ATTENTION SINKS IN TRANSFORMER-BASED LANGUAGE MODELS: EMERGENCE MECHANISMS AND BENIGN EFFECTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention sinks, in which large language models (LLMs) disproportionately attend to specific tokens such as the initial token or weak semantic tokens like punctuation and function words, have become an important focus in understanding Transformer mechanisms. This phenomenon has been shown to affect model performance and efficiency, motivating methods that either mitigate or leverage it to improve architectures. However, its theoretical underpinnings remain insufficiently understood, particularly how these patterns emerge spontaneously during training. In this work, we address this gap by theoretically analyzing a two-block Transformer trained on synthetic text datasets. We show that causal masking causes parameters associated with the first token to accumulate disproportionate gradients, driving the formation of a first-token attention sink, while high-frequency semantic tokens receive significantly larger gradients than other tokens and consequently absorb most of the attention. We further show that these attention sinks are compatible with the Transformer's predictive capacity, even when attention is heavily concentrated on a few tokens. Overall, this work provides a training dynamics explanation for the emergence of attention sinks, offering insights that inform the design of new algorithms and architectures.

1 INTRODUCTION

Large Language Models (LLMs) based on the Transformer architecture have achieved remarkable success across a wide range of natural language processing tasks (Vaswani et al., 2017; Brown et al., 2020; OpenAI, 2023; Guo et al., 2025). Despite their empirical dominance, the internal working mechanisms of the self-attention layers remain a subject of intense theoretical investigation. A particularly intriguing phenomenon observed in trained Transformers is the emergence of *Attention Sinks*, where the model allocates a disproportionately large amount of attention weight to specific tokens Xiao et al. (2024). Empirical studies have consistently observed that attention heads tend to concentrate on distinctive tokens regardless of the varying context (Yu et al., 2024; Sun et al., 2024a; Cancedda, 2024; An et al., 2025). Specifically, substantial attention is frequently directed towards the **initial tokens** of a sequence, or explicitly focused on **weak semantic tokens** such as punctuation (e.g., "`,`", "`.`", "`\n`") and function words (e.g., "and", "of").

The persistent existence of attention sinks has been actively exploited to facilitate high-performance and efficient architectures. For instance, StreamingLLM (Xiao et al., 2024) identifies the critical role of initial tokens as "attention sinks": by explicitly retaining their KV states, it acts as a stability anchor to prevent the performance collapse typical of simple window attention, thereby enabling efficient infinite-context generation. Building upon this, SepLLM (Chen et al., 2025) further observes that separator tokens (e.g., punctuation) often accumulate high attention scores and utilizes them to compress sequence information. Beyond these specific architectural designs, the characterization of attention sinks has also inspired advanced model compression techniques like quantization and pruning methods (Wei et al., 2022; Nrusimha et al., 2024; Lin et al., 2024; Liu et al., 2024; Sun et al., 2024b; Yin et al., 2024). Furthermore, the paradigm of preserving high-attention anchors continues to guide various efficient inference techniques (Zhang et al., 2024; Huang et al., 2025) and KV cache optimization algorithms (Ge et al., 2024; Wan et al., 2024; Wu & Tu, 2024).

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

The theoretical mechanisms underlying the emergence of attention sinks have also attracted growing interest. For instance, researchers have investigated the sink phenomenon from diverse perspectives, including information propagation Barbero et al. (2025), positional biases Wu et al. (2025), outlier feature utilization Zhang et al. (2025), and geometric representation Ruscio et al. (2025). However, these studies primarily examine the static properties of converged models, leaving a critical gap in understanding the *training dynamics* of this phenomenon. This leads to a fundamental question regarding the temporal evolution of these sinks:

What underlying optimization incentives drive the model to progressively anchor its attention on these specific tokens during the training process?

