{(t)}$ takes dominance.

Lemma 4.4 (Informal). *Under the same conditions as Theorem 3.3, given $k > 1$, there exists $T_{2,k} = T_{1,k} + O(\frac{\log(K)K^2}{\eta})$, such that at iteration $t = T_{2,k} + 1$, we have $A_k^{(T_{2,k}+1)} \geq 0.5\log(K)$, $B_{k,1}^{(T_{2,k}+1)} \in [-0.51\log(K), -0.49\log(K)]$, and $B_{k,n}^{(T_{2,k}+1)}$ for $n \neq k, 1$ remain close to zero.*

Lemma 4.4 shows that by the end of phase II, $A_k^{(t)}$ matches the magnitude of $B_{k,1}^{(t)}$, and during phase II $B_{k,1}^{(t)}$ changes only slightly from the end of phase I. This suggests that, at certain moments in this phase, $A_k^{(t)}$ significantly increases and its growth becomes the dominant factor. We next provide some insights into the reasons behind this transition. Once $B_{k,1}^{(t)}$ decreases to $-0.5\log(K)$, we observe that $|\beta_{k,1}^{(t)}| \approx \alpha_k^{(t)} = \Theta(\frac{1}{K^2})$. After this point, it becomes challenging for $B_{k,1}^{(t)}$ to decrease significantly compared to the increase in $A_k^{(t)}$. To illustrate, let us suppose a minimal decrease of $B_{k,1}^{(t)}$ by an amount of $0.01\log(K)$. This would yield that $\mathbf{Attn}_1^{(t)} \leq O(\frac{1}{K^{0.501}})$ and $\beta_{k,1}^{(t)} \leq O(\frac{1}{K^{2.01}})$, while $\mathbf{Attn}_k^{(t)} \geq \Omega(\frac{1}{K})$ and $\alpha_k^{(t)} \geq \Omega(\frac{1}{K^2})$, establishing a situation where $\alpha_k^{(t)} \gg \beta_{k,1}^{(t)}$. Such a discrepancy leads to the switching of the dominant effect.

4.2.3. Phase III: Growth of Target Feature

After a transition phase, we observe that $A_k^{(t)}$ enjoys a larger gradient $\alpha_k^{(t)} \approx \Theta(\frac{1}{K^{1.5}})$ compared to $|\beta_{k,1}^{(t)}| \leq O(\frac{1}{K^{1.98}})$ and $|\beta_{k,n}^{(t)}| \leq O(\frac{1}{K^3})$ with $n \neq k, 1$. This gap between $\alpha_k^{(t)}$ and $\beta_{k,n}^{(t)}$ remains over the period, and the gradient $\alpha_k^{(t)}$ continues to grow, driving the rapid growth of $A_k^{(t)}$ with $B_{k,n}^{(t)}$ being relatively unchanged. The following lemma summarizes our main results in this phase.

Lemma 4.5 (Informal). *Under the same conditions as The-*

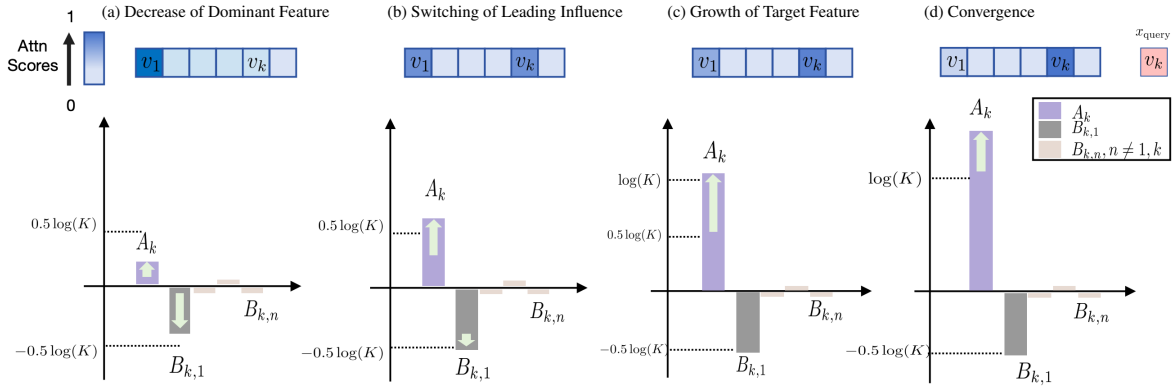


Figure 1. Overview of the dynamics of attention scores and bilinear attention weights for under-represented features. Assume the query token is v_k with $2 \le k \le K$. The *top* row depicts the trend of the attention score $\mathbf{Attn}_m^{(t)}$ for each feature v_m, where a darker color corresponds to a higher score. The *bottom* row shows the interplay and leading effect among bilinear attention weights $A_k^{(t)}$, $B_{k,1}^{(t)}$, and $B_{k,n}^{(t)}$ (where $n \neq 1, k$) in different training phases. **(a)** Phase I: $B_{k,1}^{(t)}$ significantly decreases and the attention on tokens with the dominant feature v_1 is suppressed (Section 4.2.1); **(b)** Phase II: With the suppression of $\mathbf{Attn}_1^{(t)}$, the decreasing rate for $B_{k,1}^{(t)}$ drops and the growth of $A_k^{(t)}$ becomes the leading influence (Section 4.2.2); **(c)** Phase III: $A_k^{(t)}$ rapidly grows and $\mathbf{Attn}_k^{(t)}$ reaches $\Omega(1)$ (Section 4.2.3); **(d)** Phase IV: $\mathbf{Attn}_k^{(t)}$ nearly grows to 1 and the prediction error converges to a global minimum (Section 4.2.4).

orem 3.3, given $k > 1$, there exists $T_{3,k} = O(\frac{\log(K)K^{1.5}}{\eta})$, such that for all $T_{2,k} < t \le T_{3,k}$

$$\alpha_k^{(t)} \ge \Omega\left(\frac{1}{K^{1.5}}\right), \beta_{k,1}^{(t)} \in \left[-O\left(\frac{\alpha_k^{(t)}}{K^{0.48}}\right), -\Omega\left(\frac{1}{K^{2.01}}\right)\right],$$

$$|\beta_{k,n}^{(t)}| \le O\left(\frac{\alpha_k^{(t)} + |\beta_{k,1}^{(t)}|}{K}\right) \text{ with } n \neq k, 1.$$

At time $t = T_{3,k} + 1$, we have $A_k^{(T_{3,k}+1)} \ge \log(K)$.

Lemma 4.5 follows because the continuous growth of $\alpha_k^{(t)}$ is mainly driven by $\mathbf{Attn}_k^{(t)}$, where $1 - \mathbf{Attn}_k^{(t)}$ remains at the constant order. However, as $A_k^{(t)}$ reaches $\log(K)$, $\mathbf{Attn}_k^{(t)}$ is above $\Omega(1)$, necessitating a more detailed analysis to control $\alpha_k^{(t)}$, which starts the final phase.

4.2.4. Phase IV: Convergence

After learning the target feature v_k at a certain level, the prediction error converges. We characterize this in the following lemma, where we establish a connection between $\alpha_k^{(t)}$ and the prediction error via analyzing the change of $1 - \mathbf{Attn}_k^{(t)}$ that diminishes during this phase.

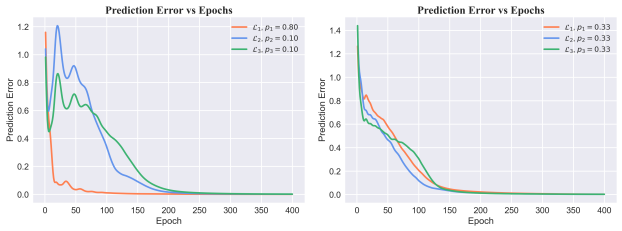
Lemma 4.6 (Informal)**.** *Under the same conditions as Theorem 3.3, given $0 < \epsilon < 1$, for each $k > 1$, there exists $T_{4,k} = T_{3,k} + O(\frac{K\log(K\epsilon^{-\frac{1}{2}})}{\eta\epsilon})$, such that for all $T_{3,k} < t \le T_{4,k}$*

$$\alpha_k^{(t)} \ge \Omega(\frac{\epsilon}{K}), \quad \beta_{k,n}^{(t)} \in [-O(\frac{\alpha_k^{(t)}}{K^{0.49}}), 0],$$

$$\beta_{k,n}^{(t)} \in [-O(\frac{\alpha_k^{(t)}}{K}), 0] \text{ with } n \neq k, 1.$$

At time $t = T_{4,k} + 1$, we have $\mathcal{L}_k(\theta^{(T_{4,k}+1)}) - \mathcal{L}_k^* < \epsilon$ and $(1 - \mathbf{Attn}_k^{(t)})^2 \le O(\epsilon)$, if $x_{query} = v_k$ and $P_{input} \in \mathcal{E}_{imbal}^*$.

The convergence result for $k > 1$ stated in Theorem 3.3 directly follows by choosing $T_k^* = T_{4,k} + 1$.



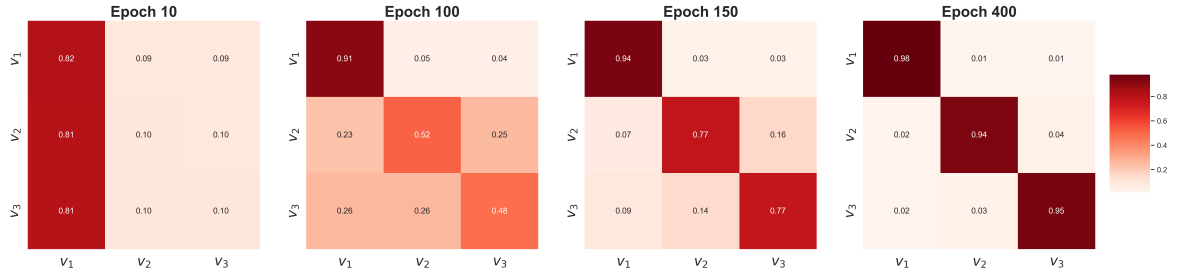
(a) Imbalanced Case (b) Balanced Case

Figure 2. The prediction error for each feature.

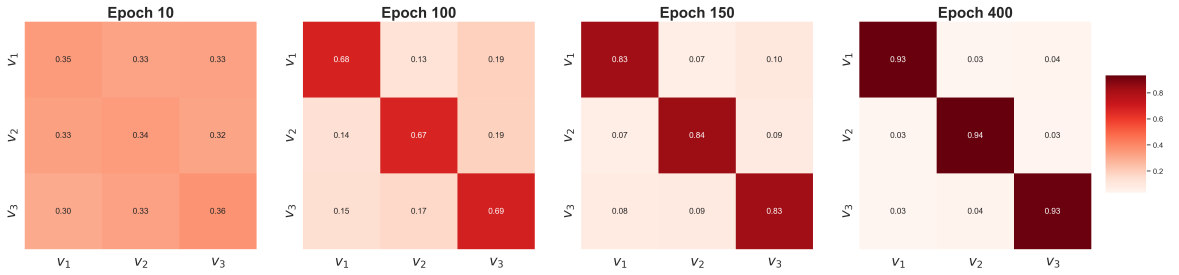
5. Experiments

In this section, we conduct experiments to demonstrate that our theoretical results are consistent with the actual dynamics during the in-context training of transformers. Detailed experimental settings are deferred to Appendix B.

Task and Data Generations. We follow the task and data distributions introduced in Section 2.1. For each task, we sample the task weight w from $\mathcal{N}(0, \mathbf{I}_{d\times d})$. Each data point is drawn from the given feature set $\{v_k \in \mathbb{R}^d, k = 1, \cdots, K\}$ with probability p_k for sampling v_k, where all features are orthonormal vectors, and $p_k \in (0, 1)$ satisfies $\sum_{k=1}^K p_k = 1$. The prompt consists of N random inputs $\{x_i\}_{i=1}^N$ with their task values given by $\{y_i\}_{i=1}^N = \{w^\top x_i\}_{i=1}^N$, and a query x_{query}. We consider the setting with



(a) Imbalanced Case



(b) Balanced Case

Figure 3. The attention heatmap during the training. For each heatmap, the i-th row represents the average attention scores of the query token attending to each feature when $x_{\text{query}} = v_i$.

$d = 16$, $N = 60$, and $K = 3$. We consider the following two types of data distributions:

- Balanced case: $p_i = \frac{1}{3}$ for $i \in [3]$;
- Imbalanced case: v_1 is the dominant feature with $p_1 = 0.8$ and $\{v_2, v_3\}$ are under-represented with $p_2 = p_3 = 0.1$.

Stage-Wise Convergence. In Figure 2, we plot the evolution of the prediction error for each feature throughout the training process. For the imbalanced case (Figure 2a), the transformer quickly converges to a model with nearly vanishing prediction error $\mathcal{L}_1$ for the dominant feature. However, the errors $\mathcal{L}_2$ and $\mathcal{L}_3$ for under-represented features initially fluctuate and then converge to zero after a considerably longer period. This behavior verifies the stage-wise convergence process characterized in our Theorem 3.3.On the other hand, in the balanced scenario (Figure 2b), the prediction errors for all features decrease in a similar manner throughout the training, which validates our theory on convergence in the balanced case in Theorem 3.2.

Attention Score Concentration. In Figure 3, we present the dynamic evolution of attention scores throughout the training process for both balanced and imbalanced scenarios. For each $k \in [3]$, and when $x_{\text{query}} = v_k$, it is observed that $\mathbf{Attn}_k$ progressively increases to be close to 1 while other $\mathbf{Attn}_{k'}$ diminishes at the end of the training. These results support the principle of attention score concentration as elaborated in Theorems 3.2 and 3.3, and demonstrate that the attention is allocated towards those tokens with the same feature as the query token.

Multi-Phase Transition during Training Process. Figure 3 also demonstrates the *multi-phase* convergence process of under-represented features, which verifies those learning phases we characterize in our proof of convergence in Section 4. We elaborate on this by taking the case with $x_{\text{query}} = v_2$ as an example. In Figure 3a and focusing on the row of $x_{\text{query}} = v_2$, from epoch 10 to 100, $\mathbf{Attn}_1$ decreases and $\mathbf{Attn}_3$ increases, which suggests that the decrease in $B_{2,1}$ is the main factor in phase I. If the increase in A_2 was the driving factor, we would expect a decrease in all off-diagonal attention scores including $\mathbf{Attn}_3$ similarly to Figure 3b, which contradicts our observation. Then moving to epoch 150, the simultaneous increase in $\mathbf{Attn}_2$ and decreases in $\mathbf{Attn}_1$ and $\mathbf{Attn}_3$ indicate a shift of dominance effect, with the rise of A_2 becomes the main factor (phases II and III). Finally, the concentration of attention scores at epoch 400 corresponds to the last phase of convergence.

6. Discussions

Practical Insights. One direct practical implication follows from our stage-wise convergence characterization for the imbalanced setting, which implies that employing an early stopping strategy for in-context (pre-)training could be advantageous when the goal is to identify and leverage dominant features quickly. Further, our insights into attention score concentration can provide useful guidance for dealing with non-stationarity in real-world applications. For example, in scenarios with task shifts, the (pre-)trained model would exhibit considerable robustness due to the in-context learning capability, allowing the model to continue to per-

form well. On the other hand, data shifts such as covariate shifts or other more complex shifts would necessitate further training of the model.

Future Directions. Our analysis focuses on an orthonormal feature model for analytical clarity, so that our characterization of the convergence and the dynamics of the attention scores will not be over-complicated by non-essential aspects, e.g., additional non-dominant terms that need to be bounded in gradient calculations. Nevertheless, our analysis can be extended to a more general setting, where the features are drawn from a subspace with K features serving as basis vectors. For such a setting, we need to further characterize how correlation among features affects attention coefficients, which we leave as future work. Furthermore, it is also important to generalize our analysis to nonlinear target functions and consider more complicated network architectures.

7. Conclusions

In this work, we investigated the training dynamics of a one-layer transformer with softmax attention trained by GD for in-context learning. We analyzed two settings respectively with balanced and imbalanced features, and proved the guaranteed convergence to a vanishing in-context prediction error by detailing the evolution of attention dynamics for both settings. Interestingly, we characterized a four-phase behavior for the imbalanced settings that sheds light on the intricate attention dynamics between dominant and target under-represented features during training. We also provide empirical results to back up our theoretical characterization. To our knowledge, this is the first work that rigorously analyzed the *softmax* attention dynamics for in-context learning. Our approach features novel ideas for phase decomposition based on the changes of the dominant role between two types of bilinear attention weights in the learning process, and has the potential to facilitate further theoretical understanding of how transformers perform in other algorithms and learning paradigms.

Acknowledgements

The work was supported in part by the U.S. National Science Foundation under the grants CCF-1900145 and DMS-2134145.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here

References

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

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

Akyürek, E., Schuurmans, D., Andreas, J., Ma, T., and Zhou, D. What learning algorithm is in-context learning? investigations with linear models. *arXiv preprint arXiv:2211.15661*, 2022.

Allen-Zhu, Z. and Li, Y. Towards understanding ensemble, knowledge distillation and self-distillation in deep learning. *arXiv preprint arXiv:2012.09816*, 2020.

Bahdanau, D., Cho, K., and Bengio, Y. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*, 2014.

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

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

Chen, L., Lu, K., Rajeswaran, A., Lee, K., Grover, A., Laskin, M., Abbeel, P., Srinivas, A., and Mordatch, I. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

Chou, H.-P., Chang, S.-C., Pan, J.-Y., Wei, W., and Juan, D.-C. Remix: rebalanced mixup. In *Computer Vision–ECCV 2020 Workshops: Glasgow, UK, August 23–28, 2020, Proceedings, Part VI 16*, pp. 95–110. Springer, 2020.

Cui, Y., Jia, M., Lin, T.-Y., Song, Y., and Belongie, S. Class-balanced loss based on effective number of samples. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9268–9277, 2019.

Dai, D., Sun, Y., Dong, L., Hao, Y., Ma, S., Sui, Z., and Wei, F. Why can gpt learn in-context? language models implicitly perform gradient descent as meta-optimizers. In *ICLR 2023 Workshop on Mathematical and Empirical Understanding of Foundation Models*, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Devroye, L. The equivalence of weak, strong and complete convergence in l1 for kernel density estimates. *The Annals of Statistics*, pp. 896–904, 1983.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Garg, S., Tsipras, D., Liang, P. S., and Valiant, G. What can transformers learn in-context? a case study of simple function classes. *Advances in Neural Information Processing Systems*, 35:30583–30598, 2022.

Giannou, A., Rajput, S., Sohn, J.-y., Lee, K., Lee, J. D., and Papailiopoulos, D. Looped transformers as programmable computers. *arXiv preprint arXiv:2301.13196*, 2023.

Han, C., Wang, Z., Zhao, H., and Ji, H. In-context learning of large language models explained as kernel regression. *arXiv preprint arXiv:2305.12766*, 2023.

He, K., Chen, X., Xie, S., Li, Y., Dollár, P., and Girshick, R. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 16000–16009, 2022.

Huang, Y., Lin, J., Zhou, C., Yang, H., and Huang, L. Modality competition: What makes joint training of multi-modal network fail in deep learning?(provably). In *International Conference on Machine Learning*, pp. 9226–9259. PMLR, 2022.

Huang, Y., Wen, Z., Chi, Y., and Liang, Y. Transformers provably learn feature-position correlations in masked image modeling. *arXiv preprint arXiv:2403.02233*, 2024.

Janner, M., Li, Q., and Levine, S. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34: 1273–1286, 2021.

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

Jiang, H. A latent space theory for emergent abilities in large language models. *arXiv preprint arXiv:2304.09960*, 2023.

Kingma, D. P. and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Kitaev, N., Kaiser, Ł., and Levskaya, A. Reformer: The efficient transformer. *arXiv preprint arXiv:2001.04451*, 2020.

Li, H., Wang, M., Liu, S., and Chen, P.-Y. A theoretical understanding of shallow vision transformers: Learning, generalization, and sample complexity. *arXiv preprint arXiv:2302.06015*, 2023a.

Li, Y., Ildiz, M. E., Papailiopoulos, D., and Oymak, S. Transformers as algorithms: Generalization and implicit model selection in in-context learning. *arXiv preprint arXiv:2301.07067*, 2023b.

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

Min, S., Lyu, X., Holtzman, A., Artetxe, M., Lewis, M., Hajishirzi, H., and Zettlemoyer, L. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

OpenAI. Gpt-4 technical report, 2023.

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

Tarzanagh, D. A., Li, Y., Thrampoulidis, C., and Oymak, S. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023.

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

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

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

Wang, X., Zhu, W., and Wang, W. Y. Large language models are implicitly topic models: Explaining and finding good demonstrations for in-context learning. *arXiv preprint arXiv:2301.11916*, 2023.

Wei, J., Wei, J., Tay, Y., Tran, D., Webson, A., Lu, Y., Chen, X., Liu, H., Huang, D., Zhou, D., et al. Larger language models do in-context learning differently. *arXiv preprint arXiv:2303.03846*, 2023.

Wies, N., Levine, Y., and Shashua, A. The learnability of in-context learning. *arXiv preprint arXiv:2303.07895*, 2023.

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

Yadlowsky, S., Doshi, L., and Tripuraneni, N. Pretraining data mixtures enable narrow model selection capabilities in transformer models. *arXiv preprint arXiv:2311.00871*, 2023.

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

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

A. Additional Related Work

In-Context Learning. Recent studies explored theoretical properties of transformers for in-context learning from various perspectives. Focusing on expressive capacity, Akyürek et al. (2022) studied linear regression tasks and showed that trained in-context learners can represent GD of ridge regression and exact least-squares regression. Giannou et al. (2023) proved the existence of a looped transformer that can emulate in-context learning algorithms. Von Oswald et al. (2023); Dai et al. (2023) also showed that transformer trained in-context implements the GD. Bai et al. (2023) further provided comprehensive results of transformers including the expressive power, in-context prediction power, and sample complexity of pre-training, and then constructed two general mechanisms for algorithm selection. Li et al. (2023b) analyzed the generalization error of trained in-context learning transformers. Another line of work considered in-context learning from a different perspective within the Bayesian framework (Xie et al., 2021; Zhang et al., 2023b; Wang et al., 2023; Jiang, 2023; Han et al., 2023; Wies et al., 2023; Ahuja et al., 2023).

Closely related to our work is the line of research by Zhang et al. (2023a); Mahankali et al. (2023); Ahn et al. (2023), which investigated the training dynamics of in-context learning. Specifically, Mahankali et al. (2023) considered linear regression tasks and showed that the one-layer transformer that minimizes the pre-training loss implements one step of GD. Zhang et al. (2023a) investigated in-context learning of transformers with a single linear self-attention layer trained by gradient flow on linear regression tasks, and showed that gradient flow finds a global minimum. Ahn et al. (2023) investigated the landscape of the loss function for linear transformers trained over random instances of linear regression. However, all those works considered only transformers with *linear* self-attention layers and do not capture the crucial role of the *softmax* mapping, which lies in the core design of transformers to be advantageous over other network architectures. Our work focuses on nonlinear transformers with *softmax attention* and characterizes their training dynamics for in-context learning.

Training Dynamics of Transformers. Jelassi et al. (2022) proposed a simplified Vision Transformers (ViT) model in which the attention matrix solely depends on the positional embeddings and showed that the trained model by GD can learn spatial structure. Li et al. (2023a) studied the training of shallow ViT for a classification task and characterized the sample complexity to achieve a desirable generalization performance. However, their analysis relied on a good initialization near the target pattern, which may not be feasible in practice. Tian et al. (2023) analyzed the SGD training dynamics for a one-layer transformer with one self-attention plus one decoder layer and showed how the self-attention layer combines input tokens during the training, but this work did not provide the convergence guarantee for SGD. Tarzanagh et al. (2023) established an equivalence between the optimization geometry of self-attention and a hard-margin SVM problem that separates optimal input tokens from non-optimal tokens using linear constraints on the outer-products of token pairs. While the mathematical setup of these problems is different from in-context learning, some of our analysis techniques may be useful for studying the training dynamics of these problems. Recently, Huang et al. (2024) studied the training dynamics of transformers trained with the masked image modeling method under a self-supervised learning framework.

B. Experimental Settings

In this section, we present additional details for experiments in Section 5.

Transformer Architecture. We consider a simplified transformer network. The model consists of one block with a single-head self-attention layer, followed by a feedforward neural network, which incorporates layer normalization and ReLU activation, and finally concludes with a linear layer for output processing.

Training Setup. We collect $M = 300$ randomly generated prompts and then train the model based on the empirical version of the training objective Equation (4) for 400 epochs using Adam (Kingma & Ba, 2014) with full batch and the learning rate of 0.002. Notice that Adam is a preferred choice for its stability in training transformers, which is also consistent with recent studies (Garg et al., 2022; Zhang et al., 2023a) to tackle the in-context learning ability of transformers over linear function classes.

Evaluations. We focus on two performance metrics. 1). Prediction error: As defined in Equation (6), the prediction error $\mathcal{L}_k$ measures the loss conditioned on the event that the query token is v_k. We evaluate $\mathcal{L}_k$ by averaging the squared loss on the prompts whose query token is v_k. 2). Attention score: We also evaluate the attention score $\mathbf{Attn}_k$ for each feature, where $\mathbf{Attn}_k$ is defined in Definition 3.1 as the average attention score for the k-th feature over the prompts with query token featuring v_k.

C. Overview of Training Phases of Other Settings

We next describe the training dynamics of other settings, which take the phases similar to those discussed in Section 4.2.

Imbalanced Setting for the Dominant Feature. For the dominant feature v_1 in the imbalanced setting, since the overall attention $\mathbf{Attn}_1^{(0)}$ to the target feature already reaches $\Omega(1)$ due to the abundance of tokens featuring v_1 in $\mathcal{E}^*_{\text{imbal}}$, the training directly enters the convergence stage, as summarized in the following lemma.

Lemma C.1 (Informal)**.** *Under the same conditions as Theorem 3.3, given $k > 1$, there exists $T_1 = O(\frac{\log(\epsilon^{-\frac{1}{2}})}{\eta\epsilon})$, such that for all $t \leq T_1$*

$$\alpha_1^{(t)} \geq \Omega(\epsilon), \quad \beta_{1,n}^{(t)} \in [-O(\tfrac{\alpha_n^{(t)}}{K}), 0] \text{ with } n > 1.$$

Further $\mathcal{L}_1(\theta^{(T_1+1)}) - \mathcal{L}_1^* < \epsilon$, and $(1 - \mathbf{Attn}_1^{(T_1+1)})^2 \leq O(\epsilon)$ if $x_{query} = v_1$ and $P_{input} \in \mathcal{E}^*_{imbal}$.

Balanced Scenarios. Similarly to imbalanced settings, we can show that for balanced data, with high probability, P_{input} belongs to $\mathcal{E}^*_{\text{bal}} := \left\{P_{\text{input}} : |\mathcal{V}_k| = \Theta(\frac{N}{K}) \text{ for all } k \in [K]\right\}$. At initialization, the transformer uniformly assigns attention to each token, i.e., $\mathbf{attn}_i^{(0)} = \frac{1}{N}$ for $i \in [N]$. Unlike the imbalanced case, here, due to $P_{\text{input}} \in \mathcal{E}^*_{\text{bal}}$, we have that $\mathbf{Attn}_m^{(0)} = \Theta(\frac{1}{K})$ for $m \in [K]$, indicating nearly equal attention to each feature. Consequently, as Lemma 4.2, we observe a significantly larger gradient in $A_k^{(t)}$ at the outset, with $\alpha_k^{(0)} \approx \Theta(\frac{1}{K^2})$, compared to $|\beta_{k,n}^{(0)}| \approx \Theta(\frac{1}{K^3})$ for $n \neq k$. This behavior mirrors the observations from phase III for under-represented features, allowing us to directly generalize the analysis.

D. Preliminary Development for Main Proofs

In this section, we will introduce warm-up gradient computations and probabilistic lemmas that establish essential properties of the data and the loss function, which are pivotal for the technical proofs in the upcoming sections. Towards the conclusion of this section, we will also provide a summary of the key notations introduced in both the main content and these preliminary sections. These notations will be frequently adopted in our subsequent analyses.