In this work, we answer this question by investigating the training dynamics of a two-block Transformer on synthetic datasets. First, we theoretically reveal that the causal masking structure inherently biases gradient descent towards the initial position, creating a structural necessity for the formation of the *First Token Sink*. Regarding *weak semantic tokens*, we further demonstrate their high frequency in the data distribution induces larger cumulative gradients, driving the optimization process to anchor on them. Crucially, we reveal that utilizing these tokens as sinks is structurally compatible with the model's predictive capabilities.

Our main contributions are summarized as follows:

- We provide a rigorous analysis of the *First Token Sink* phenomenon. We prove that the causal masking structure inherently biases the gradient descent, specifically affecting the dynamics of the value matrix ($\boldsymbol{W}_V$). We demonstrate that this dynamic on $\boldsymbol{W}_V$ subsequently drives the query and key matrices ($\boldsymbol{W}_{QK}$) to progressively anchor attention onto the initial token, providing a mechanistic guarantee for its emergence.
- We formally define and analyze the *Weak Semantic Token Sink*. We theoretically reveal that the model will attend to these tokens during training, driven by their high frequency in the data distribution. Furthermore, we show that due to causal masking, these tokens act as *context carriers*, aggregating semantic information to support correct classification. This implies that anchoring attention on weak semantic tokens is not in conflict with achieving high prediction accuracy.
- Technically, unlike prior analyses on the training dynamics of single-block Transformers which fails to fully capture the structural effects of causal masking, we study a two-block setting which uniquely allows us to capture the structural bias inherent in causal architectures, thereby providing meaningful insights for future theoretical research on deep models.

2 Related Work

Attention Sink. The concept of "attention sinks" was first systematically defined by Xiao et al. (2024), who observed that pre-trained Transformers allocate massive attention weights to the initial token. Subsequent studies extended this finding, revealing that such sinks also emerge on some weak semantic tokens, such as separators and punctuation (Sun et al., 2024a; Yu et al., 2024; An et al., 2025). This phenomenon has significantly influenced the landscape of efficient LLM deployment. For instance, StreamingLLM (Xiao et al., 2024) and SepLLM (Chen et al., 2025) leverage initial tokens and separators as stability anchors to enable efficient infinite-context generation and context compression. Beyond these specific designs, the distributional properties of sinks have also guided specialized model compression methods (Wei et al., 2022; Nrusimha et al., 2024; Lin et al., 2024; Liu et al., 2024; Sun et al., 2024b; Yin et al., 2024) and KV cache optimization methods (Ge et al., 2024; Wan et al., 2024; Wu & Tu, 2024) that explicitly accommodate these critical outliers to preserve precision. Regarding their mechanisms, recent static analyses offer diverse interpretations: Wu et al. (2025) attribute sinks to inherent positional biases; Barbero et al. (2025) argue they prevent feature over-mixing; Zhang et al. (2025) propose a Catch-Tag-Release" information retrieval mechanism; and Ruscio et al. (2025) characterize them as geometric reference frames. However, these studies primarily examine converged models, leaving the training dynamics governing their spontaneous emergence largely unexplored.

Training Dynamics of Transformers. Toward understanding the training dynamics of Transformers, Zhang et al. (2020) theoretically justified the convergence benefits of adaptive gradient methods.

Focusing on structural learning, Jelassi et al. (2022) analyzed simplified ViTs to demonstrate that they implicitly learn spatial locality from data, while Li et al. (2023) provided a mechanistic understanding of how BERT-like architectures Devlin et al. (2019) acquire semantic structure. Furthermore, Tian et al. (2023) characterized the SGD dynamics of single-layer Transformers, revealing how attention heads naturally evolve to focus on informative tokens. Extending to complex regimes, recent studies investigated ICL dynamics (Siyu et al., 2024; Li et al., 2024) and multilayer hierarchies (Tian et al., 2024). Unlike existing research, this work focuses on the training dynamics of attention sink.

3 Problem Setup

This section introduce the two-block Transformer model and the training algorithm we use in this paper.

Two-block Transformer. Consider an input sequence $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ with the next token to be predicted denoted as $\boldsymbol{x}_{T+1}$, where $\boldsymbol{x}_i \in \mathbb{R}^d$. Let the input embedding matrix be

$$\boldsymbol{H}^{(0)} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]^\top \in \mathbb{R}^{T\times d}. \tag{1}$$

First Block. The output of the first self-attention block, denoted as $\boldsymbol{O}^{(1)}$, is computed as:

$$\boldsymbol{O}^{(1)} = \mathcal{S}\left(\boldsymbol{H}^{(0)}\boldsymbol{W}_{QK}^{(1)}\boldsymbol{H}^{(0)\top} + \boldsymbol{M}\right)\boldsymbol{H}^{(0)}\boldsymbol{W}_{V}^{(1)} \in \mathbb{R}^{T\times d}, \tag{2}$$

where $\boldsymbol{W}_{QK}^{(1)}, \boldsymbol{W}_{V}^{(1)} \in \mathbb{R}^{d\times d}$. $\mathcal{S}(\cdot)$ denotes the softmax function. $\boldsymbol{M} \in \mathbb{R}^{T\times T}$ is the causal mask matrix where $M_{ij} = 0$ if $i \geq j$ and $-\infty$ otherwise (setting strictly upper triangular entries to 0 after softmax).

Second Block. We concatenate the input embeddings and the first block's output to form the input for the second block:

$$\boldsymbol{H}^{(1)} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T]^\top = \text{concat}\left([\boldsymbol{H}^{(0)}, \boldsymbol{O}^{(1)}]\right) \in \mathbb{R}^{T\times 2d}. \tag{3}$$

The output of the second block is similarly defined:

$$\boldsymbol{O}^{(2)} = \mathcal{S}\left(\boldsymbol{H}^{(1)}\boldsymbol{W}_{QK}^{(2)}\boldsymbol{H}^{(1)\top} + \boldsymbol{M}\right)\boldsymbol{H}^{(1)}\boldsymbol{W}_{V}^{(2)} \in \mathbb{R}^{T\times d}, \tag{4}$$

where the weights are adapted to the concatenated dimension: $\boldsymbol{W}_{QK}^{(2)} \in \mathbb{R}^{2d\times 2d}$ and $\boldsymbol{W}_{V}^{(2)} \in \mathbb{R}^{2d\times d}$.

Prediction. To predict $\boldsymbol{x}_{T+1}$, we utilize the representation of the last token $\boldsymbol{x}_T$ from the second block. The predicted embedding corresponds to the last row of $\boldsymbol{O}^{(2)}$:

$$\hat{\boldsymbol{x}}_{T+1}^\top = \mathcal{S}\left(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}\boldsymbol{H}^{(1)\top}\right)\boldsymbol{H}^{(1)}\boldsymbol{W}_{V}^{(2)}, \tag{5}$$

where $\boldsymbol{h}_T \in \mathbb{R}^{2d}$ is the last row of $\boldsymbol{H}^{(1)}$ (the query vector).

Training algorithm. We consider the Transformer is trained by cross-entropy loss with softmax. We denote $\ell(\hat{\boldsymbol{x}}_{T+1}, \boldsymbol{x}_{T+1}) = -\log\left(\mathcal{S}(\hat{\boldsymbol{x}}_{T+1}^\top)\boldsymbol{x}_{T+1}\right)$, and the loss is given by

$$\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \mathbb{E}_{\mathcal{D}}\left[\ell(\hat{\boldsymbol{x}}_{T+1}, \boldsymbol{x}_{T+1})\right], \tag{6}$$

where $\boldsymbol{\theta}$ denotes the parameter of the Transformer, $\mathcal{D}$ denotes the dataset. We further use gradient descent to train the second block of the Transformer:

$$\boldsymbol{W}_{QK}^{(2)}(t+1) = \boldsymbol{W}_{QK}^{(2)}(t) - \eta\nabla_{\boldsymbol{W}_{QK}^{(2)}(t)}\mathcal{L}_D\big(\boldsymbol{\theta}(t)\big), \tag{7}$$

$$\boldsymbol{W}_{V}^{(2)}(t+1) = \boldsymbol{W}_{V}^{(2)}(t) - \eta\nabla_{\boldsymbol{W}_{V}^{(2)}(t)}\mathcal{L}_D\big(\boldsymbol{\theta}(t)\big). \tag{8}$$