D.1. Gradient Computations

We first calculate the gradient with respect to Q (note that we do not update the parameter ν during the training). We omit the superscript '(t)' and write $L(\theta)$ as L here for simplicity.

Lemma D.1. *The gradient of the loss function with respect to Q is given by*

$$\nabla_Q L = \mathbb{E}\left[(\widehat{y}_{query} - \langle w, x_{query}\rangle) \sum_{i,j\in[N]} \mathbf{attn}_i \, \mathbf{attn}_j (E_i^x - E_j^x) {E_{N+1}^x}^\top y_i\right].$$

Proof. We obtain:

$$\nabla_Q L = \mathbb{E}[(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle) \frac{\partial \widehat{y}_{\text{query}}}{\partial Q}] = \mathbb{E}\left[(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle) \sum_{i\in[N]} \frac{\partial \, \mathbf{attn}_i}{\partial Q} y_i\right]. \tag{7}$$

Denote $Q_{j,k}$ as the entry in j-th row and k-th column of Q, and define $f : \mathbb{R}^{d\times d} \to \mathbb{R}^N$ as $f(Q) = \left(e^{{E_1^x}^\top Q E_{N+1}^x}, \cdots, e^{{E_N^x}^\top Q E_{N+1}^x}\right)^\top$, and $g : \mathbb{R}^N \to \mathbb{R}$ as $g(y) = \frac{y_i}{\sum_{j\in[N]} y_j}$. By the chain rule, we have

$$\begin{aligned}
\frac{\partial \, \mathbf{attn}_i}{\partial Q_{j,k}} &= \text{Tr}\left[\left(\frac{\partial g(y)}{\partial y}\Big|_{y=f(Q)}\right)^\top \frac{\partial f(Q)}{\partial Q_{j,k}}\right] \\
&= \sum_{n\neq i} -\frac{e^{{E_i^x}^\top Q E_{N+1}^x}}{\left(\sum_{n\in[N]} e^{{E_n^x}^\top Q E_{N+1}^x}\right)^2} \cdot e^{{E_n^x}^\top Q E_{N+1}^x} (E_n^x)_j (E_{N+1}^x)_k
\end{aligned}$$

$$+\frac{\sum_{n\in[N]} e^{{E_n^x}^\top Q E_{N+1}^x} - e^{{E_i^x}^\top Q E_{N+1}^x}}{\left(\sum_{n\in[N]} e^{{E_n^x}^\top Q E_{N+1}^x}\right)^2} \cdot e^{{E_i^x}^\top Q E_{N+1}^x} (E_i^x)_j (E_{N+1}^x)_k$$
$$= \mathbf{attn}_i \left((E_i^x)_j (E_{N+1}^x)_k - \sum_{n=1}^{N} \mathbf{attn}_n (E_n^x)_n = j (E_{N+1}^x)_k \right)$$
$$= \mathbf{attn}_i \left(\sum_{n=1}^{N} \mathbf{attn}_n \left((E_i^x)_j - (E_n^x)_j \right) (E_{N+1}^x)_k \right).$$

Then we reorganize these derivatives into a matrix, and have

$$\frac{\partial \mathbf{attn}_i}{\partial Q} = \mathbf{attn}_i \sum_{j\in[N]} \mathbf{attn}_j (E_i^x - E_j^x) {E_{N+1}^x}^\top.$$

Substituting the above equation into Equation (7), we have

$$\nabla_Q L = \mathbb{E}\left[(\widehat{y}_{\text{query}} - \langle w_\tau, x_{\text{query}} \rangle) \sum_{i,j\in[N]} \mathbf{attn}_i \, \mathbf{attn}_j (E_i^x - E_j^x) {E_{N+1}^x}^\top y_i \right].$$

□

Recall that the quantities A_k and $B_{k,n}$ are defined in Definition 4.1. These quantities are associated with the attention weights for each token, and they play a crucial role in our analysis of learning dynamics. We will restate their definitions here for clarity.

Definition D.2. For $k, n \in [K]$ and $n \neq k$, define the following quantities for $t \geq 0$:

$$A_k^{(t)} := v_k^\top Q^{(t)} v_k \qquad \alpha_k^{(t)} = -v_k^\top \nabla_Q L(Q^{(t)}) v_k$$
$$B_{k,n}^{(t)} := v_n^\top Q^{(t)} v_k \qquad \beta_{k,n}^{(t)} = -v_n^\top \nabla_Q L(Q^{(t)}) v_k$$

By GD update, we have

$$A_k^{(t+1)} := A_k^{(t)} + \eta \alpha_k^{(t)}$$
$$B_{k,n}^{(t+1)} := B_{k,n}^{(t)} + \eta \beta_{k,n}^{(t)}$$

Moreover, by our initialization of $Q^{(0)} = \mathbf{0}_{d\times d}$, we have $A_k^{(0)} = B_{k,n}^{(0)} = 0$ for all $k, n \in [K]$ with $n \neq k$.

Next, we apply the expression in Lemma D.1 to compute the gradient projected onto the feature directions, i.e., $\alpha_k^{(t)}$ and $\beta_{k,n}^{(t)}$.

Lemma D.3. *For $k, k' \in [K]$, where $k \neq k'$, we have*

$$\alpha_k^{(t)} = \mathbb{E}\left[\mathbf{1}\{x_{query} = v_k\} \, \mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1 - \mathbf{Attn}_k^{(t)})^2 \right) \right]$$
$$\beta_{k,k'}^{(t)} = \mathbb{E}\left[\mathbf{1}\{x_{query} = v_k\} \, \mathbf{Attn}_{k'}^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 - \mathbf{Attn}_{k'}^{(t)} - \mathbf{Attn}_k^{(t)} (1 - \mathbf{Attn}_k^{(t)}) \right) \right].$$

Proof. For any $k, k' \in [K]$, apply the previous gradient expression in Lemma D.1, and note that only when $E_{N+1}^x = x_{\text{query}} = v_k$, we have ${E_{N+1}^x}^\top v_k \neq 0$. Thus, we obtain

$$v_{k'}^\top \nabla_Q L v_k$$

$$
\begin{aligned}
&= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right) \sum_{i,j\in[N]} \mathbf{attn}_i\, \mathbf{attn}_j\, y_i v_{k'}^\top (E_i^x - E_j^x)\right] \\
&= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right) \sum_{m,n\in[K]} \sum_{i\in\mathcal{V}_m} \sum_{j\in\mathcal{V}_n} \mathbf{attn}_i\, \mathbf{attn}_j\, y_i v_{k'}^\top (v_m - v_n)\right] \\
&= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right) \sum_{n\in[K]} \sum_{i\in\mathcal{V}_{k'}} \sum_{j\in\mathcal{V}_n} \mathbf{attn}_i\, \mathbf{attn}_j\, y_i v_{k'}^\top (v_{k'} - v_n)\right] \\
&\quad + \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right) \sum_{m\in[K]} \sum_{i\in\mathcal{V}_m} \sum_{j\in\mathcal{V}_{k'}} \mathbf{attn}_i\, \mathbf{attn}_j\, y_i v_{k'}^\top (v_m - v_{k'})\right] \\
&= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right) \mathbf{Attn}_{k'} \langle w, v_{k'}\rangle \sum_{n\in[K]} \mathbf{Attn}_n\right] \\
&\quad - \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right) \mathbf{Attn}_{k'} \sum_{m\in[K]} \mathbf{Attn}_m \langle w, v_m\rangle\right] \\
&= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right) \mathbf{Attn}_{k'} \sum_{m\in[K]} \mathbf{Attn}_m \langle w, v_{k'} - v_m\rangle\right].
\end{aligned}
$$

Note that

$$
\widehat{y}_{\text{query}} = \sum_{i\in[N]} \mathbf{attn}_i\, y_i = \sum_{m\in[K]} \mathbf{Attn}_m \langle w, v_m\rangle.
$$

Thus when $x_{\text{query}} = v_k$, we have

$$
\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle = -\sum_{m\in[K]} \mathbf{Attn}_m \langle w, v_k - v_m\rangle.
$$

Substituting this into the above equation, we have

$$
\begin{aligned}
&v_{k'}^\top \nabla_Q L v_k \\
&= -\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\, \mathbf{Attn}_{k'} \left(\sum_{n\in[K]} \mathbf{Attn}_n \langle w, v_k - v_n\rangle\right)\left(\sum_{m\in[K]} \mathbf{Attn}_m \langle w, v_{k'} - v_m\rangle\right)\right] \\
&= -\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\, \mathbf{Attn}_{k'} \left(\sum_{n\in[K]} \sum_{m\in[K]} \mathbf{Attn}_m\, \mathbf{Attn}_n \langle w, v_k - v_n\rangle \langle w, v_{k'} - v_m\rangle\right)\right] \\
&= -\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\, \mathbf{Attn}_{k'} \left(\sum_{n\in[K]} \sum_{m\in[K]} \mathbf{Attn}_m\, \mathbf{Attn}_n (v_k - v_n)^\top w w^\top (v_{k'} - v_m)\right)\right] \\
&= -\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\, \mathbf{Attn}_{k'} \cdot \right. \\
&\qquad \left. \left(\sum_{n\in[K]} \sum_{m\in[K]} \mathbf{Attn}_m\, \mathbf{Attn}_n (v_k - v_n)^\top \mathbb{E}[w w^\top \mid P_{\text{input}} \cup \{x_{\text{query}}\}](v_{k'} - v_m)\right)\right] \\
&= -\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\, \mathbf{Attn}_{k'} \left(\sum_{n\in[K]} \sum_{m\in[K]} \mathbf{Attn}_m\, \mathbf{Attn}_n (v_k - v_n)^\top (v_{k'} - v_m)\right)\right] \\
&= -\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\, \mathbf{Attn}_{k'} \left((v_k - \sum_{n\in[K]} \mathbf{Attn}_n\, v_n)^\top (v_{k'} - \sum_{m\in[K]} \mathbf{Attn}_m\, v_m)\right)\right].
\end{aligned}
$$

When $k' = k$, we obtain

$$\alpha_k = -v_k^\top \nabla_Q L v_k = \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\, \mathbf{Attn}_k\, \|v_k - \sum_n \mathbf{Attn}_n\, v_n\|^2\right]$$
$$= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\, \mathbf{Attn}_k \left((1 - \mathbf{Attn}_k)^2 + \sum_{m \neq k} \mathbf{Attn}_m^2\right)\right].$$

When $k' \neq k$, we have

$$\beta_{k,k'} = -v_{k'}^\top \nabla_Q L v_k$$
$$= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\, \mathbf{Attn}_{k'} \left(\sum_{m \neq k,k'} \mathbf{Attn}_m^2 - \mathbf{Attn}_k(1 - \mathbf{Attn}_k) - \mathbf{Attn}_{k'}(1 - \mathbf{Attn}_{k'})\right)\right]$$
$$= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\, \mathbf{Attn}_{k'} \left(\sum_{m \neq k} \mathbf{Attn}_m^2 - \mathbf{Attn}_k(1 - \mathbf{Attn}_k) - \mathbf{Attn}_{k'}\right)\right].$$

□

D.2. Useful Probabilistic Lemmas for Prompt

Recall that given a prompt $P = (x_1, y_1, \ldots, x_N, y_N, x_{\text{query}})$, we denote P_{input} as the collection of input tokens, i.e., $\{x_i\}_{i=1}^N$. It is worth noting that, based on our data distribution, the occurrence count of the k-th feature in the first N input tokens from P_{input}, denoted as $|\mathcal{V}_k|$, follows a multinomial distribution. Leveraging the concentration property inherent to multinomial distributions, we can identify a high-probability event to which P_{input} belongs. This event constitutes the crux of our subsequent analysis.

We first introduce the following tail bound for multinomial distributions.

Lemma D.4 (Tail Bound of Multinomial Distribution (Devroye, 1983))**.** *Let $(X_1, \cdots, X_K)$ be a multinomial $(N, p_1, \cdots, p_K)$ random vector. For all $\varepsilon \in (0,1)$ and all K satisfying $K/N \leq \varepsilon^2/20$, we have*

$$P\left(\sum_{i=1}^K |X_i - \mathbb{E}(X_i)| > N\varepsilon\right) \leq 3\exp\left(-N\varepsilon^2/25\right).$$

Now we present our characterization of a high-probability event for P_{input}.

Lemma D.5 (High-probability Event for Balanced Data)**.** *Suppose that $p_k = \Theta\left(\frac{1}{K}\right)$ for any $k \in [K]$ and $K^3 \ll N$. For some constant $c_{bal} \geqslant \sqrt{\frac{20K^3}{N}}$, define*

$$\mathcal{E}_{bal}^* := \left\{P_{input} : |\mathcal{V}_k| \in \left[p_k N - \frac{c_{bal} N}{K}, p_k N + \frac{c_{bal} N}{K}\right] \text{ for } k \in [K]\right\}.$$

Then , we have

$$\mathbb{P}(P_{input} \in \mathcal{E}_{bal}^*) \geq 1 - 3\exp\left(-\frac{c_{bal}^2 N}{25K^2}\right).$$

Let us denote $L_k^{bal} = p_k K - c_{bal}$ and $U_k^{bal} = p_k K + c_{bal}$. Note that L_k^{bal}, U_k^{bal} are at the order of the constant level since $p_k = \Theta\left(\frac{1}{K}\right)$. Then for any P_{input} belonging to $\mathcal{E}_{bal}^*$, $|\mathcal{V}_k| \in [\frac{L_k^{bal} N}{K}, \frac{U_k^{bal} N}{K}] = \Theta(\frac{N}{K})$. Note that we can properly choose c_{bal} to guarantee $L_k^{bal} > 0$ for $k \in [K]$.

Proof. Denote $|\mathcal{V}_k| = X_k$. Then $(X_1, \cdots, X_K) \sim$ multinomial $(N, p_1, \cdots, p_K)$. Noting that $\frac{c_{\text{bal}}^2}{20K^2} \geq \frac{K}{N}$ by our choice of c_{bal}, and then letting $\epsilon = \frac{c_{\text{bal}}}{K}$, we have $\epsilon^2/20 \geq \frac{K}{N}$. By multinomial tail bound in Lemma D.4, we obtain

$$P\left(\sum_{i=1}^{K} |X_i - \mathbb{E}(X_i)| > c_{\text{bal}} \frac{N}{K}\right) \leq 3\exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right).$$

Then, since $\mathbb{E}(X_i) = p_i N$, we have

$$\begin{aligned} P\left(\cap_{i=1}^{K}\left\{|X_i - p_i N| > \frac{c_{\text{bal}} N}{K}\right\}\right) &\leq P\left(\sum_{i=1}^{K} |X_i - \mathbb{E}(X_i)| > c_{\text{bal}} \frac{N}{K}\right) \\ &\leq 3\exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right). \end{aligned}$$

□

Lemma D.6 (High-probability Event for Imbalanced Data)**.** *Suppose that $p_1 = \Theta(1)$, $p_k = \Theta\left(\frac{1}{K}\right)$ for $2 \leq k \leq K$, and $K^3 \ll N$. Then for some constant $c_{im} \geqslant \sqrt{\frac{20K^3}{N}}$, there exist constants $U_k^{im} > L_k^{im} > 0$ for any $k \in [K]$, such that letting*

$$\mathcal{E}_{imbal}^* := \left\{P_{input} : |\mathcal{V}_1| \in [L_1^{im} N, U_1^{im} N] \text{ and } |\mathcal{V}_k| \in \left[\frac{L_k^{im} N}{K}, \frac{U_k^{im} N}{K}\right] \text{ for } 2 \leq k \leq K\right\},$$

we have

$$\mathbb{P}(P_{input} \in \mathcal{E}_{imbal}^*) \geq 1 - 3\exp\left(-\frac{c_{im}^2 N}{25K^2}\right).$$

Proof. Similarly to the proof for Lemma D.5, we have

$$P\left(\cap_{i=1}^{K}\left\{|X_i - p_i N| > \frac{c_{\text{im}} N}{K}\right\}\right) \leq 3\exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right).$$

For $k > 1$, let us denote $L_k^{\text{im}} = p_k K - c_{\text{im}}$ and $U_k^{\text{im}} = p_k K + c_{\text{im}}$. Since $p_k = \Theta\left(\frac{1}{K}\right)$, we can easily conclude that $L_k^{\text{im}}, U_k^{\text{im}}$ for $k > 1$ are constant level. Furthermore, for $k = 1$, let $L_1^{\text{im}} = p_1 - 0.01 c_{\text{im}}$ and $U_1^{\text{im}} = p_1 + 0.01 c_{\text{im}}$. Since p_1 is at the order of the $\Theta(1)$, we have

$$\left[p_1 N - \frac{c_{\text{im}} N}{K}, p_1 N + \frac{c_{\text{im}} N}{K}\right] = \left[(p_1 - \frac{c_{\text{im}}}{K})N, (p_1 + \frac{c_{\text{im}}}{K})p_1 N\right] \subset \left[L_1^{\text{im}} N, U_1^{\text{im}} N\right]$$

for sufficiently large K. □

D.3. Properties of Loss Function and Prediction Error

Recall the population loss we consider is given by:

$$L(\theta) = \frac{1}{2}\mathbb{E}\left[(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}} \rangle)^2\right]. \tag{8}$$

In this part, we will present several important lemmas for such a training objective. We first introduce the following lemma, which connects the loss form with the attention score when the query token takes a certain feature.

Lemma D.7 (Loss Calculation)**.** *The population loss $L(\theta)$ can be decomposed into the following form:*

$$L(\theta) = \frac{1}{2}\sum_{k=1}^{K} \mathbb{E}\left[\mathbf{1}\{x_{query} = v_k\}\left(\sum_{m \neq k} \mathbf{Attn}_m^2 + (1 - \mathbf{Attn}_k)^2\right)\right].$$

Proof. Following the calculations similar to those in Lemma D.3, we have

$$\begin{aligned}
L(\theta) &= \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right)^2\right] \\
&= \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\left(\sum_{n\in[K]}\mathbf{Attn}_n\langle w, v_k - v_n\rangle\right)\left(\sum_{m\in[K]}\mathbf{Attn}_m\langle w, v_k - v_m\rangle\right)\right] \\
&= \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\|v_k - \sum_{n\in[K]}\mathbf{Attn}_n\, v_n\|^2\right] \\
&= \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\left((1-\mathbf{Attn}_k)^2 + \sum_{m\neq k}\mathbf{Attn}_m^2\right)\right].
\end{aligned}$$

□

D.3.1. Loss Characterization for the Balanced Case

We first introduce some additional crucial notations for the loss objectives.

Notations for the balanced case.

$$L^* = \min_\theta L(\theta) = \min_\theta \frac{1}{2}\mathbb{E}\left[\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right)^2\right], \tag{9}$$

$$L^{\text{low}} = \frac{1}{2}\left(1 + \frac{1}{K-1}\right)\sum_{k=1}^{K}\mathbb{P}\left(x_{\text{query}} = v_k \cap |\mathcal{V}_k| = 0\right). \tag{10}$$

L^* denotes the minimum value of the population loss in Equation (8) by minimizing over θ in the form of $\{1, Q\}$, and L^{low} represents the sum of unavoidable errors for each $k \in [K]$, given that the query token is the k-th feature but has not been seen in the first N training samples. We will show that L^{low} serves as a lower bound for L^*, and demonstrate that the network trained with GD will attain nearly zero error compared to L^{low}. Our convergence will be established by the suboptimality gap with respect to L^{low}, which necessarily implies the convergence to L^*. (It also implies $L^* - L^{\text{low}}$ is small.) We further introduce the following quantities to facilitate our analysis of the loss function.

$$\begin{aligned}
&L(\theta) = \sum_{k=1}^{K} L_k(\theta), \\
\text{where } &L_k(\theta) = \frac{1}{2}\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right)^2\right], \\
&L_k^{\text{low}} = \frac{1}{2}\left(1 + \frac{1}{K-1}\right)\mathbb{P}\left(x_{\text{query}} = v_k \cap |\mathcal{V}_k| = 0\right), \\
&\widetilde{L}_k(\theta) = \frac{1}{2}\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}^*_{\text{bal}}\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right)^2\right].
\end{aligned}$$

Lemma D.8. *For L^* and L^{low} defined in Equation (9) and Equation (10), respectively, we have $L^{low} \le L^*$ and they are both at the order of $\Theta(e^{-\text{poly}(K)})$ for the balanced data.*

Proof. We first prove $L^{\text{low}} \le L^*$:

$$\begin{aligned}
L^* &= \min_\theta \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right)^2\right] \\
&\ge \min_\theta \frac{1}{2}\sum_{k=1}^{K}\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap |\mathcal{V}_k| = 0\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right)^2\right]
\end{aligned}$$

$$= \min_{\theta} \frac{1}{2} \sum_{k=1}^{K} \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap |\mathcal{V}_k| = 0\}\left(\sum_{m\neq k} \mathbf{Attn}_m^2 + (1-\mathbf{Attn}_k)^2\right)\right]$$

Notice that when the query token is the k-th feature but has not been seen in the first N training samples, $\mathbf{Attn}_k = 0$. Moreover, $\sum_{m\neq k} \mathbf{Attn}_m^2 \geq \frac{1}{K-1}$ by Cauchy–Schwarz inequality. Thus

$$L^* \geq \frac{1}{2}\left(1 + \frac{1}{K-1}\right) \sum_{k=1}^{K} \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap |\mathcal{V}_k| = 0\}\right] = L^{\text{low}}.$$

Furthermore, since x_{query} and P_{input} are independently sampled,

$$L^{\text{low}} = K \cdot \Theta\left(\frac{1}{K}\right) \cdot \left(1 - \Theta\left(\frac{1}{K}\right)\right)^N = \Theta\left(e^{-\text{poly}(K)}\right).$$

where the last equality follows because $N \gg K^3$, and hence $(1 - \Theta\left(\frac{1}{K}\right))^N = \Theta\left(e^{-\text{poly}(K)}\right)$.

We next only need to show $L^* = O(e^{-\text{poly}(K)})$. We have

$$\begin{aligned} L^* = &\min_{\theta} \left(\frac{1}{2} \sum_{k=1}^{K} \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap |\mathcal{V}_k| > 0\}\left(\sum_{m\neq k} \mathbf{Attn}_m^2 + (1-\mathbf{Attn}_k)^2\right)\right]\right. \\ &\left. + \frac{1}{2} \sum_{k=1}^{K} \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap |\mathcal{V}_k| = 0\}\left(\sum_{m\neq k} \mathbf{Attn}_m^2 + 1\right)\right]\right) \end{aligned}$$

Consider $Q = \sigma \mathbf{I}_{d\times d}$. If $x_{\text{query}} = v_k \cap |\mathcal{V}_k| > 0$ holds, we have

$$\begin{aligned} &\sum_{m\neq k} \mathbf{Attn}_m^2 + (1-\mathbf{Attn}_k)^2 \\ &\quad \leq (1-\mathbf{Attn}_k) \max_{m\neq k} \mathbf{Attn}_m + (1-\mathbf{Attn}_k)^2 \\ &\quad \leq 2(1-\mathbf{Attn}_k)^2 = 2\left(\frac{N - |\mathcal{V}_k|}{N - |\mathcal{V}_k| + |V_k| e^{\sigma}}\right)^2 \leq 2\left(\frac{N}{N+e^{\sigma}}\right)^2 \end{aligned}$$

Taking $\sigma = \text{poly}(N)$, then we have

$$L^* \leq O(e^{-\text{poly}(N)}) + O(e^{-\text{poly}(K)}) = O(e^{-\text{poly}(K)}).$$

□

Lemma D.9. *For the balanced data, given $k \in [K]$, for any θ, we have*

$$\widetilde{L}_k(\theta) \leq L_k(\theta) - L_k^{low} \leq \widetilde{L}_k(\theta) + 3p_k \exp\left(-\frac{c_{bal}^2 N}{25K^2}\right).$$

Proof. We proceed the derivation as follows.

$$\begin{aligned} L_k(\theta) - \widetilde{L}_k(\theta) &= \frac{1}{2}\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}_{\text{bal}}^{*\,c}\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right)^2\right] \\ &= \frac{1}{2}\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}_{\text{bal}}^{*\,c}\}\left(\sum_{m\neq k} \mathbf{Attn}_m^2 + (1-\mathbf{Attn}_k)^2\right)\right] \\ &\overset{(a)}{\leq} \frac{1}{2} \cdot 2\mathbb{P}\left(x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}_{\text{bal}}^{*\,c}\right) \end{aligned}$$

$$\overset{(b)}{\leq} p_k \cdot 3\exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right)$$
$$= 3p_k \exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right).$$

where (a) follows from the fact that

$$\sum_{m\neq k} \mathbf{Attn}_m^2 + (1-\mathbf{Attn}_k)^2 \leq (1-\mathbf{Attn}_k)\max_{m\neq k}\mathbf{Attn}_m + (1-\mathbf{Attn}_k)^2 \leq 2,$$

and (b) holds by Lemma D.5.

On the other hand,

$$L_k(\theta) - \widetilde{L}_k(\theta) \geq \frac{1}{2}\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap |\mathcal{V}_k| = 0\}\left(\sum_{m\neq k}\mathbf{Attn}_m^2 + (1-\mathbf{Attn}_k)^2\right)\right]$$
$$\geq \frac{1}{2}\frac{K}{K-1}\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap |\mathcal{V}_k| = 0\}\right] = L_k^{\text{low}}.$$

□

Consequently, for each $k \in [K]$, $\widetilde{L}_k(\theta)$ closely tracks the deviation between $L_k(\theta)$ and L_k^{low}, which is what we will primarily focus on bounding in the subsequent analysis.

D.3.2. Loss Characterization for the Imbalanced Case

Notations for the imbalanced case. In the imbalanced case, we are interested in the prediction error for the query corresponding to each given feature $k \in [K]$. Thus we consider the following conditional prediction error for each $k \in [K]$:

$$\mathcal{L}_k(\theta) = \frac{1}{2}\mathbb{E}\left[\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right)^2 \middle| x_{\text{query}} = v_k\right]. \tag{11}$$

Similarly, we define the minimum and the unavoidable values for such conditional prediction error:

$$\mathcal{L}_k^* = \min_\theta \frac{1}{2}\mathbb{E}\left[\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right)^2 \middle| x_{\text{query}} = v_k\right], \tag{12}$$
$$\mathcal{L}_k^{\text{low}} = \frac{1}{2}\left(1 + \frac{1}{K-1}\right)\mathbb{P}\left(|\mathcal{V}_k| = 0\right), \tag{13}$$

$$\widetilde{\mathcal{L}}_k(\theta) = \frac{1}{2}\mathbb{E}\left[\mathbf{1}\{P_{\text{input}} \in \mathcal{E}_{\text{imbal}}^*\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right)^2 \middle| x_{\text{query}} = v_k\right].$$

Lemma D.10. *Given $k \in [K]$, for $\mathcal{L}_k^*$ and $\mathcal{L}_k^{low}$ defined in Equation (12) and Equation (13), respectively, we have $\mathcal{L}_k^{low} \leq \mathcal{L}_k^*$ and they are both at the order of $\Theta(e^{-\text{poly}(K)})$ for the imbalanced data.*

Proof. The analysis is similar as Lemma D.8, we only show $\mathcal{L}_k^{\text{low}} = \Theta(e^{-\text{poly}(K)})$.

$$\mathcal{L}_k^{\text{low}} = \frac{1}{2}\left(1 + \frac{1}{K-1}\right)\mathbb{P}\left(|\mathcal{V}_k| = 0\right)$$
$$= \Theta(1)(1-p_k)^N.$$

For $k = 1$, $(1-p_1)^N = \Theta(\exp(-N)) = \Theta\left(e^{-\text{poly}(K)}\right)$. For $k > 1$, since $N \gg K^3$, then $(1-p_k)^N = (1-\Theta\left(\frac{1}{K}\right))^N = \Theta\left(e^{-\text{poly}(K)}\right)$, which completes the proof. □

Table 1. Summary of Notations

Notations	Descriptions
$\mathbf{attn}_i^{(t)}$, $\mathbf{Attn}_k^{(t)}$	The attention scores for the i-token and k-th feature, where $i \in [N]$ and $k \in [K]$.
$A_k^{(t)}$, $B_{k,n}^{(t)}$	The bilinear attention weights when $x_{\text{query}} = v_k$: $A_k^{(t)} = e^{v_k^\top Q^{(t)} v_k}$, $B_{k,n}^{(t)} = e^{v_n^\top Q^{(t)} v_k}$ for $n \neq k$.
$\alpha_k^{(t)}$, $\beta_{k,n}^{(t)}$	The gradient updates respectively for $A_k^{(t)}$ and $B_{k,n}^{(t)}$.
P_{input}	The input tokens in the prompt, i.e., $\{x_i\}_{i=1}^N$.
$\mathcal{E}^*_{\text{bal}}$, $\mathcal{E}^*_{\text{imbal}}$	The high-probability events that P_{input} belongs to respectively for the balanced and imbalanced data.
L^*, L^{low}	The minimum value and lower bound on the population loss $L(\theta)$ (8).
$L_k(\theta)$, $\widetilde{L}_k(\theta)$, L_k^{low}	The loss functions on the event $\{x_{\text{query}} = v_k\}$, $\{x_{\text{query}} = v_k\} \cap \{P_{\text{input}} \in \mathcal{E}^*_{\text{bal}}\}$, and the lower bound on L_k.
$\mathcal{L}_k^*$, $\mathcal{L}_k^{\text{low}}$ (**Imbalanced**)	The minimum value and lower bound of prediction error conditioned on $x_{\text{query}} = v_k$, i.e., $\mathcal{L}_k(\theta)$ (11).
$\widetilde{\mathcal{L}}_k(\theta)$ (**Imbalanced**)	The conditional prediction error on the event $\{P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\}$.

Lemma D.11. *For the imbalanced data, given $k \in [K]$, for any θ, we have*

$$\widetilde{\mathcal{L}}_k(\theta) \leq \mathcal{L}_k(\theta) - \mathcal{L}_k^{low} \leq \widetilde{\mathcal{L}}_k(\theta) + 3\exp\left(-\frac{c_{im}^2 N}{25K^2}\right).$$

Proof. The proof of the first inequality is similar to that for Lemma D.9. We next show the second inequality.

$$\begin{aligned}
\mathcal{L}_k(\theta) - \mathcal{L}_k^{\text{low}} &\leq \widetilde{\mathcal{L}}_k(\theta) + \frac{1}{2}\mathbb{E}\left[\mathbf{1}\{P_{\text{input}} \in {\mathcal{E}^*_{\text{imbal}}}^c\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right)^2 \mid x_{\text{query}} = v_k\right] \\
&= \widetilde{\mathcal{L}}_k(\theta) + \frac{1}{2}\mathbb{E}\left[\mathbf{1}\{P_{\text{input}} \in {\mathcal{E}^*_{\text{imbal}}}^c\}\left(\sum_{m\neq k}\mathbf{Attn}_m^2 + (1-\mathbf{Attn}_k)^2\right) \mid x_{\text{query}} = v_k\right] \\
&\leq \widetilde{\mathcal{L}}_k(\theta) + \mathbb{P}\left(P_{\text{input}} \in {\mathcal{E}^*_{\text{imbal}}}^c\right) \\
&\leq \widetilde{\mathcal{L}}_k(\theta) + 3\exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right).
\end{aligned}$$

□

D.4. Notations and Parameters

In Table 1, we summarize the notations introduced throughout the main content and in the preliminary section. Throughout all the proofs in our paper, we consider $N = \text{poly}(K) \gg K^3$, and K is sufficiently large.

E. Analysis for the Balanced Case (Proof of Theorem 3.2)

In this section, we present the analysis for the balanced case, we first discuss the outline of our proof.

E.1. Roadmap of the Proof

We will analyze the convergence of the training process via two phases of dynamics. At the beginning of each phase, we will establish an induction hypothesis, which we expect to remain valid throughout that phase. Subsequently, we will analyze the dynamics under such a hypothesis within the phase, aiming to provide proof of the hypothesis by the end of the phase.

The main idea of the proof lies in analyzing the GD dynamics of $A_k^{(t)}$ and $B_{k,n}^{(t)}$. From Definition D.2 and Lemma D.3, we have

$$A_k^{(t+1)} = A_k^{(t)} + \eta \alpha_k^{(t)},$$
$$B_{k,n}^{(t+1)} = B_{k,n}^{(t)} + \eta \beta_{k,n}^{(t)},$$

where

$$\alpha_k^{(t)} = \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\, \mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m \neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1 - \mathbf{Attn}_k^{(t)})^2\right)\right],$$
$$\beta_{k,n}^{(t)} = \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\, \mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m \neq k} {\mathbf{Attn}_m^{(t)}}^2 - \mathbf{Attn}_n^{(t)} - \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right)\right].$$

We divide the learning process of any feature k in the balanced case into the following two phases.

- **Phase I** ($t \in [0, T_{1,k}]$, Appendix E.2): At initialization, $A_k^{(t)}$ keeps growing at a rate at least of $\frac{\eta}{K^2}$, while $B_{k,n}^{(t)}$ oscillates with a smaller rate of $\frac{\eta}{K^3}$. Therefore, the increase in $A_k^{(t)}$ will dominate the learning dynamics during phase I.
- **Phase II** ($t \in (T_{1,k}, T_{2,k}^{\epsilon}]$, Appendices E.3 and E.4): After rapid growth of self-attention module parameters in phase I, the query token featuring v_k is aligned with these input tokens also featuring v_k effectively and disregards other features. Then the process proceeds to the convergence phase, where $A_k^{(t)}$ monotonically increases and $B_{k,n}^{(t)}$ monotonically decreases, which finally contributes to the convergence of the loss. Based on the variation rates of $A_k^{(t)}$ and $B_{k,n}^{(t)}$, the convergence phase further has two sub-stages as follows.
  - **Stage I** ($t \in (T_{1,k}, \widetilde{T}_{2,k}^{\epsilon}]$, Appendix E.3): the increase of $A_k^{(t)}$ is as fast as $\Omega(\frac{\epsilon}{K})$ while the decrease of $B_{k,n}^{(t)}$ is slow, and the gap $A_k^{(t)} - \max_{m \neq k} B_{k,m}^{(t)}$ stays within $O(\log(\frac{K}{\epsilon^{\frac{1}{2}}}))$.
  - **Stage II** ($t \in (\widetilde{T}_{2,k}^{\epsilon}, T_{2,k}^{\epsilon}]$, Appendix E.4): the increase of $A_k^{(t)}$ and the decrease of $B_{k,n}^{(t)}$ both are relatively steady and the attention nearly focuses on the target feature, leading to the convergence of the loss.

We finally combine all results in the above two phases to prove the convergence of the training process given in Theorem 3.2 (Appendix E.5).

E.2. Phase I: Growth of Target Feature

In this section, we shall study the initial phase of learning the relationship between the query token and its corresponding feature. Firstly, we present the induction hypothesis in this phase. For the k-th feature v_k, we define the **Phase I** as all iterations $0 \leq t \leq T_{1,k}$, where

$$T_{1,k} \triangleq \max\left\{t : A_k^{(t)} \leq \log(K)\right\}.$$

We state the following induction hypothesis, which will hold throughout Phase I. This hypothesis is ultimately proved in Appendix E.2.3.

Induction Hypothesis E.1. For each $0 \leq t \leq T_{1,k}$, the following holds:

a. $A_k^{(t)}$ is monotonically increasing and $A_k^{(t)} \in [0, \log(K)]$;

b. $|B_{k,n}^{(t)}| = O(\frac{A_k^{(t)}}{K})$ for any $n \neq k$.

E.2.1. TECHNICAL LEMMAS

We first introduce several useful technical lemmas.

Lemma E.1. *Suppose Induction Hypothesis E.1 holds at iteration $0 \le t \le T_{k,1}$. If $x_{query} = v_k$ and $P_{input} \in \mathcal{E}^*_{bal}$, the following holds*

1. $\mathbf{Attn}_k^{(t)} = \Omega\left(\frac{1}{K}\right)$;
2. $1 - \mathbf{Attn}_k^{(t)} \ge \Omega(1)$.

Proof. Since $x_{\text{query}} = v_k$, then we have

$$\begin{aligned}\mathbf{Attn}_k^{(t)} &= \frac{|\mathcal{V}_k| e^{{v_k}^\top Q^{(t)} v_k}}{\sum_{j\in[N]} e^{{E_j^x}^\top Q^{(t)} v_k}} \\ &= \frac{|\mathcal{V}_k| \exp(A_k^{(t)})}{\sum_{m\neq k} |\mathcal{V}_m| \exp(B_{k,m}^{(t)}) + |\mathcal{V}_k| \exp(A_k^{(t)})} \\ &= \frac{1}{\sum_{m\neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} \exp(B_{k,m}^{(t)} - A_k^{(t)}) + 1}.\end{aligned}$$

By Induction Hypothesis E.1,

$$e^{-\left(\log(K)+O(\frac{\log(K)}{K})\right)} \le \exp(B_{k,m}^{(t)} - A_k^{(t)}) \le e^{O(\frac{\log(K)}{K})}.$$

Thus

$$\mathbf{Attn}_k^{(t)} \ge \frac{1}{e^{O(\frac{\log(K)}{K})}(\frac{N}{|\mathcal{V}_k|} - 1) + 1} \ge \frac{1}{e^{O(\frac{\log(K)}{K})}(K/L_k^{\text{bal}} - 1) + 1} = \Omega\left(\frac{1}{K}\right),$$

where the second inequality follows because $P_{\text{input}} \in \mathcal{E}^*_{\text{bal}}$.

On the other hand,

$$\mathbf{Attn}_k^{(t)} \le \frac{1}{e^{-\left(\log(K)+O(\frac{\log(K)}{K})\right)}(\frac{N}{|\mathcal{V}_k|} - 1) + 1} \le \frac{1}{e^{-1}(\frac{1}{U_k^{\text{bal}}} - \frac{1}{K}) + 1}.$$

Considering $U^{\text{bal}} = \Theta(1)$, we have

$$1 - \mathbf{Attn}_k^{(t)} \ge \frac{(\frac{1}{U_k^{\text{bal}}} - \frac{1}{K})}{(\frac{1}{U_k^{\text{bal}}} - \frac{1}{K}) + e} \ge \Omega(1).$$

□

Lemma E.2. *Suppose Induction Hypothesis E.1 holds at iteration $0 \le t \le T_{1,k}$. If $x_{query} = v_k$ and $P_{input} \in \mathcal{E}^*_{bal}$, for $n \neq k$, the following holds*

$$\mathbf{Attn}_n^{(t)} = \Theta\left(\frac{1 - \mathbf{Attn}_k^{(t)}}{K}\right) = \Theta\left(\frac{1}{K}\right).$$

Proof. To show the first equality, since $x_{\text{query}} = v_k$, we have

$$\begin{aligned}\mathbf{Attn}_n^{(t)} &= \frac{|\mathcal{V}_n| e^{{v_n}^\top Q^{(t)} v_k}}{\sum_{j\in[N]} e^{{E_j^x}^\top Q^{(t)} v_k}} \\ &= \frac{|\mathcal{V}_n| \exp(B_{k,n}^{(t)})}{\sum_{m\neq k} |\mathcal{V}_m| \exp(B_{k,m}^{(t)}) + |\mathcal{V}_k| \exp(A_k^{(t)})}.\end{aligned}$$

By Induction Hypothesis E.1, $e^{-O(\frac{\log(K)}{K})} \leq \exp(B_{k,m}^{(t)} - B_{k,n}^{(t)}) \leq e^{O(\frac{\log(K)}{K})}$. Combining with the fact that $\frac{|\mathcal{V}_m|}{|\mathcal{V}_n|} = \Theta(1)$ when $P_{\text{input}} \in \mathcal{E}_{\text{bal}}^*$, we have

$$\frac{\mathbf{Attn}_n^{(t)}}{1 - \mathbf{Attn}_k^{(t)}} = \frac{|\mathcal{V}_n| \exp(B_{k,n}^{(t)})}{\sum_{m \neq k} |\mathcal{V}_m| \exp(B_{k,m}^{(t)})} = \frac{1}{\sum_{m \neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_n|} \exp(B_{k,m}^{(t)} - B_{k,n}^{(t)})} = \Theta\left(\frac{1}{K}\right).$$

Combining with the Lemma E.1, we immediately have $\mathbf{Attn}_n^{(t)} = \Theta\left(\frac{1}{K}\right)$. □

E.2.2. Controlling Gradient Updates in Phase I

Lemma E.3. *Given any fixed $k \in [K]$, if Induction Hypothesis E.1 holds at iteration $0 \leq t \leq T_{1,k}$, then $\alpha_k^{(t)} \geq 0$ and satisfies*

$$\alpha_k^{(t)} \geq \Omega\left(\frac{1}{K^2}\right).$$

Proof. By the gradient expression in Lemma D.3,

$$\begin{aligned}
\alpha_k^{(t)} &= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\} \mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m \neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1 - \mathbf{Attn}_k^{(t)})^2\right)\right] \\
&= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}_{\text{bal}}^*\} \mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m \neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1 - \mathbf{Attn}_k^{(t)})^2\right)\right] \\
&\quad + \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in {\mathcal{E}_{\text{bal}}^*}^c\} \mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m \neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1 - \mathbf{Attn}_k^{(t)})^2\right)\right] \\
&\overset{(a)}{\geq} p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}_{\text{bal}}^*) \\
&\quad \times \mathbb{E}\left[\mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m \neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1 - \mathbf{Attn}_k^{(t)})^2\right) \middle| \{x_{\text{query}} = v_k\} \cap \{P_{\text{input}} \in \mathcal{E}_{\text{bal}}^*\}\right] \\
&\geq p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}_{\text{bal}}^*) \times \mathbb{E}\left[\mathbf{Attn}_k^{(t)} \cdot (1 - \mathbf{Attn}_k^{(t)})^2 \middle| \{x_{\text{query}} = v_k\} \cap \{P_{\text{input}} \in \mathcal{E}_{\text{bal}}^*\}\right] \qquad (14) \\
&\overset{(b)}{\geq} \Omega\left(\frac{1}{K^2}\right),
\end{aligned}$$

where (a) follows from the fact that x_{query} is independent with P_{input} and the second term is non-negative, (b) follows from Lemma D.5, Lemma E.1 and the fact that $p_k = \Theta\left(\frac{1}{K}\right)$ in the balanced case and $N \gg K^3$. □

Lemma E.4. *Given any fixed $k \in [K]$, if Induction Hypothesis E.1 holds at iteration $0 \leq t \leq T_{1,k}$, then for any $n \neq k$, $\beta_{k,n}^{(t)}$ satisfies*

$$|\beta_{k,n}^{(t)}| \leq O\left(\frac{\alpha_k^{(t)}}{K}\right).$$

Proof. By the gradient expression in Lemma D.3, we have

$$\beta_{k,n}^{(t)} \leq \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\} \mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m \neq k} {\mathbf{Attn}_m^{(t)}}^2\right)\right], \qquad (15)$$

$$-\beta_{k,n}^{(t)} \leq \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\} \mathbf{Attn}_n^{(t)} \cdot \left(\mathbf{Attn}_n^{(t)} + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right)\right]. \qquad (16)$$

For Equation (15), we further derive

$$\begin{aligned}
\beta_{k,n}^{(t)} &\leq \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}_{\text{bal}}^*\}\,\mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m\neq k} \mathbf{Attn}_m^{(t)^2}\right)\right] \\
&\quad + \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}_{\text{bal}}^{*\,c}\}\,\mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m\neq k} \mathbf{Attn}_m^{(t)^2}\right)\right] \\
&\overset{(a)}{\leq} p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}_{\text{bal}}^*) \cdot \mathbb{E}\left[\mathbf{Attn}_n^{(t)} \cdot \left(\max_{m\neq k} \mathbf{Attn}_m^{(t)}\right) \middle| \{x_{\text{query}} = v_k\} \cap \{P_{\text{input}} \in \mathcal{E}_{\text{bal}}^*\}\right] \\
&\quad + p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}_{\text{bal}}^{*\,c}) \\
&\overset{(b)}{\leq} p_k \mathbb{E}\left[\mathbf{Attn}_n^{(t)} \cdot \left(\max_{m\neq k} \mathbf{Attn}_m^{(t)}\right) \middle| \{x_{\text{query}} = v_k\} \cap \{P_{\text{input}} \in \mathcal{E}_{\text{bal}}^*\}\right] + 3p_k \exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right) \\
&\overset{(c)}{\leq} O\left(\frac{1}{K^3}\right),
\end{aligned} \tag{17}$$

where (a) follows from the fact that x_{query} is independent with P_{input}, $\mathbf{Attn}_n^{(t)} \leq 1$ and $\sum_{m\neq k} \mathbf{Attn}_m^{(t)^2} \leq \max_{m\neq k} \mathbf{Attn}_m^{(t)} \cdot \sum_{m\neq k} \mathbf{Attn}_m^{(t)} \leq \max_{m\neq k} \mathbf{Attn}_m^{(t)}$, (b) follows from Lemma D.5, and (c) follows from Lemma E.2 and the fact that $p_k = \Theta\left(\frac{1}{K}\right)$ and $N \gg K^3$.

For Equation (16), similarly to the derivation above, we have

$$\begin{aligned}
&-\beta_{k,n}^{(t)} \\
&\leq p_k \mathbb{E}\left[\mathbf{Attn}_n^{(t)} \cdot \left(\mathbf{Attn}_n^{(t)} + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}_{\text{bal}}^*\right] + 2p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}_{\text{bal}}^{*\,c}) \\
&\overset{(a)}{=} 2p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}_{\text{bal}}^{*\,c}) + p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}_{\text{bal}}^*) \times \\
&\quad \mathbb{E}\left[\Theta(\frac{1 - \mathbf{Attn}_k^{(t)}}{K}) \cdot \left(\Theta(\frac{1 - \mathbf{Attn}_k^{(t)}}{K}) + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}_{\text{bal}}^*\right] \\
&\overset{(b)}{\leq} p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}_{\text{bal}}^*) \mathbb{E}\left[O(\frac{\mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})^2}{K}) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}_{\text{bal}}^*\right] + 6p_k \exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right) \\
&\overset{(c)}{\leq} O\left(\frac{\alpha_k^{(t)}}{K} + \frac{1}{K}\exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right)\right)
\end{aligned} \tag{18}$$

where (a) follows from Lemma E.2 and (b) follows from Lemma E.1 and Lemma D.5, and (c) follows from Equation (14).

From Lemma E.3 and the choice of $N \gg K^3$, we have

$$\alpha_k^{(t)} \geq \Omega\left(\frac{1}{K^2}\right) \gg 6\exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right). \tag{19}$$

Thus, combining Equations (17) to (19), we have

$$|\beta_{n,k}^{(t)}| \leq \max\left\{O(\frac{\alpha_k^{(t)}}{K}), O\left(\frac{1}{K^3}\right)\right\} = O\left(\frac{\alpha_k^{(t)}}{K}\right).$$

□

E.2.3. End of Phase I

Lemma E.5. *Given any fixed $k \in [K]$, Induction Hypothesis E.1 holds for all iterations $0 \leq t \leq T_{1,k}$, where $T_{1,k}$ is at most $O(\frac{\log(K)K^2}{\eta})$, and at iteration $t = T_{1,k} + 1$, we have*

a. $A_k^{(T_{1,k}+1)} \geq \log(K)$*;*

b. $\mathbf{Attn}_k^{(T_{1,k}+1)} = \Omega(1)$ *if* $x_{query} = v_k$ *and* $P_{input} \in \mathcal{E}^*_{bal}$.

Proof. If Induction Hypothesis E.1 holds, the existence of $T_{1,k} = O(\frac{\log(K)K^2}{\eta})$ directly follows from Lemma E.3.

We next prove Induction Hypothesis E.1. It is easy to verify Induction Hypothesis E.1 holds at $t = 0$. Now we suppose Induction Hypothesis E.1 holds for all iterations $\leq t-1$, and prove it holds at t.

By Lemma E.3, we have $\alpha_k^{(t-1)} \geq 0$. Thus $A_k^{(t)} = A_k^{(t-1)} + \eta\alpha_k^{(t-1)} \geq 0$. Moreover, by the definition of $T_{1,k}$, we immediately obtain $A_k^{(t)} \leq \log(K)$.

By Lemma E.4, we have $|\beta_{k,n}^{(t-1)}| \leq O\left(\frac{\alpha_k^{(t-1)}}{K}\right)$. Thus,

$$\begin{aligned}
|B_{k,n}^{(t)}| &\leq |B_{k,n}^{(t-1)}| + \eta O\left(\frac{\alpha_k^{(t-1)}}{K}\right) \\
&\leq O\left(\frac{A_k^{(t-1)}}{K}\right) + \eta O\left(\frac{\alpha_k^{(t-1)}}{K}\right) \\
&\leq O\left(\frac{A_k^{(t)}}{K}\right).
\end{aligned}$$

The first statement follows the definition of $T_{1,k}$. Moreover, $\mathbf{Attn}_k^{(T_{1,k}+1)} = \Omega(1)$ can be derived from Lemma E.6 in the subsequent section.

□

E.3. Phase II: Convergence: Stage I

After rapid growth of self-attention module parameters in phase I, the query token featuring v_k is aligned with these input tokens also featuring v_k effectively and disregards other features. Then the process proceeds to the convergence phase, where $A_k^{(t)}$ monotonically increases and $B_{k,n}^{(t)}$ monotonically decreases, which finally contributes to the convergence of the loss. Based on the variation rates of $A_k^{(t)}$ and $B_{k,n}^{(t)}$, the convergence phase further has two sub-stages as follows.

Given any $0 < \epsilon < 1$, for $k \in [K]$, define

$$\widetilde{T}_{2,k}^{\epsilon} := \max\left\{t > T_{1,k} : A_k^{(t)} - \max_{m\neq k} B_{k,m}^{(t)} \leq \log\left(\left(\frac{K}{L_k^{\text{bal}}} - 1\right)\left(\left(\frac{3}{\epsilon}\right)^{\frac{1}{2}} - 1\right)\right)\right\}.$$

Induction Hypothesis E.2. For $T_{1,k} < t \leq \widetilde{T}_{2,k}^{\epsilon}$, suppose $\text{polylog}(K) \gg \log(\frac{1}{\epsilon})$, and the following holds

a. $A_k^{(t)}$ is monotonically increasing and $A_k^{(t)} \in [\log(K), O(\log(K/\epsilon))]$;

b. $B_{k,n}^{(t)}$ is monotonically decreasing and $|B_{k,n}^{(t)}| = O(\frac{A_k^{(t)}}{K})$ for any $n \neq k$.

E.3.1. Technical Lemmas

We first introduce several useful technical lemmas.

Lemma E.6. *Suppose Induction Hypothesis E.2 holds at iteration* $T_{1,k} < t \leq \widetilde{T}_{2,k}^{\epsilon}$. *If* $x_{query} = v_k$ *and* $P_{input} \in \mathcal{E}^*_{bal}$, *the following holds*

1. $\mathbf{Attn}_k^{(t)} = \Omega(1)$*;*
2. $(1 - \mathbf{Attn}_k^{(t)})^2 \geq \Omega(\epsilon) = \Omega(\exp(-\text{polylog}(K)))$.

Proof. Since $x_{\text{query}} = v_k$, we have

$$\begin{aligned}\mathbf{Attn}_k^{(t)} &= \frac{|\mathcal{V}_k|\exp(A_k^{(t)})}{\sum_{m\neq k}|\mathcal{V}_m|\exp(B_{k,m}^{(t)}) + |\mathcal{V}_k|\exp(A_k^{(t)})} \\ &= \frac{1}{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_k|}\exp(B_{k,m}^{(t)} - A_k^{(t)}) + 1}.\end{aligned}$$

By Induction Hypothesis E.2, we obtain

$$\exp(B_{k,m}^{(t)} - A_k^{(t)}) \leq e^{O(\frac{\log(K/\epsilon)}{K}) - \log(K)} \leq e^{O(\frac{\log(K)+\text{polylog}(K)}{K}) - \log(K)} \leq O\left(\frac{1}{K}\right).$$

Therefore,

$$\mathbf{Attn}_k^{(t)} \geq \frac{1}{O\left(\frac{1}{K}\right)\left(\frac{N}{|\mathcal{V}_k|} - 1\right) + 1} \geq \frac{1}{O(\frac{1}{L_k^{\text{bal}}} - \frac{1}{K}) + 1} \geq \Omega(1).$$

On the other hand, by the definition of $\widetilde{T}_{2,k}^{\epsilon}$, we have

$$\begin{aligned}1 - \mathbf{Attn}_k^{(t)} &= \frac{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_k|}\exp(B_{k,m}^{(t)} - A_k^{(t)})}{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_k|}\exp(B_{k,m}^{(t)} - A_k^{(t)}) + 1} \\ &\geq \frac{\exp(\min_{m\neq k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{N}{|\mathcal{V}_k|} - 1)}{\exp(\min_{m\neq k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{N}{|\mathcal{V}_k|} - 1) + 1} \\ &\geq \frac{\exp(\min_{m\neq k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{K}{U_k^{\text{bal}}} - 1)}{\exp(\min_{m\neq k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{K}{U_k^{\text{bal}}} - 1) + 1} \\ &= \frac{\exp(\max_{m\neq k} B_{k,m}^{(t)} - A_k^{(t)} - \Delta B_k^{(t)})(\frac{K}{U_k^{\text{bal}}} - 1)}{\exp(\max_{m\neq k} B_{k,m}^{(t)} - A_k^{(t)} - \Delta B_k^{(t)})(\frac{K}{U_k^{\text{bal}}} - 1) + 1} \\ &\geq \frac{(\frac{K}{L_k^{\text{bal}}} - 1)^{-1}(\epsilon^{-\frac{1}{2}} - 1)^{-1}\cdot e^{-O(\frac{\text{polylog}(K)}{K})}(\frac{K}{U_k^{\text{bal}}} - 1)}{(\frac{K}{L_k^{\text{bal}}} - 1)^{-1}(\epsilon^{-\frac{1}{2}} - 1)^{-1}e^{-O(\frac{\text{polylog}(K)}{K})}(\frac{K}{U_k^{\text{bal}}} - 1) + 1} \\ &\geq \Omega(\epsilon^{\frac{1}{2}}),\end{aligned}$$

where $\Delta B_k^{(t)} = \max_{m\neq k} B_{k,m}^{(t)} - \min_{m\neq k} B_{k,m}^{(t)} = O(\frac{A_k^{(t)}}{K})$, and the first and second inequalities follow from the fact that $\frac{x}{1+x}$ monotonically increases w.r.t. $x \geq 0$, and the third inequality follows from the definition of $\widetilde{T}_{2,k}^{\epsilon}$ and Induction Hypothesis E.2. □

Lemma E.7. *Suppose Induction Hypothesis E.2 holds at iteration $T_{1,k} < t \leq \widetilde{T}_{2,k}^{\epsilon}$. If $x_{query} = v_k$ and $P_{input} \in \mathcal{E}_{bal}^*$, for $n \neq k$, then the following holds*

$$\mathbf{Attn}_n^{(t)} = \Theta\left(\frac{1 - \mathbf{Attn}_k^{(t)}}{K}\right).$$

Proof. By definition,

$$\mathbf{Attn}_n^{(t)} = \frac{|\mathcal{V}_n|\exp(B_{k,n}^{(t)})}{\sum_{m\neq k}|\mathcal{V}_m|\exp(B_{k,m}^{(t)}) + |\mathcal{V}_k|\exp(A_k^{(t)})}.$$

By Induction Hypothesis E.2, we have

$$e^{-O(\frac{\log(K)-\log(\epsilon)}{K})} \leq \exp(B_{k,m}^{(t)} - B_{k,n}^{(t)}) \leq e^{O(\frac{\log(K)-\log(\epsilon)}{K})}.$$

Further combining with the fact that $-\log(\epsilon) \ll \mathrm{polylog}(K)$, we have

$$\frac{\mathbf{Attn}_n^{(t)}}{1-\mathbf{Attn}_k^{(t)}} = \frac{|\mathcal{V}_n|\exp(B_{k,n}^{(t)})}{\sum_{m\neq k}|\mathcal{V}_m|\exp(B_{k,m}^{(t)})} = \frac{1}{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_n|}\exp(B_{k,m}^{(t)}-B_{k,n}^{(t)})} = \Theta\left(\frac{1}{K}\right).$$