4 Training Dynamics Analysis of Attention Sinks

In this section, we present the training dynamics results for the *first token sink* and the *weak semantic token sink*, respectively. Furthermore, we analyze the underlying reasons why the emergence of these sinks does not conflict with correct prediction.

4.1 First Token Sink

This subsection provides a theoretical analysis of the *first token sink* phenomenon. We begin by formally defining the dataset structure and initialization assumptions, followed by the presentation of our main convergence results.

Dataset. We construct a synthetic dataset $\mathcal{D}_1$ consisting of N sequences, denoted as $\{(\boldsymbol{X}_i, \boldsymbol{x}_{i,T+1})\}_{i=1}^N$. Each sequence contains T context tokens $\boldsymbol{X}_i = [\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,T}]^\top \in \mathbb{R}^{T\times d}$, and the task is to predict the next token $\boldsymbol{x}_{i,T+1}$. We enforce a *global uniqueness* constraint on the tokens: every token instance in the dataset is distinct across all sequences and positions. Formally, $\boldsymbol{x}_{i,j} \neq \boldsymbol{x}_{i',j'}$ for any $(i,j) \neq (i',j')$. All tokens are represented as one-hot vectors. To satisfy this uniqueness, the embedding dimension d is set such that $d \geq \Omega(N(T+1))$.

Extending Eq.(3) to a dataset of N sequences, we define the input representation for the second block corresponding to the i-th sequence $\boldsymbol{X}_i = \{\boldsymbol{x}_{i,1}, \ldots, \boldsymbol{x}_{i,T}\}$ as:

$$\boldsymbol{H}_i^{(1)} = [\boldsymbol{h}_{i,1}, \ldots, \boldsymbol{h}_{i,T}]^\top \in \mathbb{R}^{T\times 2d}. \tag{9}$$

Here, $\boldsymbol{H}_i^{(1)}$ serves as the intermediate hidden state matrix for the i-th sequence.

Assumption 4.1. The model parameters are initialized as follows: $\boldsymbol{W}_{QK}^{(1)}, \boldsymbol{W}_{QK}^{(2)}$, and $\boldsymbol{W}_V^{(2)}$ are initialized to zero matrices $\mathbf{0}$. The first-layer value matrix is initialized as the identity matrix, $\boldsymbol{W}_V^{(1)}(0) = \boldsymbol{I}_d$.

We set $\boldsymbol{W}_V^{(1)}(0) = \boldsymbol{I}_d$ to ensure that signal propagates effectively through the first layer at initialization, while setting other parameters to zero allows us to analyze the gradient dynamics of the attention mechanism starting from a neutral state.

Next, we present the main results regarding the first token sink phenomenon.

Theorem 4.2 (First Token Sink)**.** *Under Assumption 4.1, if the Transformer is trained on dataset $\mathcal{D}_1$ with a learning rate $\eta = \Theta(NT\ln d)$, then the attention weight allocated to the first token at the second layer satisfies:*

$$\frac{\exp(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,1})}{\sum_{j=1}^T \exp(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j})} \geq 1 - \frac{1}{d}$$

for all sequences $i \in [N]$ and step $t \geq 2$.