□

E.3.2. Controlling Gradient Updates in Stage I of Phase II

Lemma E.8. *At each iteration $T_{1,k} < t \le \widetilde{T}_{2,k}^{\epsilon}$, if Induction Hypothesis E.2 holds, then $\alpha_k^{(t)} \ge 0$ and satisfies*

$$\alpha_k^{(t)} \ge \Omega\left(\frac{\epsilon}{K}\right).$$

Proof. The analysis is similar to that for Lemma E.3, but we need to be more careful about the lower bound of $1-\mathbf{Attn}_k^{(t)}$. By gradient expression in Lemma D.3, we obtain

$$\begin{aligned}
\alpha_k^{(t)} &= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\,\mathbf{Attn}_k^{(t)}\cdot\left(\sum_{m\neq k}{\mathbf{Attn}_m^{(t)}}^2 + (1-\mathbf{Attn}_k^{(t)})^2\right)\right] \\
&\ge p_k\cdot\mathbb{P}(P_{\text{input}}\in\mathcal{E}_{\text{bal}}^*)\mathbb{E}\left[\mathbf{Attn}_k^{(t)}\cdot\left(\sum_{m\neq k}{\mathbf{Attn}_m^{(t)}}^2 + (1-\mathbf{Attn}_k^{(t)})^2\right)\mid\{x_{\text{query}}=v_k\}\cap\mathcal{E}_{\text{bal}}^*\right] \\
&\ge p_k\cdot\mathbb{P}(P_{\text{input}}\in\mathcal{E}_{\text{bal}}^*)\mathbb{E}\left[\mathbf{Attn}_k^{(t)}\cdot(1-\mathbf{Attn}_k^{(t)})^2\mid\{x_{\text{query}}=v_k\}\cap\mathcal{E}_{\text{bal}}^*\right] \\
&\ge \Omega(\frac{\epsilon}{K}),
\end{aligned}$$

where the last inequality follows from Lemmas D.5 and E.6 and the fact that $p_k = \Theta\left(\frac{1}{K}\right)$ in the balanced case. □

Lemma E.9. *At each iteration $T_{1,k} < t \le \widetilde{T}_{2,k}^{\epsilon}$, if Induction Hypothesis E.2 holds, then given $k\in[K]$, for any $n\neq k$, $\beta_{k,n}^{(t)}$ satisfies*

$$-O\left(\frac{\alpha_k^{(t)}}{K}\right) \le \beta_{k,n}^{(t)} \le 0.$$

Proof. Note that conditioned on the event $\{x_{\text{query}}=v_k\}\cap\{P_{\text{input}}\in\mathcal{E}_{\text{bal}}^*\}$, by Lemmas E.6 and E.7, we have $\mathbf{Attn}_k^{(t)} = \Omega(1)$, $\max_{m\neq k}\mathbf{Attn}_m = O\left(\frac{1}{K}\right)$, and thus

$$\begin{aligned}
\sum_{m\neq k}{\mathbf{Attn}^{(t)}}_m^2 - \mathbf{Attn}_n^{(t)} - \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)}) &\le \max_{m\neq k}\mathbf{Attn}_m^{(t)}\sum_{m\neq k}\mathbf{Attn}_m^{(t)} - \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)}) \\
&= -(1-\mathbf{Attn}_k^{(t)})(\mathbf{Attn}_k^{(t)} - \max_{m\neq k}\mathbf{Attn}_m^{(t)}) \\
&\le -\Omega(1-\mathbf{Attn}_k^{(t)}). \qquad (20)
\end{aligned}$$

Therefore, by combining with Lemma D.3, we obtain

$$\begin{aligned}
\beta_{k,n}^{(t)} \le\ &\mathbb{E}\left[\mathbf{1}\{x_{\text{query}}=v_k\cap\mathcal{E}_{\text{bal}}^*\}\,\mathbf{Attn}_n^{(t)}\cdot\left(\sum_{m\neq k}{\mathbf{Attn}_m^{(t)}}^2 - \mathbf{Attn}_n^{(t)} - \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)})\right)\right] \\
&+\mathbb{E}\left[\mathbf{1}\{x_{\text{query}}=v_k\cap{\mathcal{E}_{\text{bal}}^*}^c\}\,\mathbf{Attn}_n^{(t)}\cdot\left(\sum_{m\neq k}{\mathbf{Attn}_m^{(t)}}^2\right)\right]
\end{aligned}$$

$$\overset{(a)}{\leq} p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{bal}}) \cdot \mathbb{E}\left[-\Omega(\frac{(1-\mathbf{Attn}_k^{(t)})^2}{K}) \mid \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{bal}}\right] + p_k \cdot \mathbb{P}({\mathcal{E}^*_{\text{bal}}}^c)$$
$$\overset{(b)}{\leq} p_k \cdot \left(-\Omega(\frac{\epsilon}{K})\right) + 3p_k \exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right)$$
$$\leq 0,$$

where (a) follows from Equation (20) and Lemma E.7, (b) follows from Lemmas D.5 and E.6, and the last inequality holds since

$$\frac{\epsilon}{K} \gg \frac{\exp(-\operatorname{polylog}(K))}{K} \gg \exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right).$$

Moreover, following the analysis similar to that for Lemma E.4, we have

$$\begin{aligned}
-\beta_{k,n}^{(t)} &\leq p_k \mathbb{E}\left[\mathbf{Attn}_n^{(t)} \cdot \left(\mathbf{Attn}_n^{(t)} + \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)})\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{bal}}\right] + p_k \mathbb{P}({\mathcal{E}^*_{\text{bal}}}^c) \\
&\leq p_k \mathbb{E}\left[\Theta(\frac{1-\mathbf{Attn}_k^{(t)}}{K}) \cdot O\left(\mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)})\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{bal}}\right] \\
&\quad + 6p_k \exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right) \\
&= p_k \mathbb{E}\left[O(\frac{\mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)})^2}{K}) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{bal}}\right] + 6p_k \exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right) \\
&\leq O(\frac{\alpha_k^{(t)}}{K}).
\end{aligned}$$

□

E.3.3. End of Stage I of Phase II

Lemma E.10. *Given $k \in [K]$, and $0 < \epsilon < 1$, suppose $\operatorname{polylog}(K) \gg \log(\frac{1}{\epsilon})$. Then Induction Hypothesis E.2 holds for at least all $T_{1,k} < t \leq \widetilde{T}_{2,k}^{\epsilon} = T_{1,k} + O\left(\frac{K\log(K\epsilon^{-\frac{1}{2}})}{\eta\epsilon}\right)$, and at iteration $t = \widetilde{T}_{2,k}^{\epsilon} + 1$, we have $A_k^{(\widetilde{T}_{2,k}^{\epsilon}+1)} \geq \Omega\left(\log(\frac{K}{\epsilon})\right)$.*

Proof. We first prove the existence of $\widetilde{T}_{2,k}^{\epsilon}$. Recall that

$$\widetilde{T}_{2,k}^{\epsilon} := \max\left\{t > T_{1,k} : A_k^{(t)} - \max_{m \neq k} B_{k,m}^{(t)} \leq \log\left(\left(\frac{K}{L_k^{\text{bal}}} - 1\right)\left((\frac{3}{\epsilon})^{\frac{1}{2}} - 1\right)\right)\right\}.$$

When $t \in (T_{1,k}, \widetilde{T}_{2,k}^{\epsilon}]$, consider

$$\begin{aligned}
&\left(A_k^{(t+1)} - \max_{m \neq k} B_{k,m}^{(t+1)}\right) - \left(A_k^{(t)} - \max_{m \neq k} B_{k,m}^{(t)}\right) \\
&\geq \eta(1 - O\left(\frac{1}{K}\right))\alpha_k^{(t)} = \Omega\left(\frac{\eta\epsilon}{K}\right),
\end{aligned}$$

where the inequality follows from Lemma E.9 and the last equation follows from Lemma E.8. Therefore, at most

$$\widetilde{T}_{2,k}^{\epsilon} - T_{1,k} = O(\frac{K\log\left((\frac{K}{L_k^{\text{bal}}} - 1)((\frac{3}{\epsilon})^{\frac{1}{2}} - 1)\right)}{\eta\epsilon}) = O(\frac{K\log(K\epsilon^{-\frac{1}{2}})}{\eta\epsilon})$$

iterations are needed before $A_k^{(t)} - \max_{m \neq k} B_{k,m}^{(t)}$ exceeds $\log\left(\left(\frac{K}{L_k^{\text{bal}}} - 1\right)\left((\frac{3}{\epsilon})^{\frac{1}{2}} - 1\right)\right)$.

It is easy to verify Induction Hypothesis E.2 holds at $t = T_{1,k} + 1$. Now we suppose Induction Hypothesis E.2 holds for all iterations in $[T_{1,k} + 1, t - 1]$, and prove it holds at t.

By Lemma E.8, we have $\alpha_k^{(t-1)} \geq 0$. Thus $A_k^{(t)} \geq A_k^{(t-1)} \geq \log(K)$. By Lemma E.9, we have $-O\left(\frac{\alpha_k^{(t-1)}}{K}\right) \leq \beta_{k,n}^{(t-1)} \leq 0$. Thus,

$$\begin{aligned}
|B_{k,n}^{(t)}| &\leq |B_{k,n}^{(t-1)}| + \eta O\left(\frac{\alpha_k^{(t-1)}}{K}\right) \\
&\leq O\left(\frac{A_k^{(t-1)}}{K}\right) + \eta O\left(\frac{\alpha_k^{(t-1)}}{K}\right) \\
&\leq O\left(\frac{A_k^{(t)}}{K}\right).
\end{aligned}$$

Moreover, by the definition of $\widetilde{T}_{2,k}^{\epsilon}$, for any $T_{1,k} < t \leq \widetilde{T}_{2,k}^{\epsilon}$ we immediately have

$$\left(1 - O\left(\frac{1}{K}\right)\right) A_k^{(t)} \leq A_k^{(t)} - \max_{m \neq k} B_{k,m}^{(t)} \leq \log\left(\left(\frac{K}{L_k^{\text{bal}}} - 1\right)\left(\left(\frac{3}{\epsilon}\right)^{\frac{1}{2}} - 1\right)\right).$$

Therefore, $A_k^{(t)} \leq O(\log(\frac{K}{\epsilon}))$ for any $T_{1,k} < t \leq \widetilde{T}_{2,k}^{\epsilon}$.

At iteration $t = \widetilde{T}_{2,k}^{\epsilon} + 1$, we have $A_k^{(\widetilde{T}_{2,k}^{\epsilon}+1)} - \max_{m \neq k} B_{k,m}^{(\widetilde{T}_{2,k}^{\epsilon}+1)} > \log\left((\frac{K}{L_k^{\text{bal}}} - 1)((\frac{3}{\epsilon})^{\frac{1}{2}} - 1)\right)$. Thus $A_k^{(\widetilde{T}_{2,k}^{\epsilon}+1)} \geq \Omega(\log(\frac{K}{\epsilon}))$.

When $\{x_{\text{query}} = v_k\} \cap \{P_{\text{input}} \in \mathcal{E}_{\text{bal}}^*\}$, we obtain

$$\begin{aligned}
1 - \mathbf{Attn}_k^{(\widetilde{T}_{2,k}^{\epsilon}+1)} &= \frac{\sum_{m \neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} \exp(B_{k,m}^{(t)} - A_k^{(t)})}{\sum_{m \neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} \exp(B_{k,m}^{(t)} - A_k^{(t)}) + 1} \\
&\leq \frac{\exp(\max_{m \neq k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{N}{|\mathcal{V}_k|} - 1)}{\exp(\max_{m \neq k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{N}{|\mathcal{V}_k|} - 1) + 1} \\
&\leq \frac{\exp(\max_{m \neq k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{K}{L_k^{\text{bal}}} - 1)}{\exp(\max_{m \neq k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{K}{L_k^{\text{bal}}} - 1) + 1} \\
&\leq \frac{\left((\frac{K}{L_k^{\text{bal}}} - 1)((\frac{3}{\epsilon})^{\frac{1}{2}} - 1)\right)^{-1}(\frac{K}{L_k^{\text{bal}}} - 1)}{\left((\frac{K}{L_k^{\text{bal}}} - 1)((\frac{3}{\epsilon})^{\frac{1}{2}} - 1)\right)^{-1}(\frac{K}{L_k^{\text{bal}}} - 1) + 1} \\
&= (\epsilon/3)^{\frac{1}{2}},
\end{aligned}$$

where the first inequality follows from the fact that $\frac{x}{1+x}$ monotonically increases w.r.t. $x \geq 0$.

□

E.4. Phase II: Convergence: Stage II

Given $k \in [K]$, define

$$T_{2,k}^{\epsilon} := \widetilde{T}_{2,k}^{\epsilon} + O\left(\frac{K \log\left(K\epsilon^{-\frac{1}{2}}\right)}{\epsilon\eta}\right).$$

Induction Hypothesis E.3. Suppose $\mathrm{polylog}(K) \gg \log(\frac{1}{\epsilon})$ for $t \in (\widetilde{T}^{\epsilon}_{2,k}, T^{\epsilon}_{2,k}]$. The following holds:

a. $A_k^{(t)}$ is monotonically increasing but cannot exceed $O(\log(K/\epsilon))$;

b. $B_{k,m}^{(t)}$ is monotonically decreasing and $|B_{k,m}^{(t)}| = O(\frac{A_k^{(t)}}{K})$ for any $m \neq k$.

E.4.1. Technical Lemmas

We first introduce several useful technical lemmas.

Lemma E.11. *Suppose Induction Hypothesis E.3 holds at iteration $t \in (\widetilde{T}^{\epsilon}_{2,k}, T^{\epsilon}_{2,k}]$. If $x_{query} = v_k$ and $P_{input} \in \mathcal{E}^*_{bal}$, the following holds*

1. $\mathbf{Attn}_k^{(t)} = \Omega(1)$;
2. $(1 - \mathbf{Attn}_k^{(t)})^2 \in [\Omega(\exp(-\mathrm{polylog}(K))), \epsilon]$.

Proof. Since $x_{\text{query}} = v_k$, we have

$$\begin{aligned}\mathbf{Attn}_k^{(t)} &= \frac{|\mathcal{V}_k| \exp(A_k^{(t)})}{\sum_{m\neq k} |\mathcal{V}_m| \exp(B_{k,m}^{(t)}) + |\mathcal{V}_k| \exp(A_k^{(t)})} \\ &= \frac{1}{\sum_{m\neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} \exp(B_{k,m}^{(t)} - A_k^{(t)}) + 1}.\end{aligned}$$

By Induction Hypothesis E.3,

$$\exp(B_{k,m}^{(t)} - A_k^{(t)}) \leq e^{O(\frac{\log(K/\epsilon)}{K}) - \log(K)} \leq e^{O(\frac{\log(K)+\mathrm{polylog}(K)}{K}) - \log(K)} \leq O\left(\frac{1}{K}\right).$$

Therefore,

$$\mathbf{Attn}_k^{(t)} \geq \frac{1}{O\left(\frac{1}{K}\right)\left(\frac{N}{|\mathcal{V}_k|} - 1\right) + 1} \geq \frac{1}{O(\frac{1}{L_k^{\text{bal}}} - \frac{1}{K}) + 1} \geq \Omega(1).$$

We first upper-bound $1 - \mathbf{Attn}_k^{(t)}$ as

$$\begin{aligned}1 - \mathbf{Attn}_k^{(t)} &= \frac{\sum_{m\neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} \exp(B_{k,m}^{(t)} - A_k^{(t)})}{\sum_{m\neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} \exp(B_{k,m}^{(t)} - A_k^{(t)}) + 1} \\ &\leq \frac{\exp(\max_{m\neq k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{N}{|\mathcal{V}_k|} - 1)}{\exp(\max_{m\neq k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{N}{|\mathcal{V}_k|} - 1) + 1} \\ &\overset{(a)}{\leq} \frac{\exp(\max_{m\neq k} B_{k,m}^{(\widetilde{T}^{\epsilon}_{2,k}+1)} - A_k^{(\widetilde{T}^{\epsilon}_{2,k}+1)})(\frac{N}{|\mathcal{V}_k|} - 1)}{\exp(\max_{m\neq k} B_{k,m}^{(\widetilde{T}^{\epsilon}_{2,k}+1)} - A_k^{(\widetilde{T}^{\epsilon}_{2,k}+1)})(\frac{N}{|\mathcal{V}_k|} - 1) + 1} \\ &\overset{(b)}{<} \left(\frac{\epsilon}{3}\right)^{\frac{1}{2}},\end{aligned}$$

where (a) holds since $\max_{m\neq k} B_{k,m}^{(t)} - A_k^{(t)}$ is non-increasing by Induction Hypothesis E.3, and (b) follows from the definition of $\widetilde{T}^{\epsilon}_{2,k}$.

Then we lower-bound $1 - \mathbf{Attn}_k^{(t)}$ following the analysis similar to that for Lemma E.6:

$$1 - \mathbf{Attn}_k^{(t)} = \frac{\sum_{m\neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} \exp(B_{k,m}^{(t)} - A_k^{(t)})}{\sum_{m\neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} \exp(B_{k,m}^{(t)} - A_k^{(t)}) + 1}$$

$$
\begin{aligned}
&\geq \frac{\exp(\min_{m\neq k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{N}{|\mathcal{V}_k|}-1)}{\exp(\min_{m\neq k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{N}{|\mathcal{V}_k|}-1)+1} \\
&\geq \frac{\exp(\min_{m\neq k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{K}{U_k^{\text{bal}}}-1)}{\exp(\min_{m\neq k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{K}{U_k^{\text{bal}}}-1)+1} \\
&\geq \frac{\frac{1}{e^{O(\log(K/\epsilon))}}(\frac{K}{U_k^{\text{bal}}}-1)}{\frac{1}{e^{O(\log(K/\epsilon))}}(\frac{K}{U_k^{\text{bal}}}-1)+1} \\
&\geq \frac{\frac{1}{e^{O(\text{polylog}(K))}}(\frac{K}{U_k^{\text{bal}}}-1)}{\frac{1}{e^{O(\text{polylog}(K))}}(\frac{K}{U_k^{\text{bal}}}-1)+1} \\
&\geq \Omega(\exp(-\text{polylog}(K))),
\end{aligned}
$$

where the first three inequalities follow from the fact that $\frac{x}{1+x}$ monotonically increases w.r.t. $x \geq 0$ and $A_k^{(t)} \leq O(\log(K/\epsilon))$. □

Lemma E.12. *Suppose Induction Hypothesis E.3 holds at iteration $t \in (\widetilde{T}_{2,k}^{\epsilon}, T_{2,k}^{\epsilon}]$. If $x_{query} = v_k$ and $P_{input} \in \mathcal{E}_{bal}^*$, for $n \neq k$, the following holds*

$$
\mathbf{Attn}_n^{(t)} = \Theta\left(\frac{1-\mathbf{Attn}_k^{(t)}}{K}\right).
$$

Proof. By definition,

$$
\mathbf{Attn}_n^{(t)} = \frac{|\mathcal{V}_n|\exp(B_{k,n}^{(t)})}{\sum_{m\neq k}|\mathcal{V}_m|\exp(B_{k,m}^{(t)}) + |\mathcal{V}_k|\exp(A_k^{(t)})}.
$$

By Induction Hypothesis E.3,

$$
e^{-O(\frac{\log(K)-\log(\epsilon)}{K})} \leq \exp(B_{k,m}^{(t)} - B_{k,n}^{(t)}) \leq e^{O(\frac{\log(K)-\log(\epsilon)}{K})}.
$$

Combining with the fact that $-\log(\epsilon) \ll \text{polylog}(K)$, we obtain

$$
\frac{\mathbf{Attn}_n^{(t)}}{1-\mathbf{Attn}_k^{(t)}} = \frac{|\mathcal{V}_n|\exp(B_{k,n}^{(t)})}{\sum_{m\neq k}|\mathcal{V}_m|\exp(B_{k,m}^{(t)})} = \frac{1}{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_n|}\exp(B_{k,m}^{(t)} - B_{k,n}^{(t)})} = \Theta\left(\frac{1}{K}\right).
$$

□

E.4.2. Controlling Gradient Updates in Stage II of Phase II

Lemma E.13. *At each iteration $t \in (\widetilde{T}_{2,k}^{\epsilon}, T_{2,k}^{\epsilon}]$. If Induction Hypothesis E.3 holds for t, then $\alpha_k^{(t)} \geq 0$ and satisfies*

$$
\alpha_k^{(t)} \leq O\left(\frac{\epsilon}{K}\right).
$$

Proof. By the gradient expression in Lemma D.3,

$$
\begin{aligned}
\alpha_k^{(t)} &= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\,\mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m\neq k}\mathbf{Attn}_m^{(t)^2} + (1-\mathbf{Attn}_k^{(t)})^2\right)\right] \\
&\leq p_k \mathbb{E}\left[\mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m\neq k}\mathbf{Attn}_m^{(t)^2} + (1-\mathbf{Attn}_k^{(t)})^2\right) \mid \{x_{\text{query}} = v_k\} \cap \mathcal{E}_{\text{bal}}^*\right]
\end{aligned}
$$

$$
\begin{aligned}
&+ 6p_k \exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right) \\
&\le p_k \cdot O(\epsilon) + 6p_k \exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right) \\
&\le O\left(\frac{\epsilon}{K}\right),
\end{aligned}
$$

where the second inequality follows from Lemmas E.11 and E.12, and the last inequality follows from the fact that $p_k = \Theta\left(\frac{1}{K}\right)$ and $\epsilon = \Omega(\exp(-\operatorname{polylog}(K))) \gg 6\exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right)$. □

Lemma E.14. *At each iteration $t \in (\widetilde{T}_{2,k}^{\epsilon}, T_{2,k}^{\epsilon}]$, if Induction Hypothesis E.3 holds for t, for any $n \neq k$, $\beta_{k,n}^{(t)}$ satisfies*

$$
-O\left(\frac{\alpha_k^{(t)}}{K}\right) \le \beta_{k,n}^{(t)} \le 0.
$$

Proof. Note that conditioned on the event $\{x_{\text{query}} = v_k\} \cap \{P_{\text{input}} \in \mathcal{E}_{\text{bal}}^*\}$, $\mathbf{Attn}_k^{(t)} = \Omega(1)$, and $\max_{m\neq k} \mathbf{Attn}_m^{(t)} = O(\frac{\epsilon^{\frac{1}{2}}}{K})$. Thus,

$$
\begin{aligned}
\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 - \mathbf{Attn}_n^{(t)} - \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)}) &\le \max_{m\neq k} \mathbf{Attn}_m^{(t)} \sum_{m\neq k} \mathbf{Attn}_m^{(t)} - \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)}) \\
&= -(1 - \mathbf{Attn}_k^{(t)})(\mathbf{Attn}_k^{(t)} - \max_{m\neq k} \mathbf{Attn}_m^{(t)}) \\
&\le -\Omega(1 - \mathbf{Attn}_k^{(t)}) \le -\Omega\left(\exp(-\operatorname{polylog}(K))\right).
\end{aligned}
$$

Therefore, by the gradient expression in Lemma D.3 and the fact that $N \gg K^3$,

$$
\beta_{k,n}^{(t)} \le 6\exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right) - \Omega(\exp(-\operatorname{polylog}(K))) < 0.
$$

Moreover, following the analysis similar to that for Lemma E.9, we have

$$
\begin{aligned}
-\beta_{k,n}^{(t)} &\le p_k \mathbb{E}\left[\mathbf{Attn}_n^{(t)} \cdot \left(\mathbf{Attn}_n^{(t)} + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right) \mid \{x_{\text{query}} = v_k\} \cap \mathcal{E}_{\text{bal}}^*\right] + p_k \mathbb{P}({\mathcal{E}_{\text{bal}}^*}^c) \\
&\le p_k \mathbb{E}\left[\Theta(\frac{1 - \mathbf{Attn}_k^{(t)}}{K}) \cdot O\left(\mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right) \mid \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*\right] \\
&\qquad + 6p_k \exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right) \\
&= p_k \mathbb{E}\left[O(\frac{\mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})^2}{K}) \mid \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*\right] + 6p_k \exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right) \\
&\le O\left(\frac{\alpha_k^{(t)}}{K}\right),
\end{aligned}
$$

where the last inequality follows from the gradient expression of $\alpha_k^{(t)}$ in Lemma D.3 and because $\alpha_k^{(t)} \gg 6\exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right)$. □

E.4.3. Controlling Loss in Stage II of Phase II

Lemma E.15. *Given $k \in [K]$, and $0 < \epsilon < 1$, suppose $\operatorname{polylog}(K) \gg \log(\frac{1}{\epsilon})$. At each iteration $t \in (\widetilde{T}_{2,k}^{\epsilon}, T_{2,k}^{\epsilon}]$, if Induction Hypothesis E.3 holds for t, then we have $\widetilde{L}_k(\theta^{(t)}) < \frac{p_k \epsilon}{2}$.*

Proof. By the gradient expression in Lemma D.3, we have

$$\begin{aligned}
\widetilde{L}_k(\theta^{(t)}) &= \frac{1}{2}\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}^*_{\text{bal}}\}\left(\widehat{y}_{\text{query}} - \langle w, x_{\text{query}}\rangle\right)^2\right] \\
&= \frac{1}{2}\mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}^*_{\text{bal}}\}\left(\sum_{m\neq k}\mathbf{Attn}_m^{(t)^2} + (1-\mathbf{Attn}_k^{(t)})^2\right)\right] \\
&\leq \frac{1}{2}p_k\mathbb{P}\left(P_{\text{input}} \in \mathcal{E}^*_{\text{bal}}\right)\cdot\mathbb{E}\left[\left(O\left(\frac{1}{K}\right)+1\right)\left(1-\mathbf{Attn}_k^{(t)}\right)^2\middle| x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}^*_{\text{bal}}\right] \\
&\leq \frac{1}{2}p_k\cdot\left(1+O\left(\frac{1}{K}\right)\right)\cdot\epsilon \\
&\leq \frac{2p_k\epsilon}{3},
\end{aligned}$$

where the first inequality follows from Lemma E.12, and the second inequality follows from Lemma E.11. □

E.4.4. End of Stage II of Phase II

Lemma E.16. *Given $k \in [K]$, and $0 < \epsilon < 1$, suppose $\text{polylog}(K) \gg \log(\frac{1}{\epsilon})$. Then Induction Hypothesis E.3 holds for all $\widetilde{T}^\epsilon_{2,k} < t \leq T^\epsilon_{2,k} = \widetilde{T}^\epsilon_{2,k} + O\left(\frac{K\log\left(K\epsilon^{-\frac{1}{2}}\right)}{\epsilon\eta}\right)$.*

Proof. It is easy to verify Induction Hypothesis E.3 holds at $t = \widetilde{T}^\epsilon_{2,k} + 1$. Now we suppose Induction Hypothesis E.3 holds for all iterations $\widetilde{T}^\epsilon_{2,k} \leq t-1$, and prove it holds at t.

For the first claim, we can upper-bound the update of $A_k^{(t)}$ by Lemma E.13 as follows:

$$\begin{aligned}
A_k^{(t)} &\leq A_k^{(t-1)} + \eta\cdot O(\frac{\epsilon}{K}) \\
&\leq A_k^{(\widetilde{T}^\epsilon_{2,k}+1)} + \eta(t-\widetilde{T}^\epsilon_{2,k}-1)\cdot O(\frac{\epsilon}{K}) \\
&\leq O(\log(K/\epsilon)) + \eta O(\frac{K\log\left(K\epsilon^{-\frac{1}{2}}\right)}{\epsilon\eta})\cdot O(\frac{\epsilon}{K}) \\
&= O(\log(K/\epsilon)).
\end{aligned}$$

The second claim follows from Lemma E.14 and the analysis similar to that for Lemma E.10.

□

E.5. Proof of Theorem 3.2 for Balanced Case

Theorem E.17 (Restatement of Theorem 3.2 for balanced features)**.** *Suppose $p_k = \Theta\left(\frac{1}{K}\right)$ for each $k \in [K]$. For any $0 < \epsilon < 1$, suppose $N \geq \text{poly}(K)$ and $\text{polylog}(K) \gg \log(\frac{1}{\epsilon})$. We apply GD to train the loss function given in Equation (4). Then with at most $T^* = O(\frac{\log(K)K^2}{\eta} + \frac{K\log\left(K\epsilon^{-\frac{1}{2}}\right)}{\epsilon\eta})$ iterations, we have*

1. *The loss converges: $L(\theta^{(T^*)}) - L^* \leq \epsilon$, where $L^* = \Theta(e^{-\text{poly}(K)})$ is the global minimum of the population loss in Equation (4).*
2. *Attention score concentrates: if $x_{query} = v_k$, with probability at least $1 - e^{-\Omega(\text{poly}(K))}$[3], the one-layer transformer nearly "pays all attention" to input tokens featuring v_k, i.e., $(1-\mathbf{Attn}_k^{(T^*)})^2 \leq O(\epsilon)$.*

[3]The randomness originates from the first N input tokens in the test prompt.

Proof. Denote $T^* = \max_{k\in[K]} \widetilde{T}^{\epsilon}_{2,k} + 1 = O(\frac{\log(K)K^2}{\eta} + \frac{K\log\left(K\epsilon^{-\frac{1}{2}}\right)}{\epsilon\eta})$. Thus for any k, at iteration T^*, it is in stage II of the convergence phase, i.e., $T^* \in (\widetilde{T}^{\epsilon}_{2,k}, T^{\epsilon}_{2,k}]$. Then by Lemmas E.15 and E.16, for any $k \in [K]$, we obtain:

$$\widetilde{L}_k(\theta^{(T^*)}) \leq \frac{2p_k\epsilon}{3}.$$

Therefore

$$\begin{aligned}
L(\theta^{(T^*)}) - L^{\text{low}} &= \sum_{k=1}^{K}(L_k(\theta^{(T^*)}) - L_k^{\text{low}}) \\
&\leq \sum_{k=1}^{K}\left(\widetilde{L}_k(\theta^{(T^*)}) + 3p_k\exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right)\right) \\
&\leq \sum_{k=1}^{K}\frac{2p_k\epsilon}{3} + 3\exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right) \\
&\leq \frac{2\epsilon}{3} + 3\exp\left(-\frac{c_{\text{bal}}^2 N}{25K^2}\right) \\
&\leq \epsilon,
\end{aligned}$$

where the first inequality follows from Lemma D.9.

Finally, by Lemma D.8,

$$L(\theta^{(T^*)}) - L^* \leq L(\theta^{(T^*)}) - L^{\text{low}} \leq \epsilon.$$

□

F. Analysis for the Imbalanced Case: Under-represented Features (Proof of Theorem 3.3 Part I)

In this section, we present the analysis of the prediction error when the query token features an under-represented feature v_k with $k > 1$ in the imbalanced case. We first discuss the outline of our proof.

F.1. Roadmap of the Proof

We will analyze the convergence of the training process via four phases of dynamics. At the beginning of each phase, we will establish an induction hypothesis, which we expect to remain valid throughout that phase. Subsequently, we will analyze the dynamics under such a hypothesis within the phase, aiming to provide proof of the hypothesis by the end of the phase.

The main idea of the proof lies in analyzing the GD dynamics of $A_k^{(t)}$ and $B_{k,n}^{(t)}$. From Definition D.2 and Lemma D.3, we have

$$A_k^{(t+1)} = A_k^{(t)} + \eta\alpha_k^{(t)},$$
$$B_{k,n}^{(t+1)} = B_{k,n}^{(t)} + \eta\beta_{k,n}^{(t)},$$

where

$$\alpha_k^{(t)} = \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\,\mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1-\mathbf{Attn}_k^{(t)})^2\right)\right],$$
$$\beta_{k,n}^{(t)} = \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\,\mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 - \mathbf{Attn}_n^{(t)} - \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)})\right)\right].$$

We divide the learning process of the under-represented feature v_k with $k > 1$ into the following four phases.

- **Phase I** ($t \in [0, T_{1,k}]$, Appendix F.2): At initialization, $B_{k,1}^{(t)}$ enjoys a much larger reduction rate, i.e., $\beta_{k,1} < 0$ and $|\beta_{k,1}|$ is large. Therefore, the decrease of $B_{k,1}^{(t)}$ will dominate the dynamics during phase I.
- **Phase II** ($t \in (T_{1,k}, T_{2,k}]$, Appendix F.3): At time $T_{2,k}+1$, the decrease of $B_{k,1}^{(t)}$ becomes slower, and the same happens to $|\beta_{k,1}^{(t)}|$. Their decreasing rate drops to be closer to the increasing rate of $\alpha_k^{(t)}$. This marks the beginning of phase II. Shortly after entering this phase, the previous dominance of reduction of $B_{k,1}^{(t)}$ diminishes, as $|\beta_{k,1}^{(t)}|$ approaches a comparable order of the magnitude to $\alpha_k^{(t)}$. At this point, there is a shift in the leading influence, with the growth of $A_k^{(t)}$ taking over.
- **Phase III** ($t \in (T_{2,k}, T_{3,k}]$, Appendix F.4): Following the transitional phase, $\alpha_k^{(t)}$ grows from the value of $\Theta(\frac{1}{K^{1.5}})$, whereas $|\beta_{k,1}^{(t)}|$ and $|\beta_{k,n}^{(t)}|$ for $n \neq k, 1$ stay at much lower values ($\leq O(\frac{1}{K^{1.98}})$ and $\leq O\left(\frac{1}{K^3}\right)$ respectively). This consistent gap in magnitude between $\alpha_k^{(t)}$ and $\beta_{k,n}^{(t)}$ leads to the continuously rapid growth of $A_k^{(t)}$, while $B_{k,n}^{(t)}$ remains relatively unchanged.
- **Phase IV** ($t \in (T_{3,k}, T_{4,k}^{\epsilon}]$, Appendix F.5): At $t = T_{3,k}+1$, we achieve the desired attention structures for query tokens featuring the under-represented feature v_k. Then we establish a connection between $\alpha_k^{(t)}$ and the prediction error via analyzing the change of $1 - \mathbf{Attn}_k^{(t)}$ that diminishes, leading to the subsequent proof of convergence.

We finally combine all results in the above four phases to prove the main Theorem 3.3 for underrepresented features (Appendix F.6).

F.2. Phase I: Decrease of Dominant Feature

In this section, we will delve into the initial phase of learning dynamics, aiming at mitigating the high occurrence bias of the dominant feature v_1. Specifically, for $k > 1$, $B_{k,1}$ will undergo significant decrease during this phase. Let us begin by defining phase I.

For the k-th feature v_k with $k > 1$, we define phase I as all iterations $t \leq T_{1,k}$, where

$$T_{1,k} \triangleq \max\left\{t : B_{k,1}^{(t)} \geq -0.49\log(K)\right\}.$$

We state the following induction hypothesis, which will hold throughout phase I:

Induction Hypothesis F.1. Given $k > 1$, for each $0 \le t \le T_{1,k}$, the following holds:

a. $A_k^{(t)}$ is monotonically increasing and $A_k^{(t)} \in [0, O(\frac{\log(K)}{K^{0.02}})]$;

b. $B_{k,1}^{(t)}$ is monotonically decreasing and $B_{k,1}^{(t)} \in [-0.49\log(K), 0]$;

c. $|B_{k,n}^{(t)}| = O(\frac{A_k^{(t)} - B_{k,1}^{(t)}}{K})$ and $B_{k,n}^{(t)} > B_{k,1}^{(t)}$ for any $n \neq k, 1$.

F.2.1. Technical Lemmas

We first introduce several technical lemmas that will be used for the proof of Induction Hypothesis F.1.

Lemma F.1. *If Induction Hypothesis F.1 holds at iteration $0 \le t \le T_{1,k}$, for the prompt satisfying $x_{query} = v_k$ and $P_{input} \in \mathcal{E}^*_{imbal}$, the following holds*

1. $\mathbf{Attn}_k^{(t)} = \Theta\left(\frac{1}{K}\right)$;
2. $\mathbf{Attn}_1^{(t)} = \Omega\left(\frac{1}{K^{0.49}}\right)$;
3. $1 - \mathbf{Attn}_1^{(t)} - \mathbf{Attn}_k^{(t)} \ge \Omega(1)$.

Proof. Since $x_{\text{query}} = v_k$, and $|\mathcal{V}_k| > 0$ for $P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}$, we have

$$\begin{aligned}\mathbf{Attn}_k^{(t)} &= \frac{|\mathcal{V}_k| e^{{v_k}^\top Q^{(t)} v_k}}{\sum_{j\in[N]} e^{{E_j^x}^\top Q^{(t)} v_k}} \\ &= \frac{|\mathcal{V}_k| \exp(A_k^{(t)})}{\sum_{m\neq k} |\mathcal{V}_m| \exp(B_{k,m}^{(t)}) + |\mathcal{V}_k| \exp(A_k^{(t)})} \\ &= \frac{1}{\sum_{m\neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} \exp(B_{k,m}^{(t)} - A_k^{(t)}) + 1}.\end{aligned}$$

By Induction Hypothesis F.1, we have

- for $m \neq 1, k$, $e^{-O(\frac{\log(K)}{K^{0.02}})} \le \exp(B_{k,m}^{(t)} - A_k^{(t)}) \le e^{O(\frac{\log(K)}{K})}$;
- for $m = 1$, $e^{\left(-0.49\log(K) - O(\frac{\log(K)}{K^{0.02}})\right)} \le \exp(B_{k,1}^{(t)} - A_k^{(t)}) \le e^0$.

Combining with the fact that $\sum_{m\neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} = \Theta(K)$ for $P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}$, we have

$$\mathbf{Attn}_k^{(t)} \ge \Omega\left(\frac{1}{K}\right).$$

On the other hand, since $\frac{N - |\mathcal{V}_1|}{|\mathcal{V}_k|}$ is still $\Theta(K)$, we have

$$\mathbf{Attn}_k^{(t)} \le \frac{1}{e^{-O(\frac{\log(K)}{K^{0.02}})}(\frac{N-|\mathcal{V}_1|}{|\mathcal{V}_k|} - 1) + e^{\left(-0.49\log(K) - O(\frac{\log(K)}{K^{0.02}})\right)}\frac{|\mathcal{V}_1|}{|\mathcal{V}_k|} + 1} \le O\left(\frac{1}{K}\right).$$

By similar analysis, we have

$$\begin{aligned}\mathbf{Attn}_1^{(t)} &= \frac{|\mathcal{V}_1| \exp(B_{k,1}^{(t)})}{\sum_{m\neq k} |\mathcal{V}_m| \exp(B_{k,m}^{(t)}) + |\mathcal{V}_k| \exp(A_k^{(t)})} \\ &= \frac{1}{\sum_{m\neq 1,k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} \exp(B_{k,m}^{(t)} - B_{k,1}^{(t)}) + \frac{|\mathcal{V}_k|}{|\mathcal{V}_1|} \exp(A_k^{(t)} - B_{k,1}^{(t)}) + 1}.\end{aligned}$$

By Induction Hypothesis F.1,

- for $m \neq 1, k$, we have $e^0 \leq \exp(B_{k,m}^{(t)} - B_{k,1}^{(t)}) \leq e^{0.49\log(K)+O(\frac{\log(K)}{K})}$;
- $e^0 \leq \exp(A_k^{(t)} - B_{k,1}^{(t)}) \leq e^{0.49\log(K)+O(\frac{\log(K)}{K})}$.

Hence,

$$\mathbf{Attn}_1^{(t)} \geq \frac{1}{e^{0.49\log(K)+O(\frac{\log(K)}{K})}(\frac{N}{|\mathcal{V}_1|}-1)+1} \geq \Omega\left(\frac{1}{K^{0.49}}\right),$$

where the last inequality holds since $\frac{N}{|\mathcal{V}_1|} = \Theta(1)$ for $P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}$.

For the last statement,

$$1 - \mathbf{Attn}_1^{(t)} \geq \frac{e^0(\frac{N}{|\mathcal{V}_1|}-1)}{e^0(\frac{N}{|\mathcal{V}_1|}-1)+1} \geq \Omega(1).$$

Combining with the fact that $\mathbf{Attn}_k^{(t)} = \Theta\left(\frac{1}{K}\right)$, we have

$$1 - \mathbf{Attn}_k^{(t)} - \mathbf{Attn}_1^{(t)} \geq \Omega(1).$$

□

Lemma F.2. *If Induction Hypothesis F.1 holds at iteration $0 \leq t \leq T_{1,k}$, for the prompt satisfying $x_{query} = v_k$ and $P_{input} \in \mathcal{E}^*_{imbal}$, the following holds*

$$\mathbf{Attn}_n^{(t)} = O\left(\frac{1 - \mathbf{Attn}_k^{(t)} - \mathbf{Attn}_1^{(t)}}{K}\right).$$

Proof. Since $x_{\text{query}} = v_k$, we have

$$\begin{aligned}\mathbf{Attn}_n^{(t)} &= \frac{|\mathcal{V}_n| e^{{v_n}^\top Q^{(t)} v_k}}{\sum_{j\in[N]} e^{{E_j^x}^\top Q^{(t)} v_k}} \\ &= \frac{|\mathcal{V}_n| \exp(B_{k,n}^{(t)})}{\sum_{m\neq k} |\mathcal{V}_m| \exp(B_{k,m}^{(t)}) + |\mathcal{V}_k| \exp(A_k^{(t)})}.\end{aligned}$$

By Induction Hypothesis F.1, for $m, n \neq 1$,

$$e^{-O(\frac{\log(K)}{K})} \leq \exp(B_{k,m}^{(t)} - B_{k,n}^{(t)}) \leq e^{O(\frac{\log(K)}{K})}.$$

Combining with the fact that $\frac{|\mathcal{V}_m|}{|\mathcal{V}_n|} = \Theta(1)$ for $P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}$, we have

$$\begin{aligned}\frac{\mathbf{Attn}_n^{(t)}}{1 - \mathbf{Attn}_k^{(t)} - \mathbf{Attn}_1^{(t)}} &= \frac{|\mathcal{V}_n| \exp(B_{k,n}^{(t)})}{\sum_{m\neq 1,k} |\mathcal{V}_m| \exp(B_{k,m}^{(t)})} \\ &= \frac{1}{\sum_{m\neq k,1} \frac{|\mathcal{V}_m|}{|\mathcal{V}_n|} \exp(B_{k,m}^{(t)} - B_{k,n}^{(t)})} \\ &\leq O\left(\frac{1}{K}\right).\end{aligned}$$

□

F.2.2. Controlling Gradient Updates in Phase I

Lemma F.3. *Given $k > 1$, if Induction Hypothesis F.1 holds at iteration $0 \le t \le T_{1,k}$, then $\alpha_k^{(t)} \ge 0$ and satisfies*

$$\alpha_k^{(t)} = \Theta\left(\frac{1}{K^2}\right).$$

Proof. By the gradient expression in Lemma D.3, we have

$$\begin{aligned}
\alpha_k^{(t)} &= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\,\mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1-\mathbf{Attn}_k^{(t)})^2\right)\right] \\
&= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap \mathcal{E}^*_{\text{imbal}}\}\,\mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1-\mathbf{Attn}_k^{(t)})^2\right)\right] \\
&\quad + \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap {\mathcal{E}^*_{\text{imbal}}}^c\}\,\mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1-\mathbf{Attn}_k^{(t)})^2\right)\right] \\
&\overset{(a)}{\le} p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}})\mathbb{E}\left[\mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1-\mathbf{Attn}_k^{(t)})^2\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
&\quad + 2p_k \cdot \mathbb{P}(P_{\text{input}} \in {\mathcal{E}^*_{\text{imbal}}}^c) \\
&\overset{(b)}{\le} p_k \cdot \mathbb{E}\left[\mathbf{Attn}_k^{(t)} \cdot \left(O\left(\frac{1}{K}\right) + {\mathbf{Attn}_1^{(t)}}^2 + (1-\mathbf{Attn}_k^{(t)})^2\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
&\quad + 2p_k \cdot \mathbb{P}(P_{\text{input}} \in {\mathcal{E}^*_{\text{imbal}}}^c) \\
&\overset{(c)}{\le} O\left(\frac{1}{K^2}\right),
\end{aligned}$$

where (a) follows from the fact that x_{query} and P_{input} are independently sampled, and $\mathbf{Attn}_k^{(t)} \cdot (\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1 - \mathbf{Attn}_k^{(t)})^2)$ is upper-bounded by 2 on the event $\{P_{\text{input}} \in {\mathcal{E}^*_{\text{imbal}}}^c\}$, (b) follows by applying Lemma F.2 to $\mathbf{Attn}_m^{(t)}$ for $m \neq 1, k$, and (c) follows from Lemma F.1, our choice of p_k, Lemma D.6, and the evident bound:

$$3\exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \ll O\left(\frac{1}{K}\right).$$

Similarly, we can show that $\alpha_k^{(t)} \ge \Omega\left(\frac{1}{K^2}\right)$. □

Lemma F.4. *Given $k > 1$, if Induction Hypothesis F.1 holds at iteration $0 \le t \le T_{k,1}$, then $\beta_{k,1}^{(t)} < 0$ satisfies*

$$|\beta_{k,1}^{(t)}| \ge \Omega\left(\frac{1}{K^{1.98}}\right).$$

Proof. We first derive

$$\begin{aligned}
&\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 - \mathbf{Attn}_1^{(t)} - \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)}) \\
&\quad = \sum_{m\neq 1,k} {\mathbf{Attn}_m^{(t)}}^2 - \mathbf{Attn}_1^{(t)}(1-\mathbf{Attn}_1^{(t)}) - \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)}) \\
&\quad \le \max_{m\neq 1,k} \mathbf{Attn}_m^{(t)}(1-\mathbf{Attn}_1^{(t)} - \mathbf{Attn}_k^{(t)}) - \mathbf{Attn}_1^{(t)}(1-\mathbf{Attn}_1^{(t)}) - \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)}) \\
&\quad \le -(1-\mathbf{Attn}_k^{(t)} - \mathbf{Attn}_1^{(t)})(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)} - \max_{m\neq 1,k} \mathbf{Attn}_m^{(t)}). \qquad (21)
\end{aligned}$$

Therefore, by the gradient expression in Lemma D.3, we have

$$\begin{aligned}
\beta_{k,1}^{(t)} \leq & \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\}\,\mathbf{Attn}_1^{(t)} \cdot \left(\sum_{m\neq k} \mathbf{Attn}_m^{(t)\,2} - \mathbf{Attn}_1^{(t)} - \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)})\right)\right] \\
& + \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}^{*\,c}_{\text{imbal}}\}\,\mathbf{Attn}_1^{(t)} \cdot \left(\sum_{m\neq k} \mathbf{Attn}_m^{(t)\,2}\right)\right] \\
\overset{(a)}{\leq} & p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^{*\,c}_{\text{imbal}}) + p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}) \\
& \cdot \mathbb{E}\left[-\Omega\left(\frac{(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)} - \max_{m\neq 1,k} \mathbf{Attn}_m^{(t)})}{K^{0.49}}\right)\middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
\leq & p_k \cdot \left(-\Omega\left(\frac{1}{K^{0.98}}\right)\right) + 3p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\
= & -\Omega\left(\frac{1}{K^{1.98}}\right),
\end{aligned}$$

where (a) follows from Equation (21) and Lemma F.1, and the last equality holds since

$$\frac{1}{K^{0.98}} \gg \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right).$$

□

Lemma F.5. *If Induction Hypothesis F.1 holds at iteration $0 \leq t \leq T_{k,1}$, for any $n \neq 1, k$, $\beta_{k,n}^{(t)}$ satisfies*

$$|\beta_{k,n}^{(t)}| \leq O\left(\frac{\alpha_k^{(t)} - \beta_{k,1}^{(t)}}{K}\right).$$

Proof. By the gradient expression in Lemma D.3, we have

$$\beta_{k,n}^{(t)} \leq \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\,\mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m\neq k} \mathbf{Attn}_m^{(t)\,2}\right)\right] \tag{22}$$

$$-\beta_{k,n}^{(t)} \leq \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\,\mathbf{Attn}_n^{(t)} \cdot \left(\mathbf{Attn}_n^{(t)} + \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)})\right)\right] \tag{23}$$

We further upper-bound Equation (22) as,

$$\begin{aligned}
\beta_{k,n}^{(t)} \leq & \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\}\,\mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m\neq k} \mathbf{Attn}_m^{(t)\,2}\right)\right] \\
& + \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}^{*\,c}_{\text{imbal}}\}\,\mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m\neq k} \mathbf{Attn}_m^{(t)\,2}\right)\right] \\
\leq & p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}) \cdot \mathbb{E}\left[\mathbf{Attn}_n^{(t)} \cdot \left(\mathbf{Attn}_1^{(t)\,2} + O\left(\frac{1}{K}\right)\right) \mid \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
& + p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^{*\,c}_{\text{imbal}}) \\
\overset{(a)}{\leq} & O\left(\frac{1}{K^3}\right) + O\left(\frac{|\beta_{k,1}^{(t)}|}{K}\right) + 3p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\
\leq & O\left(\frac{1}{K^3}\right) + O\left(\frac{|\beta_{k,1}^{(t)}|}{K}\right).
\end{aligned}$$

where (a) follows from the following two observations from Lemma F.2:

$$|\beta_{k,1}^{(t)}| \geq p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}) \cdot \mathbb{E}\left[\Omega\left(\mathbf{Attn}_1^{(t)^2}\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right],$$

and $\mathbf{Attn}_n^{(t)} \leq O\left(\frac{1}{K}\right)$.

To further upper-bound Equation (23), we have

$$\begin{aligned}
&-\beta_{k,n}^{(t)} \\
&\leq p_k \mathbb{E}\left[\mathbf{Attn}_n^{(t)} \cdot \left(\mathbf{Attn}_n^{(t)} + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] + p_k \cdot \mathbb{P}(\mathcal{E}^*_{\text{imbal}}{}^c) \\
&\overset{(a)}{\leq} p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}{}^c) + p_k \cdot \mathbb{P}(\mathcal{E}^*_{\text{imbal}}) \mathbb{E}\left[O\left(\frac{1 - \mathbf{Attn}_k^{(t)}}{K}\right)\right. \\
&\qquad \left. \cdot \left(O\left(\frac{1 - \mathbf{Attn}_k^{(t)}}{K}\right) + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
&\leq p_k \cdot \mathbb{P}(\mathcal{E}^*_{\text{imbal}}) \mathbb{E}\left[O(\frac{\mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})^2}{K}) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] + 3p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\
&\leq O\left(\frac{\alpha_k^{(t)}}{K}\right),
\end{aligned}$$

where (a) follows from Lemma F.2, and the last inequality follows from the analysis in the proof of Lemma F.3, and from the fact that

$$\alpha_k^{(t)} \geq \Omega\left(\frac{1}{K^2}\right) \gg 3 \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right).$$

Thus, we obtain

$$|\beta_{k,n}^{(t)}| \leq O\left(\frac{\alpha_k^{(t)} - \beta_{k,1}^{(t)}}{K}\right).$$

□

F.2.3. End of Phase I

Lemma F.6. *Given $k \geq 2$, Induction Hypothesis F.1 holds for all $t \leq T_{1,k} = O(\frac{\log(K)K^{1.98}}{\eta})$, and at iteration $t = T_{1,k} + 1$, we have*

a. $B_{k,1}^{(T_{1,k}+1)} \leq -0.49 \log(K)$;

b. $\mathbf{Attn}_1 = O\left(\frac{1}{K^{0.49}}\right)$ *if* $x_{query} = v_k$ *and* $P_{input} \in \mathcal{E}^*_{imbal}$.

Proof. The existence of $T_{1,k} = O(\frac{\log(K)K^{1.98}}{\eta})$ directly follows from Lemma F.3.

It is easy to verify that Induction Hypothesis F.1 holds at $t = 0$. Now we suppose Induction Hypothesis F.1 holds for all iterations $\leq t - 1$, and prove it holds at t.

By Lemma F.3, we have $\alpha_k^{(t-1)} \geq 0$. Thus $A_k^{(t)} = A_k^{(t-1)} + \eta\alpha_k^{(t-1)} \geq 0$. Moreover, combining Lemmas F.3 and F.4, we obtain $A_k^{(t)} - A_k^{(0)} \leq O(\frac{|B_{k,1}^{(t)} - B_{k,1}^{(0)}|}{K^{0.02}})$ which further implies $A_k^{(t)} \leq O(\log(K)/K^{0.02})$.

For $m \neq 1, k$, by Lemma F.5, we have

$$|B_{k,m}^{(t)}| \leq O(\frac{A_k^{(t)} - A_k^{(0)} + |B_{k,1}^{(t)} - B_{k,1}^{(0)}|}{K}) \leq O(\log(K)/K).$$

The proof for the second statement is deferred to the next phase (Lemma F.7). □

F.3. Phase II: Switching of Leading Influence

During phase I, $B_{k,1}^{(t)}$ significantly decreases, resulting in a decrease in $\mathbf{Attn}_1^{(t)}$, while other $\mathbf{Attn}_n^{(t)}$ with $n > 1$ remain approximately at the order of $\Theta\left(\frac{1}{K}\right)$. By the end of phase I, $(\mathbf{Attn}_1^{(t)})^2$ decreases to $O(\frac{1}{K^{0.98}})$, leading to a decrease in $|\beta_{k,1}^{(t)}|$ as it approaches towards $\alpha_k^{(t)}$. At this point, phase II begins. Shortly after entering this phase, the prior dominant role of the decrease of $B_{k,1}^{(t)}$ in learning dynamics diminishes as $|\beta_{k,1}^{(t)}|$ reaches the same order of magnitude as $\alpha_k^{(t)}$.

For $k > 1$, define

$$T_{2,k} \triangleq \max\{t > T_{1,k} : A_k^{(t)} - B_{k,1}^{(t)} \leq 1.01\log(K)\}.$$

We next state the following induction hypothesis which holds during phase II.

Induction Hypothesis F.2. For $T_{1,k} < t \leq T_{2,k}$, the following holds

a. $A_k^{(t)}$ is monotonically increasing and $A_k^{(t)} \in [0, 0.52\log(K)]$;

b. $B_{k,1}^{(t)}$ is monotonically decreasing and $B_{k,1}^{(t)} \in [-0.51\log(K), -0.49\log(K)]$;

c. $|B_{k,n}^{(t)}| = O(\frac{A_k^{(t)} + |B_{k,1}^{(t)}|}{K})$ for any $n \neq 1, k$.

F.3.1. Technical Lemmas

We first introduce several technical lemmas that will be used for the proof of Induction Hypothesis F.2.

Lemma F.7. *Suppose Induction Hypothesis F.2 holds at iteration $T_{1,k} < t \leq T_{2,k}$. If $x_{query} = v_k$ and $P_{input} \in \mathcal{E}^*_{imbal}$, the following holds*

1. $\mathbf{Attn}_k^{(t)} \in [\Omega\left(\frac{1}{K}\right), O(\frac{1}{K^{0.48}})]$;
2. $\mathbf{Attn}_1^{(t)} \in [\Omega\left(\frac{1}{K^{0.51}}\right), O\left(\frac{1}{K^{0.49}}\right)]$;
3. $1 - \mathbf{Attn}_1^{(t)} - \mathbf{Attn}_k^{(t)} \geq \Omega(1)$.

Proof. Since $x_{\text{query}} = v_k$, and $|\mathcal{V}_k| > 0$ for $P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}$, we have

$$\begin{aligned}
\mathbf{Attn}_k^{(t)} &= \frac{|\mathcal{V}_k| e^{{v_k}^\top Q^{(t)} v_k}}{\sum_{j\in[N]} e^{{E_j^x}^\top Q^{(t)} v_k}} \\
&= \frac{|\mathcal{V}_k| \exp(A_k^{(t)})}{\sum_{m\neq k} |\mathcal{V}_m| \exp(B_{k,m}^{(t)}) + |\mathcal{V}_k| \exp(A_k^{(t)})} \\
&= \frac{1}{\sum_{m\neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} \exp(B_{k,m}^{(t)} - A_k^{(t)}) + 1}.
\end{aligned}$$

By Induction Hypothesis F.2,

- for $m \neq 1, k$, we have $e^{-O(\frac{\log(K)}{K}) - 0.52\log(K)} \leq \exp(B_{k,m}^{(t)} - A_k^{(t)}) \leq e^{O(\frac{\log(K)}{K})}$;
- for $m = 1$, $e^{-1.01\log(K)} \leq \exp(B_{k,1}^{(t)} - A_k^{(t)}) \leq e^0$.