Theorem 4.2 provides a theoretical guarantee for the *first token sink* phenomenon. It demonstrates that after training, the attention mechanism in the second block becomes dominated by the first token. Specifically, when using the last token representation $\boldsymbol{h}_{i,T}$ as the query, the attention distribution assigns an attention weight close to 1 to the first token representation $\boldsymbol{h}_{i,1}$ (which corresponds to the input $\boldsymbol{x}_{i,1}$), effectively suppressing contributions from other positions.

To understand the formation of the **attention sink**, we analyze the underlying training dynamics. The following lemma identifies two critical properties of the optimization process that guarantee the emergence of this phenomenon.

Lemma 4.3 (Dominance and Monotonicity)**.** *Under the same assumptions as Theorem 4.2, for any sequence $i \in [N]$ and step $t \geq 0$, the following properties hold:*

1. ***Dominance in Value Projection:*** *The value matrix learns to prioritize the first token's contribution to the target logit y_i: $\boldsymbol{h}_{i,1}^\top \boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_{y_i} \geq \boldsymbol{h}_{i,j}^\top \boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_{y_i}, \quad \forall j \in [2,T]$.*
2. ***Non-decreasing Attention:*** *Let $a_{i,j}(t)$ denote the attention score on the j-th token of the i-th sequence at step t: $a_{i,j}(t) = \frac{\exp(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j})}{\sum_{k=1}^T \exp(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,k})}$. Then, the attention on the first token is monotonically non-decreasing: $a_{i,1}(t+1) \geq a_{i,1}(t)$.*

Lemma 4.3 characterizes two critical dynamics: the preferential learning of the first token by the value projection $\boldsymbol{W}_V$, and the monotonic growth of its attention weight.

This phenomenon fundamentally originates from the structural asymmetry induced by the *causal mask*. Since the first token is the only position visible to the entire sequence, its features are broadcasted to the representations of all subsequent tokens. Consequently, the first token acts as a globally shared component in the input to the second block. This ubiquitous presence exposes $\boldsymbol{W}_V$ to the first token's features at every position, driving the value projection to preferentially utilize the first token's features for prediction (Property 1).

Crucially, the bias of $\boldsymbol{W}_V^{(2)}$ towards the first token amplifies the gradient signal associated with that position. During back-propagation, the chain rule dictates that the gradient with respect to the attention weight is proportional to the magnitude of the value projection. Since the first token has the strongest value projection, it receives the largest gradient update, strictly incentivizing the query-key mechanism to increase the attention allocated to the first token during training (Property 2). A higher attention weight, in turn, amplifies the gradient descent through the first token, further reinforcing the value matrix's preference. This cyclic interaction inevitably drives the system towards the attention sink state.

4.2 Weak Semantic Token Sink

In addition to the first token sink, another prevalent phenomenon in Transformer training is the concentration of attention on tokens that serve structural rather than semantic roles, such as punctuation (e.g., "`,`", "`.`") or separators (e.g., "`\n`"). We term this phenomenon the *weak semantic token sink*. In this subsection, we investigate the theoretical mechanisms driving this behavior. We first formalize the definition of weak semantic tokens based on their statistical properties and introduce the corresponding dataset construction. Subsequently, we present our main theoretical results demonstrating the emergence of the weak semantic token sink.

Definition of Weak Semantic Tokens. Empirically, these structural tokens exhibit distinctive characteristics: high frequency, widespread occurrence across diverse sequences, and crucially, negligible correlation with the specific next-token prediction. Building upon the concept of "common tokens" proposed by Tian et al. (2023), we formally define *Weak Semantic Tokens* as tokens that maintain a fixed, shared representation throughout the dataset while lacking discriminative power for the classification task (e.g., "`,`", "`.`"). Unlike the *globally unique tokens* (defined in Section 4.1) which possess strong semantic content to drive prediction, weak semantic tokens primarily function as syntactic delimiters. We explicitly model this property in the dataset constructed below.