Combining with the fact that $\sum_{m\neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} = \Theta(K)$ for $P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}$, we obtain

$$\mathbf{Attn}_k^{(t)} \geq \Omega\left(\frac{1}{K}\right).$$

Moreover, since $\frac{N - |\mathcal{V}_1|}{|\mathcal{V}_k|}$ is still at the order of $\Theta(K)$, we have

$$\mathbf{Attn}_k^{(t)} \leq \frac{1}{e^{-O(\frac{\log(K)}{K}) - 0.52\log(K)}(\frac{N-|\mathcal{V}_1|}{|\mathcal{V}_k|} - 1) + e^{-1.01\log(K)}\frac{|\mathcal{V}_1|}{|\mathcal{V}_k|} + 1} \leq O\left(\frac{1}{K^{0.48}}\right).$$

We next analyze $\mathbf{Attn}_1^{(t)}$ as

$$\mathbf{Attn}_1^{(t)} = \frac{|\mathcal{V}_1|\exp(B_{k,1}^{(t)})}{\sum_{m\neq k}|\mathcal{V}_m|\exp(B_{k,m}^{(t)}) + |\mathcal{V}_k|\exp(A_k^{(t)})} = \frac{1}{\sum_{m\neq 1,k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_k|}\exp(B_{k,m}^{(t)} - B_{k,1}^{(t)}) + \frac{|\mathcal{V}_k|}{|\mathcal{V}_1|}\exp(A_k^{(t)} - B_{k,1}^{(t)}) + 1}.$$

By Induction Hypothesis F.2,

- for $m \neq 1, k$, we have

$$e^{0.49\log(K) - O(\frac{\log(K)}{K})} \leq \exp(B_{k,m}^{(t)} - B_{k,1}^{(t)}) \leq e^{0.51\log(K) + O(\frac{\log(K)}{K})};$$

- for $m = 1$, $e^{0.49\log(K)} \leq \exp(A_k^{(t)} - B_{k,1}^{(t)}) \leq e^{1.01\log(K)}$.

Thus, we obtain

$$\mathbf{Attn}_1^{(t)} \geq \frac{1}{e^{0.51\log(K)+O(\frac{\log(K)}{K})}(\frac{N-|\mathcal{V}_k|}{|\mathcal{V}_1|} - 1) + e^{1.01\log(K)+O(\frac{\log(K)}{K})}\frac{|\mathcal{V}_k|}{|\mathcal{V}_1|} + 1} \geq \Omega\left(\frac{1}{K^{0.51}}\right).$$

On the other hand,

$$1 - \mathbf{Attn}_1^{(t)} \geq \frac{e^{0.49\log(K)-O(\frac{\log(K)}{K})}(\frac{N}{|\mathcal{V}_1|} - 1)}{e^{0.49\log(K)-O(\frac{\log(K)}{K})}(\frac{N}{|\mathcal{V}_1|} - 1) + 1} \geq \Omega(1).$$

Thus, we obtain

$$1 - \mathbf{Attn}_k^{(t)} - \mathbf{Attn}_1^{(t)} \geq \Omega(1).$$

□

Lemma F.8. *Suppose Induction Hypothesis F.2 holds at iteration $T_{1,k} < t \leq T_{2,k}$. If $x_{query} = v_k$ and $P_{input} \in \mathcal{E}^*_{imbal}$, for $n \neq 1, k$, the following holds*

$$\mathbf{Attn}_n^{(t)} = O\left(\frac{1 - \mathbf{Attn}_k^{(t)} - \mathbf{Attn}_1^{(t)}}{K}\right).$$

Proof. Since $x_{\text{query}} = v_k$, we have

$$\mathbf{Attn}_n^{(t)} = \frac{|\mathcal{V}_n| e^{{v_n}^\top Q^{(t)} v_k}}{\sum_{j\in[N]} e^{{E_j^x}^\top Q^{(t)} v_k}} = \frac{|\mathcal{V}_n|\exp(B_{k,n}^{(t)})}{\sum_{m\neq k}|\mathcal{V}_m|\exp(B_{k,m}^{(t)}) + |\mathcal{V}_k|\exp(A_k^{(t)})}.$$

By Induction Hypothesis F.2, for $m, n \neq 1$,

$$e^{-O(\frac{\log(K)}{K})} \leq \exp(B_{k,m}^{(t)} - B_{k,n}^{(t)}) \leq e^{O(\frac{\log(K)}{K})}.$$

Combining with the fact that $\frac{|\mathcal{V}_m|}{|\mathcal{V}_n|} = \Theta(1)$ for $P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}$, we obtain

$$\frac{\mathbf{Attn}_n^{(t)}}{1 - \mathbf{Attn}_k^{(t)} - \mathbf{Attn}_1^{(t)}} = \frac{|\mathcal{V}_n|\exp(B_{k,n}^{(t)})}{\sum_{m\neq 1,k}|\mathcal{V}_m|\exp(B_{k,m}^{(t)})}$$

$$= \frac{1}{\sum_{m\neq k,1} \frac{|\mathcal{V}_m|}{|\mathcal{V}_n|} \exp(B_{k,m}^{(t)} - B_{k,n}^{(t)})}$$
$$\leq O\left(\frac{1}{K}\right).$$

□

F.3.2. Controlling Gradient Updates in Phase II

Lemma F.9. *Given $k > 1$, if Induction Hypothesis F.2 holds at iteration $T_{1,k} < t \leq T_{2,k}$, then $\alpha_k^{(t)} \geq 0$ and satisfies*

$$\alpha_k^{(t)} \geq \Omega\left(\frac{1}{K^2}\right).$$

Proof. By the gradient expression in Lemma D.3, we have

$$\begin{aligned}
\alpha_k^{(t)} &= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\,\mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1 - \mathbf{Attn}_k^{(t)})^2\right)\right] \\
&= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap \mathcal{E}^*_{\text{imbal}}\}\,\mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1 - \mathbf{Attn}_k^{(t)})^2\right)\right] \\
&\quad + \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap {\mathcal{E}^*_{\text{imbal}}}^c\}\,\mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1 - \mathbf{Attn}_k^{(t)})^2\right)\right] \\
&\geq p_k \cdot \mathbb{P}(P \in \mathcal{E}^*_{\text{imbal}})\mathbb{E}\left[\mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1 - \mathbf{Attn}_k^{(t)})^2\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
&\geq \Omega\left(\frac{1}{K^2}\right),
\end{aligned}$$

where the last inequality follows from Lemma D.6, Lemma F.7 and our choice of p_k. □

Lemma F.10. *Given $k > 1$, if Induction Hypothesis F.2 holds at iteration $T_{k,1} \leq t \leq T_{k,2}$, then $\beta_{k,1}^{(t)} < 0$ and satisfies*

$$|\beta_{k,1}^{(t)}| \in \left[\Omega\left(\frac{1}{K^{2.02}}\right), O\left(\frac{1}{K^{1.97}}\right)\right].$$

Proof. Following the computations similar to those in Lemma F.4, we have

$$\begin{aligned}
&\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 - \mathbf{Attn}_1^{(t)} - \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)}) \\
&\quad \leq -(1 - \mathbf{Attn}_k^{(t)} - \mathbf{Attn}_1^{(t)})(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)} - \max_{m\neq 1,k} \mathbf{Attn}_m^{(t)}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\beta_{k,1}^{(t)} \\
&\leq \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap \mathcal{E}^*_{\text{imbal}}\}\,\mathbf{Attn}_1^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 - \mathbf{Attn}_1^{(t)} - \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right)\right] \\
&\quad + \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap {\mathcal{E}^*_{\text{imbal}}}^c\}\,\mathbf{Attn}_1^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2\right)\right]
\end{aligned}$$

$$
\begin{aligned}
&\overset{(a)}{\leq} p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}) \cdot \mathbb{E}\left[-\Omega(\frac{(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)} - \max_{m\neq 1,k} \mathbf{Attn}_m^{(t)})}{K^{0.51}}) \mid \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
&\quad + p_k \cdot \mathbb{P}(P_{\text{input}} \in {\mathcal{E}^*_{\text{imbal}}}^c) \\
&\overset{(b)}{\leq} p_k \cdot \left(-\Omega\left(\frac{1}{K^{1.02}}\right)\right) + 3p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\
&= -\Omega\left(\frac{1}{K^{2.02}}\right),
\end{aligned}
$$

where (a) follows from Lemma F.7, (b) follows from Lemma F.7 and Lemma D.6, and the last inequality holds since

$$
\frac{1}{K^{1.02}} \gg \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right).
$$

Moreover,

$$
\begin{aligned}
-\beta_{k,1}^{(t)} \leq& \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap \mathcal{E}^*_{\text{imbal}}\}\, \mathbf{Attn}_1^{(t)} \cdot \left(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right)\right] \\
&+ \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap {\mathcal{E}^*_{\text{imbal}}}^c\}\, \mathbf{Attn}_1^{(t)} \cdot \left(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right)\right] \\
\overset{(a)}{\leq}& p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}) \cdot \mathbb{E}\left[\mathbf{Attn}_1^{(t)} \cdot O(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)}) \mid \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
&+ 2p_k \cdot \mathbb{P}(P_{\text{input}} \in {\mathcal{E}^*_{\text{imbal}}}^c) \\
\overset{(b)}{\leq}& p_k \cdot \left(O\left(\frac{1}{K^{0.97}}\right)\right) + 6p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\
=& O\left(\frac{1}{K^{1.97}}\right),
\end{aligned}
$$

where (a) follows because $\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})$ is upper-bounded by 2 on the event $\{P_{\text{input}} \in {\mathcal{E}^*_{\text{imbal}}}^c\}$, and (b) follows from Lemma F.7. □

Lemma F.11. *If Induction Hypothesis F.2 holds at iteration $T_{1,k} < t \leq T_{2,k}$, for any $n \neq 1, k$, $\beta_{k,n}^{(t)}$ satisfies*

$$
|\beta_{k,n}^{(t)}| \leq O\left(\frac{\alpha_k^{(t)} - \beta_{k,1}^{(t)}}{K}\right).
$$

Proof. By the gradient expression in Lemma D.3, we have

$$
\beta_{k,n}^{(t)} \leq \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\, \mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2\right)\right], \tag{24}
$$

$$
-\beta_{k,n}^{(t)} \leq \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\, \mathbf{Attn}_n^{(t)} \cdot \left(\mathbf{Attn}_n^{(t)} + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right)\right]. \tag{25}
$$

To further bound Equation (24), we have

$$
\begin{aligned}
\beta_{k,n}^{(t)} \leq& \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\}\, \mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2\right)\right] \\
&+ \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in {\mathcal{E}^*_{\text{imbal}}}^c\}\, \mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2\right)\right]
\end{aligned}
$$

$$
\begin{aligned}
&\leq p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}) \cdot \mathbb{E}\left[\mathbf{Attn}_n^{(t)} \cdot \left(\mathbf{Attn}_1^{(t)^2} + O\left(\frac{1}{K}\right)\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
&\quad + p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}{}^c) \\
&\leq O\left(\frac{1}{K^3}\right) + O\left(\frac{|\beta_{k,1}^{(t)}|}{K}\right) + 3p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\
&\leq O\left(\frac{1}{K^3}\right) + O\left(\frac{|\beta_{k,1}^{(t)}|}{K}\right).
\end{aligned}
$$

To further bound Equation (25), we have

$$
\begin{aligned}
&-\beta_{k,n}^{(t)} \\
&\leq p_k \mathbb{E}\left[\mathbf{Attn}_n^{(t)} \cdot \left(\mathbf{Attn}_n^{(t)} + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right) \mid \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] + 2p_k \cdot \mathbb{P}(\mathcal{E}^*_{\text{imbal}}{}^c) \\
&= 2p_k \cdot \mathbb{P}(\mathcal{E}^*_{\text{imbal}}{}^c) + p_k \cdot \mathbb{P}(\mathcal{E}^*_{\text{imbal}}) \mathbb{E}\left[O\left(\frac{1 - \mathbf{Attn}_k^{(t)}}{K}\right)\right. \\
&\qquad \left. \cdot \left(O\left(\frac{1 - \mathbf{Attn}_k^{(t)}}{K}\right) + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
&\leq p_k \cdot \mathbb{P}(\mathcal{E}^*_{\text{imbal}}) \mathbb{E}\left[O\left(\frac{\mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})^2}{K}\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] + 6p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\
&\leq O(\frac{\alpha_k^{(t)}}{K}).
\end{aligned}
$$

Following from the analysis in Lemma F.9, we have

$$
\alpha_k^{(t)} \geq \Omega\left(\frac{1}{K^2}\right) \gg 6 \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right).
$$

Thus, we obtain

$$
|\beta_{k,n}^{(t)}| \leq O\left(\frac{\alpha_k^{(t)} - \beta_{k,1}^{(t)}}{K}\right).
$$

□

F.3.3. End of Phase II

Lemma F.12. *Given $k \geq 2$, Induction Hypothesis F.2 holds for all $T_{1,k} < t \leq T_{2,k} = T_{1,k} + O(\frac{\log(K)K^2}{\eta})$, and at iteration $t = T_{2,k} + 1$, we have*

a. $A_k^{(T_{2,k}+1)} \geq 0.5 \log(K)$;

b. $B_k^{(T_{2,k}+1)} \geq -0.51 \log(K)$.

Proof. The existence of $T_{2,k} = T_{1,k} + O(\frac{\log(K)K^2}{\eta})$ directly follows from Lemmas F.9 and F.10.

It is easy to verify that Induction Hypothesis F.2 holds at $T_{1,k} + 1$. Now we suppose Induction Hypothesis F.2 holds for all iterations $\leq t - 1$, and prove that it holds at t.

For $m \neq 1, k$, by Lemma F.11, we have

$$
|B_{k,m}^{(t)}| \leq |B_{k,m}^{(T_{1,k}+1)}| + O(\frac{A_k^{(T_{2,k})} - A_k^{(T_{1,k}+1)} + |B_{k,1}^{(T_{2,k})} - B_{k,1}^{(T_{1,k}+1)}|}{K}) \leq O(\log(K)/K).
$$

Now suppose $A_k^{(T_{2,k}+1)} < 0.5\log(K)$, then $B_{k,1}^{(T_{2,k}+1)} < -0.51\log(K)$. Denote the first time that $B_{k,1}^{(t)}$ reaches $-0.501\log(K)$ as $\widetilde{T}$. Note that $\widetilde{T} < T_{2,k}^{(t)}$ since $\beta_{k,1}^{(t)}$, the change of $B_{k,1}^{(t)}$, satisfies $|\beta_{k,1}^{(t)}| \ll \log(K)$. Then for $t \geq \widetilde{T}$, if $x_{\text{query}} = v_k$ and $P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}$, the following holds:

1. $\mathbf{Attn}_k^{(t)} \in [\Omega\left(\frac{1}{K}\right), O(\frac{1}{K^{0.5}})]$;
2. $\mathbf{Attn}_1^{(t)} \leq O(\frac{1}{K^{0.501}})$.

Therefore, following the analysis similar to those for Lemma F.10, we have

$$\begin{aligned}
|\beta_{k,1}^{(t)}| \leq& \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap \mathcal{E}^*_{\text{imbal}}\}\, \mathbf{Attn}_1^{(t)} \cdot \left(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right)\right] \\
&+ \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap {\mathcal{E}^*_{\text{imbal}}}^c\}\, \mathbf{Attn}_1^{(t)} \cdot \left(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right)\right] \\
\leq& p_k \cdot \mathbb{P}(P \in \mathcal{E}^*_{\text{imbal}}) \cdot \mathbb{E}\left[\mathbf{Attn}_1^{(t)} \cdot O(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)}) \mid \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
&+ 2p_k \cdot \mathbb{P}({\mathcal{E}^*_{\text{imbal}}}^c) \\
\leq& p_k \cdot \left(O\left(\frac{1}{K^{1.002}}\right)\right) + O\left(\frac{\alpha_k^{(t)}}{K^{0.501}}\right) + 6p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\
\leq& O\left(\frac{\alpha_k^{(t)}}{K^{0.002}}\right),
\end{aligned}$$

where the last inequality follows from Lemma F.9.

Since $|B_{k,1}^{(T_{2,k}+1)} - B_{k,1}^{(\widetilde{T})}| \geq \Omega(\log(K))$, we have

$$A_k^{(T_{2,k}+1)} \geq |B_{k,1}^{(T_{2,k}+1)} - B_{k,1}^{(\widetilde{T})}| \cdot \Omega(K^{0.002}) + A_k^{(\widetilde{T})} \gg \Omega(K^{0.002}\log(K)),$$

which contradicts the assumption that $A_k^{(T_{2,k}+1)} < 0.5\log(K)$. Therefore, $A_k^{(T_{2,k}+1)} \geq 0.5\log(K)$. Noting that once $B_{k,1}^{(t)}$ drops below $-0.501\log(K)$, it will change much smaller compared to the increase of $A_k^{(t)}$. Thus, $B_{k,1}^{(T_{2,k}+1)} \geq -0.51\log(K)$.

□

F.4. Phase III: Growth of Target Feature

After the transition phase, $A_k^{(t)}$ will experience a larger gradient, with the growth of $A_k^{(t)}$ becoming the dominant effect in this phase. For the k-th feature v_k, we define phase III as all iterations $T_{2,k} < t \leq T_{3,k}$, where

$$T_{3,k} \triangleq \max\left\{t > T_{2,k} : A_k^{(t)} \leq \log(K)\right\}.$$

We state the following induction hypothesis, which will hold throughout phase III.

Induction Hypothesis F.3. For each $T_{2,k} < t \leq T_{3,k}$, the following holds:

a. $A_k^{(t)}$ is monotonically increasing and $A_k^{(t)} \in [0.5\log(K), \log(K)]$;

b. $B_{k,1}^{(t)}$ is monotonically decreasing and $B_{k,1}^{(t)} \in [-0.51\log(K) - O(\frac{\log(K)}{K^{0.48}}), -0.49\log(K)]$;

c. $|B_{k,n}^{(t)}| = O(\frac{A_k^{(t)} + |B_{k,1}^{(t)}|}{K})$ for any $n \neq 1, k$.

F.4.1. Technical Lemmas

We first introduce several useful technical lemmas.

Lemma F.13. *Suppose Induction Hypothesis F.3 holds at iteration $T_{k,2} < t \leq T_{k,3}$. If $x_{query} = v_k$ and $P_{input} \in \mathcal{E}^*_{imbal}$, then the following holds*

1. $\mathbf{Attn}_k^{(t)} = \Omega\left(\frac{1}{K^{0.5}}\right)$;
2. $\mathbf{Attn}_1^{(t)} \in [\Omega\left(\frac{1}{K^{0.51}}\right), O\left(\frac{1}{K^{0.49}}\right)]$;
3. $1 - \mathbf{Attn}_k^{(t)} \geq \Omega(1)$.

Proof. Since $x_{\text{query}} = v_k$, and $|\mathcal{V}_k| > 0$ for $P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}$, we have

$$\begin{aligned}\mathbf{Attn}_k^{(t)} &= \frac{|\mathcal{V}_k| e^{v_k{}^\top Q^{(t)} v_k}}{\sum_{j\in[N]} e^{E_j^{x\top} Q^{(t)} v_k}} \\ &= \frac{|\mathcal{V}_k| \exp(A_k^{(t)})}{\sum_{m\neq k} |\mathcal{V}_m| \exp(B_{k,m}^{(t)}) + |\mathcal{V}_k| \exp(A_k^{(t)})} \\ &= \frac{1}{\sum_{m\neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} \exp(B_{k,m}^{(t)} - A_k^{(t)}) + 1}.\end{aligned}$$

By Induction Hypothesis F.3, we have

- for $m \neq 1$, $e^{-\left(\log(K)+O(\frac{\log(K)}{K})\right)} \leq \exp(B_{k,m}^{(t)} - A_k^{(t)}) \leq e^{O(\frac{\log(K)}{K})-0.5\log(K)}$;
- $e^{-\left(1.51\log(K)+O(\frac{\log(K)}{K})\right)} \leq \exp(B_{k,1}^{(t)} - A_k^{(t)}) \leq e^{-1.01\log(K)}$.

Thus,

$$\mathbf{Attn}_k^{(t)} \geq \frac{1}{e^{O(\frac{\log(K)}{K})-0.5\log(K)}(\frac{N-|\mathcal{V}_1|}{|\mathcal{V}_k|}-1) + e^{-1.01\log(K)}\frac{|\mathcal{V}_1|}{|\mathcal{V}_k|} + 1} \geq \Omega\left(\frac{1}{K^{0.5}}\right),$$

where the second inequality follows from the fact that $P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}$.

On the other hand,

$$\mathbf{Attn}_k^{(t)} \leq \frac{1}{e^{-\left(\log(K)+O(\frac{\log(K)}{K})\right)}(\frac{N-|\mathcal{V}_1|}{|\mathcal{V}_k|}-1)+1} \leq \frac{1}{e^{-1}(\frac{1}{U_k^{\text{im}}}-\frac{1}{K})+1}.$$

Thus,

$$1 - \mathbf{Attn}_k^{(t)} \geq \frac{e^{-\left(\log(K)+O(\frac{\log(K)}{K})\right)}(\frac{N-|\mathcal{V}_1|}{|\mathcal{V}_k|}-1) + e^{-1.01\log(K)}\frac{|\mathcal{V}_1|}{|\mathcal{V}_k|}}{e^{-\left(\log(K)+O(\frac{\log(K)}{K})\right)}(\frac{N-|\mathcal{V}_1|}{|\mathcal{V}_k|}-1) + e^{-1.01\log(K)}\frac{|\mathcal{V}_1|}{|\mathcal{V}_k|} + 1} \geq \Omega(1).$$

We next analyze $\mathbf{Attn}_1^{(t)}$ as follows.

$$\begin{aligned}\mathbf{Attn}_1^{(t)} &= \frac{|\mathcal{V}_1| \exp(B_{k,1}^{(t)})}{\sum_{m\neq k} |\mathcal{V}_m| \exp(B_{k,m}^{(t)}) + |\mathcal{V}_k| \exp(A_k^{(t)})} \\ &= \frac{1}{\sum_{m\neq 1,k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_k|} \exp(B_{k,m}^{(t)} - B_{k,1}^{(t)}) + \frac{|\mathcal{V}_k|}{|\mathcal{V}_1|}\exp(A_k^{(t)} - B_{k,1}^{(t)}) + 1}.\end{aligned}$$

By Induction Hypothesis F.3,

- for $m \neq 1, k$, we have

$$e^{0.49\log(K)-O(\frac{\log(K)}{K})} \leq \exp(B_{k,m}^{(t)} - B_{k,1}^{(t)}) \leq e^{0.51\log(K)+O(\frac{\log(K)}{K})};$$

- for $m = 1$, $e^{1.01\log(K)} \leq \exp(A_k^{(t)} - B_{k,1}^{(t)}) \leq e^{1.51\log(K)+O(\frac{\log(K)}{K})}$.

Thus,

$$\mathbf{Attn}_1^{(t)} \le \frac{1}{e^{0.49\log(K)-O(\frac{\log(K)}{K})}(\frac{N-|\mathcal{V}_k|}{|\mathcal{V}_1|}-1)+e^{1.01\log(K)}\frac{|\mathcal{V}_k|}{|\mathcal{V}_1|}+1} \le O\left(\frac{1}{K^{0.49}}\right).$$
$$\mathbf{Attn}_1^{(t)} \ge \frac{1}{e^{0}(\frac{N-|\mathcal{V}_k|}{|\mathcal{V}_1|}-1)+e^{1.51\log(K)+O(\frac{\log(K)}{K})}\frac{|\mathcal{V}_k|}{|\mathcal{V}_1|}+1} \ge \Omega\left(\frac{1}{K^{0.51}}\right).$$

□

Lemma F.14. *Suppose Induction Hypothesis F.3 holds at iteration $T_{2,k} < t \le T_{3,k}$. If $x_{query} = v_k$ and $P_{input} \in \mathcal{E}^*_{imbal}$, for $n \ne 1, k$, then the following holds*

$$\mathbf{Attn}_n^{(t)} = O\left(\frac{1-\mathbf{Attn}_k^{(t)}-\mathbf{Attn}_1^{(t)}}{K}\right).$$

Proof. By Induction Hypothesis F.3, we have

$$e^{-O(\frac{\log(K)}{K})} \le \exp(B_{k,m}^{(t)} - B_{k,n}^{(t)}) \le e^{O(\frac{\log(K)}{K})}.$$

Combining with the fact that $\frac{|\mathcal{V}_m|}{|\mathcal{V}_n|} = \Theta(1)$ when $P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}$, we have

$$\frac{\mathbf{Attn}_n^{(t)}}{1-\mathbf{Attn}_k^{(t)}-\mathbf{Attn}_1^{(t)}} = \frac{|\mathcal{V}_n|\exp(B_{k,n}^{(t)})}{\sum_{m\ne 1,k}|\mathcal{V}_m|\exp(B_{k,m}^{(t)})} = \frac{1}{\sum_{m\ne 1,k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_n|}\exp(B_{k,m}^{(t)}-B_{k,n}^{(t)})} \le O\left(\frac{1}{K}\right).$$

□

F.4.2. Controlling Gradient Updates in Phase III

Lemma F.15. *At each iteration $T_{2,k} < t \le T_{3,k}$, if Induction Hypothesis F.3 holds, then $\alpha_k^{(t)} \ge 0$ and satisfies*

$$\alpha_k^{(t)} \ge \Omega\left(\frac{1}{K^{1.5}}\right).$$

Proof. By the gradient expression in Lemma D.3, we have

$$\begin{aligned}
\alpha_k^{(t)} &= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\,\mathbf{Attn}_k^{(t)}\cdot\left(\sum_{m\ne k}{\mathbf{Attn}_m^{(t)}}^2 + (1-\mathbf{Attn}_k^{(t)})^2\right)\right] \\
&= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap \mathcal{E}^*_{\text{imbal}}\}\,\mathbf{Attn}_k^{(t)}\cdot\left(\sum_{m\ne k}{\mathbf{Attn}_m^{(t)}}^2 + (1-\mathbf{Attn}_k^{(t)})^2\right)\right] \\
&\quad + \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap {\mathcal{E}^*_{\text{imbal}}}^c\}\,\mathbf{Attn}_k^{(t)}\cdot\left(\sum_{m\ne k}{\mathbf{Attn}_m^{(t)}}^2 + (1-\mathbf{Attn}_k^{(t)})^2\right)\right] \\
&\ge p_k\cdot\mathbb{P}(P_{\text{input}}\in\mathcal{E}^*_{\text{imbal}})\mathbb{E}\left[\mathbf{Attn}_k^{(t)}\cdot\left(\sum_{m\ne k}{\mathbf{Attn}_m^{(t)}}^2 + (1-\mathbf{Attn}_k^{(t)})^2\right)\middle|\{x_{\text{query}} = v_k\}\cap\mathcal{E}^*_{\text{imbal}}\right] \\
&\ge p_k\cdot\mathbb{P}(P_{\text{input}}\in\mathcal{E}^*_{\text{imbal}})\mathbb{E}\left[\mathbf{Attn}_k^{(t)}\cdot(1-\mathbf{Attn}_k^{(t)})^2 \mid \{x_{\text{query}} = v_k\}\cap\mathcal{E}^*_{\text{imbal}}\right] \\
&\ge \Omega\left(\frac{1}{K^{1.5}}\right)
\end{aligned}$$

where the last inequality follows from Lemma D.6, Lemma F.13 and our choice of p_k. □

Lemma F.16. *Given $k > 1$, if Induction Hypothesis F.3 holds at iteration $T_{k,2} \le t \le T_{k,3}$, then $\beta_{k,1}^{(t)} < 0$ satisfies*

$$|\beta_{k,1}^{(t)}| \le \left[\Omega\left(\frac{1}{K^{2.01}}\right), O\left(\frac{\alpha_k^{(t)}}{K^{0.48}}\right)\right].$$

Proof. Following the computations similar to those for Lemma F.4, we have

$$\begin{aligned}
&\sum_{m\neq k} \mathbf{Attn}_m^{(t)^2} - \mathbf{Attn}_1^{(t)} - \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)}) \\
&\quad \le -(1 - \mathbf{Attn}_k^{(t)} - \mathbf{Attn}_1^{(t)})(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)} - \max_{m\neq 1,k} \mathbf{Attn}_m^{(t)}).
\end{aligned}$$

Therefore,

$$\begin{aligned}
\beta_{k,1}^{(t)} &\le \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap \mathcal{E}^*_{\text{imbal}}\}\, \mathbf{Attn}_1^{(t)} \cdot \left(\sum_{m\neq k} \mathbf{Attn}_m^{(t)^2} - \mathbf{Attn}_1^{(t)} - \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right)\right] \\
&\quad + \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap \mathcal{E}^*_{\text{imbal}}{}^c\}\, \mathbf{Attn}_1^{(t)} \cdot \left(\sum_{m\neq k} \mathbf{Attn}_m^{(t)^2}\right)\right] \\
&\overset{(a)}{\le} p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}{}^c) + p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}) \\
&\qquad\qquad \cdot \mathbb{E}\left[-\Omega(\frac{(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)} - \max_{m\neq 1,k} \mathbf{Attn}_m^{(t)})}{K^{0.51}}) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
&\overset{(b)}{\le} p_k \cdot \left(-\Omega\left(\frac{1}{K^{1.01}}\right)\right) + 3p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\
&= -\Omega\left(\frac{1}{K^{2.01}}\right),
\end{aligned}$$

where both (a) and (b) follow from Lemma F.13, and the last inequality holds since

$$\frac{1}{K^{1.01}} \gg \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right).$$

Moreover, we have

$$\begin{aligned}
-\beta_{k,1}^{(t)} &\le \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_1 \cap P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\}\, \mathbf{Attn}_1^{(t)} \cdot (\mathbf{Attn}_1 + \mathbf{Attn}_k(1 - \mathbf{Attn}_k))\right] \\
&\quad + \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}{}^c\}\, \mathbf{Attn}_1^{(t)} \cdot (\mathbf{Attn}_1 + \mathbf{Attn}_k(1 - \mathbf{Attn}_k))\right] \\
&\le p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}) \cdot \mathbb{E}\left[\mathbf{Attn}_1 \cdot O(\mathbf{Attn}_1 + \mathbf{Attn}_k) \mid \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
&\quad + 2p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}{}^c) \\
&\le p_k \cdot \left(O\left(\frac{1}{K^{0.98}}\right)\right) + O\left(\frac{\alpha_k^{(t)}}{K^{0.49}}\right) + 6p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\
&\le O\left(\frac{\alpha_k^{(t)}}{K^{0.48}}\right)
\end{aligned}$$

where the last inequality follows from Lemma F.15. □

Lemma F.17. *If Induction Hypothesis F.3 holds at iteration $T_{2,k} < t \le T_{3,k}$, then for any $n \neq 1, k$, $\beta_{k,n}^{(t)}$ satisfies*

$$|\beta_{k,n}^{(t)}| \le O\left(\frac{\alpha_k^{(t)} - \beta_{k,1}^{(t)}}{K}\right).$$

Proof. By the gradient computation in Lemma D.3, we have

$$\beta_{k,n}^{(t)} \leq \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\,\mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m\neq k} \mathbf{Attn}_m^{(t)\,2}\right)\right], \tag{26}$$

$$-\beta_{k,n}^{(t)} \leq \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\,\mathbf{Attn}_n^{(t)} \cdot \left(\mathbf{Attn}_n^{(t)} + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right)\right]. \tag{27}$$

We further bound Equation (26) as

$$\begin{aligned}
\beta_{k,n}^{(t)} \leq& \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\}\,\mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m\neq k} \mathbf{Attn}_m^{(t)\,2}\right)\right] \\
&+ \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}^{*\,c}_{\text{imbal}}\}\,\mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m\neq k} \mathbf{Attn}_m^{(t)\,2}\right)\right] \\
\leq& p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}) \cdot \mathbb{E}\left[\mathbf{Attn}_n^{(t)} \cdot \left(\mathbf{Attn}_1^{(t)\,2} + O\left(\frac{1}{K}\right)\right) \middle| \{x_{\text{query}} = v_k\} \cap P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\right] \\
&+ p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^{*\,c}_{\text{imbal}}) \\
\leq& O\left(\frac{1}{K^3}\right) + O\left(\frac{|\beta_{k,1}^{(t)}|}{K}\right) + 6p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\
\leq& O\left(\frac{1}{K^3}\right) + O\left(\frac{|\beta_{k,1}^{(t)}|}{K}\right).
\end{aligned}$$

We then further bound Equation (27) as

$$\begin{aligned}
-\beta_{k,n}^{(t)} \leq& p_k \mathbb{E}\left[\mathbf{Attn}_n^{(t)} \cdot \left(\mathbf{Attn}_n^{(t)} + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right) \mid \{x_{\text{query}} = v_k\} \cap P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\right] \\
&+ p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^{*\,c}_{\text{imbal}}) \\
=& p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^{*\,c}_{\text{imbal}}) + p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}) \mathbb{E}\left[O\left(\frac{1 - \mathbf{Attn}_k^{(t)}}{K}\right)\right. \\
&\left. \cdot \left(O\left(\frac{1 - \mathbf{Attn}_k^{(t)}}{K}\right) + \mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
\leq& p_k \cdot \mathbb{P}(\mathcal{E}^*_{\text{imbal}}) \mathbb{E}\left[O\left(\frac{\mathbf{Attn}_k^{(t)}(1 - \mathbf{Attn}_k^{(t)})^2}{K}\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] + 6p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\
\leq& O\left(\frac{\alpha_k^{(t)}}{K}\right).
\end{aligned}$$

Following from the analysis in Lemma F.15, we have

$$\alpha_k^{(t)} \geq \Omega\left(\frac{1}{K^{1.5}}\right).$$

Thus, we obtain

$$|\beta_{k,n}^{(t)}| \leq O\left(\frac{\alpha_k^{(t)} - \beta_{k,1}^{(t)}}{K}\right).$$

□

F.4.3. End of Phase III

Lemma F.18. *Given $k > 1$, Induction Hypothesis F.3 holds for all $T_{2,k} < t \leq T_{3,k} = T_{2,k} + O(\frac{\log(K)K^{1.5}}{\eta})$, and at iteration $t = T_{3,k} + 1$, we have*

a. $A_k^{(T_{3,k}+1)} \geq \log(K)$;

b. $\mathbf{Attn}_k = \Omega(1)$ *if* $x_{query} = v_k$ *and* $P_{input} \in \mathcal{E}^*_{imbal}$.

Proof. The existence of $T_{3,k} = T_{2,k} + O(\frac{\log(K)K^{1.5}}{\eta})$ directly follows from Lemma F.15.

It is easy to verify Induction Hypothesis F.3 holds at $t = T_{2,k} + 1$. Now we suppose Induction Hypothesis F.3 holds for all iterations $\leq t-1$, and prove it holds at t.

By Lemma F.15, we have $\alpha_k^{(t-1)} \geq 0$. Thus $A_k^{(t)} = A_k^{(t-1)} + \eta\alpha_k^{(t-1)} \geq 0.5\log(K)$. Morover, by Lemma F.16, we have $|B_{k,1}^{(t)} - B_{k,1}^{(T_{2,k}+1)}| \leq O(\frac{A_k^{(t)} - A_k^{(T_{2,k}+1)}}{K^{0.48}})$ which immediately implies that

$$B_{k,1}^{(t)} \geq -O(\log(K)/K^{0.48}) - 0.51\log(K).$$

For $m \neq 1, k$, by Lemma F.17, we have

$$|B_{k,m}^{(t)}| \leq O(\frac{A_k^{(t)} - A_k^{(T_{2,k}+1)} + |B_{k,1}^{(t)} - B_{k,1}^{(T_{2,k}+1)}|}{K}) \leq O(\log(K)/K).$$

The proof for the second statement is deferred to the next phase (Lemma F.19). □

F.5. Phase IV: Convergence

At $t = T_{3,k} + 1$, the desired attention structure for the query token associated with feature v_k has already been achieved. In this final phase, we establish that these structures, including each under-represented feature, indeed represent the solutions toward which the algorithm converges.

Given any $0 < \epsilon < 1$, for $k \geq 2$, define

$$T_{4,k}^{\epsilon} \triangleq \max\left\{t > T_{3,k} : A_k^{(t)} \leq \log\left(\left(\frac{e(1-L_1^{\text{im}})K + U_1^{\text{im}}K^{0.51}}{L_k^{\text{im}}} - 1\right)\left(\left(\frac{3}{\epsilon}\right)^{\frac{1}{2}} - 1\right)\right)\right\}.$$

Induction Hypothesis F.4. For $T_{3,k} < t \leq T_{4,k}^{\epsilon}$, suppose $\text{polylog}(K) \gg \log(\frac{1}{\epsilon})$. Then the following holds.

a. $A_k^{(t)}$ is monotonically increasing and $A_k^{(t)} \in [\log(K), O(\log(K/\epsilon))]$;

b. $B_{k,1}^{(t)}$ is monotonically decreasing and

$$B_{k,1}^{(t)} \in \left[-0.51\log(K) - O\left(\frac{\log(K)}{K^{0.48}}\right), -0.49\log(K)\right]$$

c. $B_{k,n}^{(t)}$ is monotonically decreasing and $|B_{k,n}^{(t)}| = O(\frac{\log(K/\epsilon)}{K})$ for any $n \neq 1, k$.

F.5.1. Technical Lemmas

We first introduce several useful technical lemmas.

Lemma F.19. *Suppose Induction Hypothesis F.4 holds at iteration $T_{3,k} < t \leq T_{4,k}^{\epsilon}$. If $x_{query} = v_k$ and $P_{input} \in \mathcal{E}^*_{imbal}$, then the following holds.*

1. $\mathbf{Attn}_k^{(t)} = \Omega(1)$;
2. $(1 - \mathbf{Attn}_k^{(t)})^2 \geq \Omega(\epsilon) = \Omega(\exp(-\text{polylog}(K)))$.

Proof. Since $x_{\text{query}} = v_k$, we have

$$\begin{aligned}\mathbf{Attn}_k^{(t)} &= \frac{|\mathcal{V}_k|\exp(A_k^{(t)})}{\sum_{m\neq k}|\mathcal{V}_m|\exp(B_{k,m}^{(t)})+|\mathcal{V}_k|\exp(A_k^{(t)})}\\ &= \frac{1}{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_k|}\exp(B_{k,m}^{(t)}-A_k^{(t)})+1}.\end{aligned}$$

By Induction Hypothesis F.4, we have

- for $m\neq 1,k$:

$$\exp(B_{k,m}^{(t)}-A_k^{(t)}) \leq e^{O(\frac{\log(K/\epsilon)}{K})-\log(K)} \leq e^{O(\frac{\log(K)+\text{polylog}(K)}{K})-\log(K)} \leq O\left(\frac{1}{K}\right).$$

- for $m=1$, $\exp(B_{k,1}^{(t)}-A_k^{(t)}) \leq O(\frac{1}{K^{1.49}})$.

Therefore,

$$\mathbf{Attn}_k^{(t)} \geq \frac{1}{O\left(\frac{1}{K}\right)\left(\frac{N-|\mathcal{V}_1|}{|\mathcal{V}_k|}-1\right)+O(\frac{1}{K^{1.49}})\frac{|\mathcal{V}_1|}{|\mathcal{V}_k|}+1} \geq \Omega(1).$$

On the other hand, we have

$$\begin{aligned}1-\mathbf{Attn}_k^{(t)} &= \frac{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_k|}\exp(B_{k,m}^{(t)}-A_k^{(t)})}{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_k|}\exp(B_{k,m}^{(t)}-A_k^{(t)})+1}\\ &\overset{(a)}{\geq} \frac{\exp(\min_{m\neq 1,k}B_{k,m}^{(t)}-A_k^{(t)})(\frac{N-|\mathcal{V}_1|}{|\mathcal{V}_k|}-1)+\exp(B_{k,1}^{(t)}-A_k^{(t)})\frac{|\mathcal{V}_1|}{|\mathcal{V}_k|}}{\exp(\min_{m\neq 1,k}B_{k,m}^{(t)}-A_k^{(t)})(\frac{N-|\mathcal{V}_1|}{|\mathcal{V}_k|}-1)+\exp(B_{k,1}^{(t)}-A_k^{(t)})\frac{|\mathcal{V}_1|}{|\mathcal{V}_k|}+1}\\ &\geq \frac{\exp(\min_{m\neq k}B_{k,m}^{(t)}-A_k^{(t)})(\frac{(1-U_1^{\text{im}})K}{U_k^{\text{im}}}-1)+\exp(B_{k,1}^{(t)}-A_k^{(t)})\cdot\frac{L_1^{\text{im}}}{U_k^{\text{im}}}}{\exp(\min_{m\neq k}B_{k,m}^{(t)}-A_k^{(t)})(\frac{(1-U_1^{\text{im}})K}{U_k^{\text{im}}}-1)+\exp(B_{k,1}^{(t)}-A_k^{(t)})\cdot\frac{L_1^{\text{im}}}{U_k^{\text{im}}}+1}\\ &= \frac{(\frac{(1-U_1^{\text{im}})K}{U_k^{\text{im}}}-1+\frac{L_1^{\text{im}}K^{0.49}}{U_k^{\text{im}}})\exp(-A_k^{(t)})}{(\frac{(1-U_1^{\text{im}})K}{U_k^{\text{im}}}-1+\frac{L_1^{\text{im}}K^{0.49}}{U_k^{\text{im}}})\exp(-A_k^{(t)})+1}\\ &\geq \Omega(\epsilon^{\frac{1}{2}}),\end{aligned}$$

where (a) follows from the fact that $\frac{x}{1+x}$ increases w.r.t. $x>0$. □

Lemma F.20. *Suppose Induction Hypothesis F.4 holds at iteration $T_{3,k}<t\leq T_{4,k}^{\epsilon}$. If $x_{query}=v_k$ and $P_{input}\in\mathcal{E}_{imbal}^{*}$, then the following holds.*

1. $\mathbf{Attn}_n^{(t)} = \Theta\left(\frac{1-\mathbf{Attn}_k^{(t)}}{K}\right)$ *for* $n\neq 1,k$;
2. $\mathbf{Attn}_1^{(t)} \in \left[\Omega(\frac{1-\mathbf{Attn}_k^{(t)}}{K^{0.51}}), O\left(\frac{1-\mathbf{Attn}_k^{(t)}}{K^{0.49}}\right)\right]$.

Proof. We first have

$$\frac{\mathbf{Attn}_n^{(t)}}{1-\mathbf{Attn}_k^{(t)}} = \frac{|\mathcal{V}_n|\exp(B_{k,n}^{(t)})}{\sum_{m\neq k}|\mathcal{V}_m|\exp(B_{k,m}^{(t)})}.$$

If $n\neq 1$, by Induction Hypothesis F.4, we have

- for $m \neq 1, k$, $e^{-O(\frac{\log(K)-\log(\epsilon)}{K})} \leq \exp(B_{k,m}^{(t)} - B_{k,n}^{(t)}) \leq e^{O(\frac{\log(K)-\log(\epsilon)}{K})}$,
- for $m = 1$, $e^{-0.51\log(K)-O(\frac{\log(K/\epsilon)}{K})} \leq \exp(B_{k,1}^{(t)} - B_{k,n}^{(t)}) \leq 0$.

Note that when $P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}$, we have $\frac{|\mathcal{V}_m|}{|\mathcal{V}_n|} = \Theta(1)$, and $\frac{|\mathcal{V}_1|}{|\mathcal{V}_n|} = \Theta(K)$. Then combining with the fact that $-\log(\epsilon) \ll \text{polylog}(K)$, we obtain

$$\frac{\mathbf{Attn}_n^{(t)}}{1-\mathbf{Attn}_k^{(t)}} = \frac{|\mathcal{V}_n|\exp(B_{k,n}^{(t)})}{\sum_{m\neq k}|\mathcal{V}_m|\exp(B_{k,m}^{(t)})} = \frac{1}{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_n|}\exp(B_{k,m}^{(t)} - B_{k,n}^{(t)})} = \Theta\left(\frac{1}{K}\right).$$

For $n = 1$, by Induction Hypothesis F.4, we have

$$e^{0.49\log(K)-O(\frac{\log(K/\epsilon)}{K})} \leq \exp(B_{k,m}^{(t)} - B_{k,1}^{(t)}) \leq^{0.51\log(K)+O(\frac{\log(K/\epsilon)}{K})},$$

for $m \neq 1$. Combining with the fact that $\frac{|\mathcal{V}_m|}{|\mathcal{V}_1|} = \Theta\left(\frac{1}{K}\right)$ when $P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}$, and $-\log(\epsilon) \ll \text{polylog}(K)$, we have

$$\frac{\mathbf{Attn}_1^{(t)}}{1-\mathbf{Attn}_k^{(t)}} = \frac{1}{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_n|}\exp(B_{k,m}^{(t)} - B_{k,n}^{(t)})} \leq O(\frac{1}{K\cdot\frac{1}{K}\cdot e^{0.49\log(K)-O(\frac{\log(K/\epsilon)}{K})}+1}) = O\left(\frac{1}{K^{0.49}}\right);$$

and

$$\frac{\mathbf{Attn}_1^{(t)}}{1-\mathbf{Attn}_k^{(t)}} = \frac{1}{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_n|}\exp(B_{k,m}^{(t)} - B_{k,n}^{(t)})} \geq O(\frac{1}{K\cdot\frac{1}{K}\cdot e^{0.51\log(K)+O(\frac{\log(K/\epsilon)}{K})}+1}) \geq \Omega\left(\frac{1}{K^{0.51}}\right).$$

□

F.5.2. Controlling Gradient Updates in Phase IV

Lemma F.21. *At each iteration $T_{3,k} < t \leq T_{4,k}^{\epsilon}$, if Induction Hypothesis F.4 holds, then $\alpha_k^{(t)} \geq 0$ and satisfies*

$$\alpha_k^{(t)} \geq \Omega\left(\frac{\epsilon}{K}\right).$$

Proof. The analysis is similar to that for Lemma F.15, but we need to be more careful about the lower bound of $1-\mathbf{Attn}_k^{(t)}$. By the gradient expression, we have

$$\begin{aligned}
\alpha_k^{(t)} &= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k\}\,\mathbf{Attn}_k^{(t)}\cdot\left(\sum_{m\neq k}\mathbf{Attn}_m^{(t)\,2} + (1-\mathbf{Attn}_k^{(t)})^2\right)\right] \\
&\geq p_k\cdot\mathbb{P}(P\in\mathcal{E}^*)\mathbb{E}\left[\mathbf{Attn}_k^{(t)}\cdot\left(\sum_{m\neq k}\mathbf{Attn}_m^{(t)\,2} + (1-\mathbf{Attn}_k^{(t)})^2\right) \mid \{x_{\text{query}} = v_k\}\cap\mathcal{E}^*\right] \\
&\geq p_k\cdot\mathbb{P}(P\in\mathcal{E}^*)\mathbb{E}\left[\mathbf{Attn}_k^{(t)}\cdot(1-\mathbf{Attn}_k^{(t)})^2 \mid \{x_{\text{query}} = v_k\}\cap\mathcal{E}^*\right] \\
&\geq \Omega\left(\frac{\epsilon}{K}\right)
\end{aligned}$$

where the last inequality follows from Lemma F.19 and our choice of p_k. □

Lemma F.22. *At each iteration $T_{3,k} < t \leq T_{4,k}^{\epsilon}$, if Induction Hypothesis F.4 holds, then given $k \geq 2$, $\beta_{k,1}^{(t)}$ satisfies*

$$-O\left(\frac{\alpha_k^{(t)}}{K^{0.49}}\right) \leq \beta_{k,n}^{(t)} \leq 0.$$

Proof. Following the computations similar to those for Lemma F.4, we have

$$\sum_{m\neq k} \mathbf{Attn}_m^{(t)^2} - \mathbf{Attn}_1^{(t)} - \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)}) \leq -(1-\mathbf{Attn}_k^{(t)} - \mathbf{Attn}_1^{(t)})(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)} - \max_{m\neq 1,k} \mathbf{Attn}_m^{(t)}).$$

Therefore,

$$\begin{aligned}
\beta_{k,1}^{(t)} \leq\ & \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap \mathcal{E}^*_{\text{imbal}}\}\, \mathbf{Attn}_1^{(t)} \cdot \left(\sum_{m\neq k} \mathbf{Attn}_m^{(t)^2} - \mathbf{Attn}_1^{(t)} - \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)})\right)\right] \\
& + \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap {\mathcal{E}^*_{\text{imbal}}}^c\}\, \mathbf{Attn}_1^{(t)} \cdot \left(\sum_{m\neq k} \mathbf{Attn}_m^{(t)^2}\right)\right] \\
\overset{(a)}{\leq}\ & p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}) \cdot \mathbb{E}\left[-\Omega(\frac{(1-\mathbf{Attn}_k^{(t)})^2}{K^{0.51}}) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
& + p_k \cdot \mathbb{P}({\mathcal{E}^*_{\text{imbal}}}^c) \\
\leq & p_k \cdot \left(-\Omega\left(\frac{\epsilon}{K^{0.51}}\right)\right) + 3p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\
< & 0.
\end{aligned}$$

where (a) follows from Lemma F.20, and the last inequality holds since

$$\frac{\epsilon}{K^{0.51}} \geq \frac{\exp(-\operatorname{polylog}(K))}{K^{0.51}} \gg \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right).$$

Moreover,

$$\begin{aligned}
-\beta_{k,1}^{(t)} \leq\ & \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap \mathcal{E}^*_{\text{imbal}}\}\, \mathbf{Attn}_1^{(t)} \cdot \left(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)})\right)\right] \\
& + \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap {\mathcal{E}^*_{\text{imbal}}}^c\}\, \mathbf{Attn}_1^{(t)} \cdot \left(\mathbf{Attn}_1^{(t)} + \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)})\right)\right] \\
\leq & p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}) \cdot \mathbb{E}\left[\cdot O(\mathbf{Attn}_1^{(t)}(1-\mathbf{Attn}_k^{(t)})) \mid \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\
& + 2p_k \cdot \mathbb{P}({\mathcal{E}^*_{\text{imbal}}}^c) \\
\leq & O\left(\frac{\alpha_k^{(t)}}{K^{0.49}}\right) + 6p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\
= & O\left(\frac{\alpha_k^{(t)}}{K^{0.49}}\right).
\end{aligned}$$

□

Lemma F.23. *At each iteration $T_{3,k} < t \leq T_{4,k}^{\epsilon}$, if Induction Hypothesis F.4 holds, then given $k \geq 2$, for any $n \neq 1, k$, $\beta_{k,n}^{(t)}$ satisfies*

$$-O\left(\frac{\alpha_k^{(t)}}{K}\right) \leq \beta_{k,n}^{(t)} \leq 0.$$

Proof. Note that conditioned on the event $\{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}$, by Lemmas F.19 and F.20, we have $\mathbf{Attn}_k^{(t)} = \Omega(1)$, $\max_{m\neq k} \mathbf{Attn}_m = O\left(\frac{1}{K^{0.49}}\right)$. Thus, we obtain

$$\begin{aligned}\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 - \mathbf{Attn}_n^{(t)} - \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)}) &\leq \max_{m\neq k} \mathbf{Attn}_m^{(t)} \sum_{m\neq k} \mathbf{Attn}_m^{(t)} - \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)}) \\ &= -(1-\mathbf{Attn}_k^{(t)})(\mathbf{Attn}_k^{(t)} - \max_{m\neq k} \mathbf{Attn}_m^{(t)}) \\ &\leq -\Omega(1-\mathbf{Attn}_k^{(t)}). \end{aligned} \tag{28}$$

Therefore,

$$\begin{aligned}\beta_{k,n}^{(t)} \leq& \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap \mathcal{E}^*\}\, \mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 - \mathbf{Attn}_n^{(t)} - \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)})\right)\right] \\ &+ \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_k \cap {\mathcal{E}^*_{\text{imbal}}}^c\}\, \mathbf{Attn}_n^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2\right)\right] \\ \leq& p_k \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}) \cdot \mathbb{E}\left[-\Omega(\frac{(1-\mathbf{Attn}_k)^2}{K}) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*\right] + p_k \cdot \mathbb{P}(P_{\text{input}} \in {\mathcal{E}^*_{\text{imbal}}}^c) \\ \leq& p_k \cdot \left(-\Omega\left(\frac{\epsilon}{K}\right)\right) + 3p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\ \leq& 0,\end{aligned}$$

where the last inequality holds since

$$\frac{\epsilon}{K} \gg \frac{\exp(-\operatorname{polylog}(K))}{K} \gg \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right).$$

Moreover, we have

$$\begin{aligned}-\beta_{k,n}^{(t)} \leq& p_k \mathbb{E}\left[\mathbf{Attn}_n^{(t)} \cdot \left(\mathbf{Attn}_n^{(t)} + \mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)})\right) \mid \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] + 2p_k \mathbb{P}({\mathcal{E}^*_{\text{imbal}}}^c) \\ \leq& p_k \mathbb{E}\left[\Theta(\frac{1-\mathbf{Attn}_k^{(t)}}{K}) \cdot O\left(\mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)})\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] \\ &+ 6p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\ =& p_k \mathbb{E}\left[O\left(\frac{\mathbf{Attn}_k^{(t)}(1-\mathbf{Attn}_k^{(t)})^2}{K}\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*_{\text{imbal}}\right] + 6p_k \exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) \\ \leq& O\left(\frac{\alpha_k^{(t)}}{K}\right).\end{aligned}$$

□

F.5.3. End of Phase IV

Lemma F.24. *Given $k > 1$, and $0 < \epsilon < 1$, suppose $\operatorname{polylog}(K) \gg \log(\frac{1}{\epsilon})$. Then Induction Hypothesis F.4 holds for all $T_{3,k} < t \leq T_{4,k}^\epsilon = T_{3,k} + O(\frac{K\log(K\epsilon^{-\frac{1}{2}})}{\eta\epsilon})$, and at iteration $t = T_{4,k}^\epsilon + 1$, we have*

1. $\widetilde{\mathcal{L}}_k(\theta^{T_{4,k}^\epsilon+1}) < \frac{\epsilon}{2}$;
2. *If $x_{query} = v_k$ and $P_{input} \in \mathcal{E}^*_{imbal}$, we have $(1-\mathbf{Attn}_k^{(T_{4,k}^\epsilon+1)})^2 \leq O(\epsilon)$.*

Proof. The existence of $T_{4,k}^{\epsilon} = T_{3,k} + O(\frac{K\log(K\epsilon^{-\frac{1}{2}})}{\eta\epsilon})$ directly follows from Lemma F.21.

It is easy to verify Induction Hypothesis F.4 holds at $t = T_{3,k}+1$. Now we suppose Induction Hypothesis F.4 holds for all iterations $\leq t-1$, and prove it holds at t.