Dataset Construction. We construct a synthetic dataset $\mathcal{D}_2$ with N sequences. Formally, the i-th sequence $\boldsymbol{X}_i$ comprises $C+1$ semantic segments interleaved with weak semantic tokens, structured as:

$$\boldsymbol{X}_i = [\boldsymbol{U}_{i,0}, \boldsymbol{w}, \boldsymbol{U}_{i,1}, \boldsymbol{w}, \ldots, \boldsymbol{w}, \boldsymbol{U}_{i,C}]^\top, \tag{10}$$

where $\boldsymbol{U}_{i,k} = [\boldsymbol{x}_{i,k(T+1)+1}, \cdots, \boldsymbol{x}_{i,k(T+1)+T}] \in \mathbb{R}^{T\times d}$ denotes the k-th semantic segment consisting of T tokens, and $\boldsymbol{w} \in \mathbb{R}^d$ is the weak semantic token. The total sequence length is $T' = (C+1)T + C$. Consistent with our previous analysis, we represent all input tokens as one-hot vectors.

To facilitate the theoretical analysis, we define the index set of weak tokens as $\mathcal{C} = \{k(T+1) \mid k = 1, \ldots, C\}$, and impose the following constraints on the data geometry and scale:

- **Unique Semantic Tokens ($j \notin \mathcal{C}$):** Tokens $\boldsymbol{x}_{i,j}$ within semantic segments ($\boldsymbol{U}_{i,k}$) are mutually orthogonal across the entire dataset. To satisfy this strict orthogonality, we require the embedding dimension to be sufficiently large: $d = \Omega(NT')$. These tokens are strongly correlated with the next token as they are exclusive to the i-th sequence (corresponding to the unique target $\boldsymbol{x}_{i,T'+1}$).
- **Weak Semantic Tokens ($j \in \mathcal{C}$):** Tokens $\boldsymbol{x}_{i,j}$ at indices $j \in \mathcal{C}$ share the identical embedding $\boldsymbol{w}$. These tokens exhibit only weak correlation with the next token due to their ubiquitous presence in every sequence.
- **Data Scale:** We assume the number of tokens in semantic segments, the number of weak semantic tokens, and the number of sequences (classes) are sufficiently large, satisfying: $C, T = \Omega(1)$ and $N = \Omega(T^4 \log T)$.

Prediction Objective. The goal is to predict the next token immediately following the final segment. We assign a unique class label $y_i \in [N]$ to the i-th sequence. Accordingly, the target token is defined as the one-hot vector corresponding to this class: $\boldsymbol{x}_{i,T'+1} = \boldsymbol{e}_{y_i}$.

Theorem 4.4 (Weak Semantic Token Sink)**.** *Under Assumption 4.1, if the Transformer is trained on dataset $\mathcal{D}_2$ with training rate $\eta = O(1)$, then there exists $t \leq \frac{N}{4\eta}$, such that*

$$\frac{\sum_{l\in\mathcal{C}} \exp(\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,l})}{\sum_{j=1}^{T} \exp(\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j})} \geq 1 - \frac{1}{T}.$$

Theorem 4.4 confirms that the attention mechanism will sink to the weak semantic tokens, assigning them a total attention weight close to 1.

This behavior is intuitively explained by the statistical properties of the dataset: since weak semantic tokens appear with high frequency both within and across sequences, they accumulate significantly larger gradients than the sparse unique tokens. This gradient dominance naturally drives the attention mechanism to focus on them.