By Lemma F.21, we have $\alpha_k^{(t-1)} \geq 0$. Thus $A_k^{(t)} = A_k^{(t-1)} + \eta\alpha_k^{(t-1)} \geq \log(K)$. Moreover, by Lemma F.22, we have

$$|B_{k,1}^{(t)} - B_{k,1}^{(T_{3,k}+1)}| \leq O(\frac{A_k^{(t)} - A_k^{(T_{3,k}+1)}}{K^{0.49}}),$$

which immediately implies

$$B_{k,1}^{(t)} \geq -O(A_k^{(t)}/K^{0.49}) - O(\log(K)/K^{0.48}) - 0.51\log(K).$$

For $m \neq 1, k$, by Lemma F.23, we have

$$|B_{k,m}^{(t)} - B_{k,m}^{(T_{3,k}+1)}| \leq O(\frac{A_k^{(t)} - A_k^{(T_{3,k}+1)}}{K}) \leq O(\log(K/\epsilon)/K).$$

Thus

$$|B_{k,m}^{(t)}| \leq O(\log(K/\epsilon)/K) + O(\log(K)/K) = O(\log(K/\epsilon)/K).$$

At iteration $t = T_{4,k}^{\epsilon}+1$, we have

$$A_k^{(t)} \geq \log\left(\left(\frac{e(1-L_1^{\text{im}})K + U_1^{\text{im}}K^{0.51}}{L_k^{\text{im}}} - e\right)\left(\left(\frac{3}{\epsilon}\right)^{\frac{1}{2}} - 1\right)\right).$$

Thus when $\{x_{\text{query}} = v_k\} \cap \{P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\}$, we obtain

$$\begin{aligned}
1 - \mathbf{Attn}_k^{(t)} &= \frac{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_k|}\exp(B_{k,m}^{(t)} - A_k^{(t)})}{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_k|}\exp(B_{k,m}^{(t)} - A_k^{(t)}) + 1} \\
&\leq \frac{\exp(\max_{m\neq 1,k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{N-|\mathcal{V}_1|}{|\mathcal{V}_k|} - 1) + \exp(B_{k,1}^{(t)} - A_k^{(t)})\frac{|\mathcal{V}_1|}{|\mathcal{V}_k|}}{\exp(\max_{m\neq 1,k} B_{k,m}^{(t)} - A_k^{(t)})(\frac{N-|\mathcal{V}_1|}{|\mathcal{V}_k|} - 1) + \exp(B_{k,1}^{(t)} - A_k^{(t)})\frac{|\mathcal{V}_1|}{|\mathcal{V}_k|} + 1} \\
&\leq \frac{\exp(1 - A_k^{(t)})(\frac{(1-L_1^{\text{im}})K}{L_k^{\text{im}}} - 1) + \exp(-0.49\log(K) - A_k^{(t)})\frac{U_1^{\text{im}}K}{L_k^{\text{im}}}}{\exp(1 - A_k^{(t)})(\frac{(1-L_1^{\text{im}})K}{L_k^{\text{im}}} - 1) + \exp(-0.49\log(K) - A_k^{(t)})\frac{U_1^{\text{im}}K}{L_k^{\text{im}}} + 1} \\
&= \frac{\left((\frac{e(1-L_1^{\text{im}})K + U_1^{\text{im}}K^{0.51}}{L_k} - e)\right)\exp(-A_k^{(t)})}{\left((\frac{e(1-L_1^{\text{im}})K + U_1^{\text{im}}K^{0.51}}{L_k^{\text{im}}} - e)\right)\exp(-A_k^{(t)}) + 1} \\
&\leq \frac{((\frac{3}{\epsilon})^{\frac{1}{2}} - 1)^{-1}}{((\frac{3}{\epsilon})^{\frac{1}{2}} - 1)^{-1} + 1} \\
&= (\epsilon/3)^{\frac{1}{2}}.
\end{aligned}$$

We further derive

$$\begin{aligned}
\widetilde{\mathcal{L}}_k(\theta^{(t)}) &= \frac{1}{2}\mathbb{E}\left[\mathbf{1}\{P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\}\left(\sum_{m\neq k}\mathbf{Attn}_m^{(t)^2} + (1 - \mathbf{Attn}_k^{(t)})^2\right)\middle| x_{\text{query}} = v_k\right] \\
&\leq \frac{1}{2}\mathbb{P}\left(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\right)\cdot\mathbb{E}\left[\left(O\left(\frac{1}{K^{0.49}}\right) + 1\right)(1 - \mathbf{Attn}_k^{(t)})^2\middle| x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\right] \\
&\leq \frac{1}{2}\left(1 + O\left(\frac{1}{K^{0.49}}\right)\right)\cdot\frac{\epsilon}{3}
\end{aligned}$$

$$\leq \frac{\epsilon}{2}.$$

□

F.6. Proof of Theorem 3.3 for Under-represented Features

Theorem F.25 (Restatement of Theorem 3.3 for Under-represented Features)**.** *Suppose* $p_1 = \Theta(1)$ *and* $p_k = \Theta\left(\frac{1}{K}\right)$ *for* $2 \leq k \leq K$*. For any* $0 < \epsilon < 1$*, suppose* $N \geq \mathrm{poly}(K)$*, and* $\mathrm{polylog}(K) \gg \log(\frac{1}{\epsilon})$*. We apply GD to train the loss function given in Equation* (4)*. Then the following results hold.*

1. *The prediction error for* ***under-represented*** *feature converges: for* v_k *with* $2 \leq k \leq K$*, with at most* $T_k = O(\frac{\log(K)K^2}{\eta} + \frac{K\log\left(K\epsilon^{-\frac{1}{2}}\right)}{\epsilon\eta})$ *GD iterations,* $\mathcal{L}_k(\theta^{(T_k)}) \leq \mathcal{L}_k^* + \epsilon$*, where* $\mathcal{L}_k^* = \Theta(e^{-\mathrm{poly}(K)})$ *is the global minimum of Equation* (6)*;*
2. *Attention score concentrates: for each* $2 \leq k \leq K$*, if the query token is* v_k*, then after* T_k *iterations, with probability at least* $1 - e^{-\Omega(\mathrm{poly}(K))}$*, the one-layer transformer nearly "pays all attention" to input tokens featuring* v_k*:* $(1 - \mathbf{Attn}_k^{(T_k)})^2 \leq O(\epsilon)$.

Proof. The first statement is obtained by letting $T_k = T_{4,k}^{\epsilon} + 1$, and combining Lemma F.24, Lemma D.10 and Lemma D.11, which lead to

$$\mathcal{L}_k(\theta^{(T_k)}) - \mathcal{L}_k^* \leq \mathcal{L}_k(\theta^{(T_k)}) - \mathcal{L}_k^{\text{low}} \leq \widetilde{\mathcal{L}}_k(\theta^{(T_k)}) + 3\exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) < \epsilon.$$

The second statement directly follows from Lemma F.24. □

G. Analysis for Imbalanced Case: Dominant Feature (Proof of Theorem 3.3 Part II)

In this section, we delve into the analysis of prediction error when the query token features the dominant feature v_1. The training dynamics for the dominant feature v_1 are relatively straightforward, comprising only a single phase.

Note that at the beginning $t = 0$, we already have the following lemma.

Lemma G.1. *If $x_{query} = v_1$ and $P_{input} \in \mathcal{E}^*_{imbal}$, at $t = 0$, we have $\mathbf{Attn}_1^{(0)} = \Omega(1)$, $\mathbf{Attn}_k^{(0)} = O\left(\frac{1}{K}\right)$ for $k > 1$.*

Thus, the learning process directly enters the convergence phase, which is defined as follows. Given any $0 < \epsilon < 1$, define

$$T^{\epsilon}_{1,*} \triangleq \max\left\{t \geq 0 : A_1^{(t)} - \max_{m\neq 1} B_{1,m}^{(t)} \leq \log\left(\left(\frac{1}{L_1^{\text{im}}} - 1\right)\left(\left(\frac{2}{\epsilon}\right)^{\frac{1}{2}} - 1\right)\right)\right\}.$$

Induction Hypothesis G.1. For $0 \leq t \leq T^{\epsilon}_{1,*}$, suppose $\text{polylog}(K) \gg \log(\frac{1}{\epsilon})$. Then the following holds.

a. $A_1^{(t)}$ is monotonically increasing and $A_1^{(t)} \in [0, O(\log(1/\epsilon))]$;

b. $B_{k,n}^{(t)}$ is monotonically decreasing and $-O(\frac{A_1^{(t)}}{K}) \leq B_{1,n}^{(t)} \leq 0$ for any $n \neq 1$.

G.1. Technical Lemmas

We first introduce several useful technical lemmas.

Lemma G.2. *Suppose Induction Hypothesis G.1 holds at iteration $0 < t \leq T^{\epsilon}_{1,*}$. If $x_{query} = v_1$ and $\mathcal{E}^*_{imbal} \in P_{input}$, then the following holds*

1. $\mathbf{Attn}_1^{(t)} = \Omega(1)$;
2. $(1 - \mathbf{Attn}_1^{(t)})^2 \geq \Omega(\epsilon) = \Omega(\exp(-\text{polylog}(K)))$.

Proof. Since $x_{\text{query}} = v_1$, we have

$$\begin{aligned}\mathbf{Attn}_1^{(t)} &= \frac{|\mathcal{V}_1|\exp(A_k^{(t)})}{\sum_{m\neq 1}|\mathcal{V}_m|\exp(B_{1,m}^{(t)}) + |\mathcal{V}_1|\exp(A_k^{(t)})} \\ &= \frac{1}{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_k|}\exp(B_{k,m}^{(t)} - A_k^{(t)}) + 1}.\end{aligned}$$

By Induction Hypothesis G.1, we have

$$\begin{aligned}\mathbf{Attn}_1^{(t)} &\geq \frac{1}{\sum_{m\neq k}\frac{|\mathcal{V}_m|}{|\mathcal{V}_k|}\exp(B_{k,m}^{(0)} - A_k^{(0)}) + 1} \\ &\geq \frac{1}{(\frac{N}{L_1^{\text{im}}N} - 1) + 1} \geq \Omega(1).\end{aligned}$$

On the other hand, by the definition of $T^{\epsilon}_{1,*}$, we have

$$\begin{aligned}1 - \mathbf{Attn}_1^{(t)} &= \frac{\sum_{m\neq 1}\frac{|\mathcal{V}_m|}{|\mathcal{V}_1|}\exp(B_{1,m}^{(t)} - A_k^{(t)})}{\sum_{m\neq 1}\frac{|\mathcal{V}_m|}{|\mathcal{V}_1|}\exp(B_{1,m}^{(t)} - A_1^{(t)}) + 1} \\ &\overset{(a)}{\geq} \frac{\exp(\min_{m\neq 1} B_{1,m}^{(t)} - A_1^{(t)})(\frac{N}{|\mathcal{V}_1|} - 1)}{\exp(\min_{m\neq 1} B_{1,m}^{(t)} - A_1^{(t)})(\frac{N}{|\mathcal{V}_1|} - 1) + 1} \\ &\geq \frac{\exp(\min_{m\neq 1} B_{1,m}^{(t)} - A_1^{(t)})(\frac{1}{U_1^{\text{im}}} - 1)}{\exp(\min_{m\neq 1} B_{1,m}^{(t)} - A_1^{(t)})(\frac{1}{U_1^{\text{im}}} - 1) + 1}\end{aligned}$$

$$
\begin{aligned}
&= \frac{\exp(\max_{m\neq 1} B_{1,m}^{(t)} - A_1^{(t)} - \Delta B_1^{(t)})(\frac{1}{U_1^{\text{im}}} - 1)}{\exp(\max_{m\neq 1} B_{1,m}^{(t)} - A_1^{(t)} - \Delta B_1^{(t)})(\frac{1}{U_1^{\text{im}}} - 1) + 1} \\
&\geq \frac{(\frac{1}{p_1 L_1} - 1)^{-1}((\frac{2}{\epsilon})^{\frac{1}{2}} - 1)^{-1} \cdot e^{-O(\frac{\text{polylog}(K)}{K})}(\frac{1}{U_1^{\text{im}}} - 1)}{(\frac{1}{L_1^{\text{im}}} - 1)^{-1}((\frac{2}{\epsilon})^{\frac{1}{2}} - 1)^{-1}(\frac{1}{U_1^{\text{im}}} - 1)e^{-O(\frac{\text{polylog}(K)}{K})} + 1} \\
&\geq \Omega(\epsilon^{\frac{1}{2}}).
\end{aligned}
$$

where $\Delta B_k^{(t)} = \max_{m\neq k} B_{k,m}^{(t)} - \min_{m\neq k} B_{k,m}^{(t)} = O(\frac{A_k^{(t)}}{K})$, (a) follows from the fact that $\frac{x}{1+x}$ increases w.r.t. $x > 0$. □

Lemma G.3. *Suppose Induction Hypothesis G.1 holds at iteration $0 \leq t \leq T_{1,*}^{\epsilon}$. If $x_{\tau,query} = v_1$ and $P \in \mathcal{E}^*$, for $n \neq 1$, the following holds*

$$
\mathbf{Attn}_n^{(t)} = \Theta\left(\frac{(1 - \mathbf{Attn}_1^{(t)})}{K}\right).
$$

Proof. We first have

$$
\mathbf{Attn}_n^{(t)} = \frac{|\mathcal{V}_n| \exp(B_{1,n}^{(t)})}{\sum_{m\neq k} |\mathcal{V}_m| \exp(B_{1,m}^{(t)}) + |\mathcal{V}_1| \exp(A_1^{(t)})}.
$$

By Induction Hypothesis G.1, we have

$$
e^{-O(\frac{p \log(\frac{1}{\epsilon})}{K})} \leq \exp(B_{k,m}^{(t)} - B_{k,n}^{(t)}) \leq e^{O(\frac{p \log(\frac{1}{\epsilon})}{K})}.
$$

Combining with the fact that $-\log(\epsilon) \ll \text{polylog}(K)$, we obtain

$$
\frac{\mathbf{Attn}_n^{(t)}}{1 - \mathbf{Attn}_k^{(t)}} = \frac{|\mathcal{V}_n| \exp(B_{k,n}^{(t)})}{\sum_{m\neq k} |\mathcal{V}_m| \exp(B_{k,m}^{(t)})} = \frac{1}{\sum_{m\neq k} \frac{|\mathcal{V}_m|}{|\mathcal{V}_n|} \exp(B_{k,m}^{(t)} - B_{k,n}^{(t)})} = \Theta\left(\frac{1}{K}\right).
$$

□

G.2. Controlling Gradient Updates

Lemma G.4. *At each iteration $0 \leq t \leq T_{1,*}^{\epsilon}$, if Induction Hypothesis G.1 holds then $\alpha_1^{(t)} \geq 0$ and satisfies*

$$
\alpha_k^{(t)} \geq \Omega(\epsilon).
$$

Proof. By the gradient expression, we have

$$
\begin{aligned}
\alpha_1^{(t)} &= \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_1\} \mathbf{Attn}_1^{(t)} \cdot \left(\sum_{m\neq 1} {\mathbf{Attn}_m^{(t)}}^2 + (1 - \mathbf{Attn}_1^{(t)})^2\right)\right] \\
&\geq p_1 \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}_{\text{imbal}}^*) \mathbb{E}\left[\mathbf{Attn}_k^{(t)} \cdot \left(\sum_{m\neq k} {\mathbf{Attn}_m^{(t)}}^2 + (1 - \mathbf{Attn}_k^{(t)})^2\right) \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}_{\text{imbal}}^*\right] \\
&\geq p_1 \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}_{\text{imbal}}^*) \mathbb{E}\left[\mathbf{Attn}_k^{(t)} \cdot (1 - \mathbf{Attn}_k^{(t)})^2 \middle| \{x_{\text{query}} = v_k\} \cap \mathcal{E}^*\right] \\
&\geq \Omega(\epsilon)
\end{aligned}
$$

where the last inequality follows from Lemma G.2 and our choice of p_1. □

Lemma G.5. *At each iteration $0 \le t \le T^{\epsilon}_{1,*}$, if Induction Hypothesis G.1 holds, then for any $n \neq 1$, $\beta^{(t)}_{1,n}$ satisfies*

$$-O\left(\frac{\alpha^{(t)}_1}{K}\right) \le \beta^{(t)}_{1,n} \le 0.$$

Proof. Note that conditioned on the event $\{x_{\text{query}} = v_1\} \cap \mathcal{E}^*_{\text{imbal}}$, by Lemmas G.2 and G.3, we have $\mathbf{Attn}^{(t)}_1 = \Omega(1)$, and $\max_{m\neq 1} \mathbf{Attn}^{(t)}_m = O\left(\frac{1}{K}\right)$. Thus, we further obtain

$$\begin{aligned}
&\sum_{m\neq 1} {\mathbf{Attn}^{(t)}_m}^2 - \mathbf{Attn}^{(t)}_n - \mathbf{Attn}^{(t)}_1(1-\mathbf{Attn}^{(t)}_1) \\
&\quad\le \max_{m\neq 1} \mathbf{Attn}^{(t)}_m \sum_{m\neq 1} \mathbf{Attn}^{(t)}_m - \mathbf{Attn}^{(t)}_1(1-\mathbf{Attn}^{(t)}_1) \\
&\quad= -(1-\mathbf{Attn}^{(t)}_1)(\mathbf{Attn}^{(t)}_1 - \max_{m\neq 1}\mathbf{Attn}^{(t)}_m) \\
&\quad\le -\Omega(1-\mathbf{Attn}^{(t)}_1). \qquad (29)
\end{aligned}$$

Therefore,

$$\begin{aligned}
\beta^{(t)}_{1,n} \le\ & \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_1 \cap \mathcal{E}^*_{\text{imbal}}\}\, \mathbf{Attn}^{(t)}_n \cdot \left(\sum_{m\neq 1} {\mathbf{Attn}^{(t)}_m}^2 - \mathbf{Attn}^{(t)}_n - \mathbf{Attn}^{(t)}_1(1-\mathbf{Attn}^{(t)}_1)\right)\right] \\
&+ \mathbb{E}\left[\mathbf{1}\{x_{\text{query}} = v_1 \cap {\mathcal{E}^*_{\text{imbal}}}^c\}\, \mathbf{Attn}^{(t)}_n \cdot \left(\sum_{m\neq 1} {\mathbf{Attn}^{(t)}_m}^2\right)\right] \\
\overset{(a)}{\le}\ & p_1 \cdot \mathbb{P}(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}) \cdot \mathbb{E}\left[-\Omega\left(\frac{(1-\mathbf{Attn}^{(t)}_1)^2}{K}\right) \middle| \{x_{\text{query}} = v_1\} \cap \mathcal{E}^*_{\text{imbal}}\right] + p_1 \cdot \mathbb{P}({\mathcal{E}^*_{\text{imbal}}}^c) \\
\overset{(b)}{\le}\ & p_1 \cdot \left(-\Omega\left(\frac{\epsilon}{K}\right)\right) + 3p_1 \exp\left(-\frac{c^2_{\text{im}} N}{25K^2}\right) \\
\le\ & 0
\end{aligned}$$

where (a) follows from Equation (29) and Lemma G.3, (b) follows from Lemma G.2, and the last inequality holds since

$$\frac{\epsilon}{K} \gg \frac{\exp(-\operatorname{polylog}(K))}{K} \gg \exp\left(-\frac{c^2_{\text{im}} p_2 N}{25K^2}\right).$$

Moreover, we have

$$\begin{aligned}
-\beta^{(t)}_{1,n} &\le p_1 \mathbb{E}\left[\mathbf{Attn}^{(t)}_n \cdot \left(\mathbf{Attn}^{(t)}_n + \mathbf{Attn}^{(t)}_1(1-\mathbf{Attn}^{(t)}_1)\right) \mid \{x_{\text{query}} = v_1\} \cap \mathcal{E}^*_{\text{imbal}}\right] + 2p_1 \mathbb{P}({\mathcal{E}^*_{\text{imbal}}}^c) \\
&\le p_1 \mathbb{E}\left[\Theta\left(\frac{1-\mathbf{Attn}^{(t)}_1}{K}\right) \cdot O\left(\mathbf{Attn}^{(t)}_k(1-\mathbf{Attn}^{(t)}_1)\right) \middle| \{x_{\text{query}} = v_1\} \cap \mathcal{E}^*_{\text{imbal}}\right] + 6p_1 \exp\left(-\frac{c^2_{\text{im}} N}{25K^2}\right) \\
&= p_1 \mathbb{E}\left[O\left(\frac{\mathbf{Attn}^{(t)}_1(1-\mathbf{Attn}^{(t)}_1)^2}{K}\right) \middle| \{x_{\text{query}} = v_1\} \cap \mathcal{E}^*_{\text{imbal}}\right] + 6p_1 \exp\left(-\frac{c^2_{\text{im}} p^2 N}{25K^2}\right) \\
&\le O\left(\frac{\alpha^{(t)}_1}{K}\right).
\end{aligned}$$

□

G.3. End of the Phase

Lemma G.6. *Given $0 < \epsilon < \frac{1}{2}$, suppose $\operatorname{polylog}(K) \gg \log(\frac{1}{\epsilon})$. Then Induction Hypothesis G.1 holds for all $0 \le t \le T^{\epsilon}_{1,*} = O(\frac{\log(\epsilon^{-\frac{1}{2}})}{\eta\epsilon})$, and at iteration $t = T^{\epsilon}_{1,*} + 1$, we have*

1. $\widetilde{\mathcal{L}}_1(\theta^{T^\epsilon_{1,*}+1}) < \epsilon/2$;
2. *If* $x_{query} = 1$ *and* $P_{input} \in \mathcal{E}^*_{imbal}$, *we have* $(1 - \mathbf{Attn}_1^{(T^\epsilon_{1,*}+1)})^2 \le O(\epsilon)$.

Proof. We first prove the existence of $T^\epsilon_{1,*}$. Recall that

$$T^\epsilon_{1,*} = \max\left\{t \ge 0 : A_1^{(t)} - \max_{m\neq 1} B_{1,m}^{(t)} \le \log\left(\left(\frac{1}{L_1^{\text{im}}} - 1\right)\left(\left(\frac{2}{\epsilon}\right)^{\frac{1}{2}} - 1\right)\right)\right\}.$$

When $t \in [0, T^\epsilon_{1,*}]$, we can simply lower bound the update of $A_k^{(t)} - \max_{m\neq k} B_{k,m}^{(t)}$ as

$$A_k^{(t+1)} - \max_{m\neq k} B_{k,m}^{(t+1)} \ge A_k^{(t+1)} \ge A_k^{(t)} + \Omega\left(\frac{\eta\epsilon}{K}\right).$$

Therefore, at most $T^\epsilon_{1,*} = O(\frac{\log\left((\frac{1}{L_1^{\text{im}}}-1)((\frac{2}{\epsilon})^{\frac{1}{2}}-1)\right)}{\eta\epsilon}) = O(\frac{\log(\epsilon^{-\frac{1}{2}})}{\eta\epsilon})$ iterations are needed before $A_k^{(t)} - \max_{m\neq k} B_{k,m}^{(t)}$ exceeds $\log\left((\frac{1}{L_1^{\text{im}}} - 1)((\frac{2}{\epsilon})^{\frac{1}{2}} - 1)\right)$.

It is easy to verify Induction Hypothesis G.1 holds at $t = 0$. Now we suppose Induction Hypothesis G.1 holds for all iterations $0 \le t - 1$, and prove it holds at t.

By Lemma G.4, we have $\alpha_1^{(t-1)} \ge 0$. Thus $A_1^{(t)} \ge A_1^{(t-1)} \ge 0$. By Lemma G.5, we have $-O\left(\frac{\alpha_1^{(t-1)}}{K}\right) \le \beta_{1,n}^{(t-1)} \le 0$. Thus,

$$\begin{aligned}
-B_{1,n}^{(t)} &\le -B_{1,n}^{(t-1)} + \eta O\left(\frac{\alpha_1^{(t-1)}}{K}\right) \\
&\le O\left(\frac{A_1^{(t-1)}}{K}\right) + \eta O\left(\frac{\alpha_1^{(t-1)}}{K}\right) \\
&\le O\left(\frac{A_1^{(t)}}{K}\right).
\end{aligned}$$

Moreover, by the definition of $T^\epsilon_{1,*}$, for any $t \le T^\epsilon_{1,*}$, we immediately have

$$A_1^{(t)} \le A_1^{(t)} - \max_{m\neq 1} B_{1,m}^{(t)} \le \log\left(\left(\frac{1}{L_1^{\text{im}}} - 1\right)\left(\left(\frac{2}{\epsilon}\right)^{\frac{1}{2}} - 1\right)\right).$$

Therefore, $A_1^{(t)} \le O(\log(\frac{1}{\epsilon}))$.

At iteration $t = T^\epsilon_{1,*} + 1$, we have $A_1^{(t)} - \max_{m\neq 1} B_{1,m}^{(t)} > \log\left((\frac{1}{L_1^{\text{im}}} - 1)((\frac{2}{\epsilon})^{\frac{1}{2}} - 1)\right)$. Thus, when $\{x_{\text{query}} = v_1\} \cap \{P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\}$, we obtain

$$\begin{aligned}
1 - \mathbf{Attn}_1^{(t)} &= \frac{\sum_{m\neq 1} \frac{|\mathcal{V}_m|}{|\mathcal{V}_1|} \exp(B_{1,m}^{(t)} - A_1^{(t)})}{\sum_{m\neq 1} \frac{|\mathcal{V}_m|}{|\mathcal{V}_1|} \exp(B_{1,m}^{(t)} - A_1^{(t)}) + 1} \\
&\le \frac{\exp(\max_{m\neq 1} B_{1,m}^{(t)} - A_1^{(t)})(\frac{N}{|\mathcal{V}_1|} - 1)}{\exp(\max_{m\neq 1} B_{1,m}^{(t)} - A_1^{(t)})(\frac{N}{|\mathcal{V}_1|} - 1) + 1} \\
&\le \frac{\exp(\max_{m\neq 1} B_{1,m}^{(t)} - A_1^{(t)})(\frac{1}{L_1^{\text{im}}} - 1)}{\exp(\max_{m\neq 1} B_{1,m}^{(t)} - A_1^{(t)})(\frac{1}{L_1^{\text{im}}} - 1) + 1}
\end{aligned}$$

$$\leq \frac{\left((\frac{1}{L_1^{\text{im}}}-1)((\frac{2}{\epsilon})^{\frac{1}{2}}-1)\right)^{-1}(\frac{1}{L_1^{\text{im}}}-1)}{\left((\frac{1}{L_1^{\text{im}}}-1)((\frac{2}{\epsilon})^{\frac{1}{2}}-1)\right)^{-1}(\frac{1}{L_1^{\text{im}}}-1)+1}$$
$$= (\epsilon/2)^{\frac{1}{2}}.$$

Similarly, we have

$$\begin{aligned}
\widetilde{\mathcal{L}}_1(\theta^{(t)}) &= \frac{1}{2}\mathbb{E}\left[\mathbf{1}\{P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\}\left(\sum_{m\neq k}\mathbf{Attn}_m^{(t)^2} + (1-\mathbf{Attn}_k^{(t)})^2\right)\middle| x_{\text{query}} = v_k\right] \\
&\leq \frac{1}{2}\mathbb{P}\left(P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\right)\cdot\mathbb{E}\left[\left(O\left(\frac{1}{K}\right)+1\right)(1-\mathbf{Attn}_k^{(t)})^2\middle| x_{\text{query}} = v_k \cap P_{\text{input}} \in \mathcal{E}^*_{\text{imbal}}\right] \\
&\leq \frac{1}{2}\cdot\left(1+O\left(\frac{1}{K}\right)\right)\cdot\frac{\epsilon}{2} \\
&\leq \epsilon/2.
\end{aligned}$$

□

G.4. Proof of Theorem 3.3 for Dominant Feature

Theorem G.7 (Restatement of Theorem 3.3 for Dominant Feature)**.** *Suppose $p_1 = \Theta(1)$ and $p_k = \Theta\left(\frac{1}{K}\right)$ for $2 \leq k \leq K$. For any $0 < \epsilon < 1$, suppose $N \geq \text{poly}(K)$, and $\text{polylog}(K) \gg \log(\frac{1}{\epsilon})$. We apply GD to train the loss function given in Equation (4). Then the following results hold.*

1. *The prediction error for* **dominant** *feature converges: for v_1, with at most $T_1 = O(\frac{\log(\epsilon^{-\frac{1}{2}})}{\eta\epsilon})$ GD iterations, $\mathcal{L}_1(\theta^{(T_1)}) \leq \mathcal{L}_1^* + \epsilon$, where $\mathcal{L}_1^* = \Theta(e^{-\text{poly}(K)})$ is the global minimum of Equation (6);*
2. *Attention score concentrates: $k = 1$, if the query token is v_k, then after T_k iterations, with probability at least $1-e^{-\Omega(\text{poly}(K))}$, the one-layer transformer nearly "pays all attention" to input tokens featuring v_k: $(1-\mathbf{Attn}_k^{(T_k)})^2 \leq O(\epsilon)$.*

Proof. The first statement is obtained by letting $T_1 = T^{\epsilon}_{1,*} + 1$, and combining Lemma G.6, Lemma D.10 and Lemma D.11, which lead to

$$\mathcal{L}_1(\theta^{(T_1)}) - \mathcal{L}_1^* \leq \mathcal{L}_1(\theta^{(T_1)}) - \mathcal{L}_1^{\text{low}} \leq \widetilde{\mathcal{L}}_1(\theta^{(T_1)}) + 3\exp\left(-\frac{c_{\text{im}}^2 N}{25K^2}\right) < \epsilon.$$

The second statement directly follows from Lemma G.6. □