4.3 Compatibility of Attention Sinks with Prediction Accuracy

In this section, we analyze why the concentration of attention on sink tokens does not compromise the model's prediction capacity. We begin by formalizing the final prediction step. Let $\hat{\boldsymbol{x}}_{i,T+1}$ denote the output logit vector for the i-th sequence. This vector is computed by aggregating the projected values of the hidden states $\boldsymbol{h}_{i,j}$ from the second block, weighted by the attention scores $a_{i,j}$:

$$\hat{\boldsymbol{x}}_{i,T+1}^\top = \sum_{j=1}^{T} a_{i,j} \cdot \left(\boldsymbol{h}_{i,j}^\top \boldsymbol{W}_V^{(2)}\right). \tag{11}$$

For dataset $\mathcal{D}_1$, the structural bias is naturally benign. Since all tokens in sequence i share the same target label y_i, the value projection maps them toward a class-congruent direction: $\hat{\boldsymbol{x}}_{i,T+1} = \boldsymbol{e}_{y_i}$. Consequently, the logit summation in Eq. (11) remains aligned with the correct class regardless of the attention distribution, even if it collapses entirely onto the first token.

In contrast, the emergence of sinks on class-agnostic tokens in $\mathcal{D}_2$ presents a paradox: how is accurate prediction possible when attention focuses on uninformative features? This is resolved by the causal mask, which enables these tokens to aggregate preceding discriminative signals and act as *context carriers*. We introduce the following lemma to illustrate this mechanism.

Lemma 4.5 (Divergent Projection Dynamics of Token Types)**.** *Let $\alpha_{i,j,k}(t) = [\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_k$ and $\beta_{i,j,k}(t) = [\boldsymbol{0}_{1\times d}, \boldsymbol{x}_{i,j}^\top]\boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_k$ denote the projection scores associated with the residual and attention pathways, respectively. We distinguish their behavior based on the token type of $\boldsymbol{x}_{i,j}$:*

1. ***Weak Semantic Token Uniformity:** If $\boldsymbol{x}_{i,j}$ is a weak semantic token ($j \in \mathcal{C}$), the projection magnitudes are consistent across all classes: $\alpha_{i,j,y_k}(t) = \alpha_{i,j,y_{k'}}(t), \quad \beta_{i,j,y_k}(t) = \beta_{i,j,y_{k'}}(t), \forall k, k' \in [N]$. This indicates that the embedding $\boldsymbol{w}$ does not favor any specific class label.*
2. ***Unique Semantic Token Dominance:** For $t \geq 1$, if $\boldsymbol{x}_{i,j}$ is a unique semantic token associated with class y_i ($j \notin \mathcal{C}$), the projection exhibits strong dominance for the target class: $\alpha_{i,j,y_i}(t) > \max_{k\neq y_i} \alpha_{i,j,k}(t), \quad \beta_{i,j,y_i}(t) > \max_{k\neq y_i} \beta_{i,j,k}(t)$.*

Lemma 4.5 characterizes the divergent learning dynamics of $\boldsymbol{W}_V^{(2)}$: it projects unique semantic tokens discriminatively towards their specific targets, whereas for the weak semantic token $\boldsymbol{w}$, gradients from diverse classes counterbalance each other.

Now assume that $\boldsymbol{x}_{i,j}$ is a weak semantic token ($j \in \mathcal{C}$). Its input representation for the second block, $\boldsymbol{h}_{i,j}^{(1)}$, comprises the residual connection and the output from the first attention layer:

$$\boldsymbol{h}_{i,j} = [\boldsymbol{x}_{i,j}^\top, \sum_{k=1}^{j} A_{i,j,k}\boldsymbol{x}_{i,k}^\top]^\top, \tag{12}$$

where $A_{i,j,k}$ denotes the attention weight from token j to k in sequence i.

Applying the projection $\boldsymbol{W}_V^{(2)}$, we decompose the output into two distinct components based on the token source:

$$\boldsymbol{h}_{i,j}^\top \boldsymbol{W}_V^{(2)} = \underbrace{[\boldsymbol{w}^\top, \sum_{k\le j, k\in\mathcal{C}} A_{i,j,k}\boldsymbol{w}^\top]\boldsymbol{W}_V^{(2)}}_{\text{Weak Semantic Projection}} + \underbrace{\sum_{k\le j, k\notin\mathcal{C}} A_{i,j,k}[\boldsymbol{0}, \boldsymbol{x}_{i,k}^\top]\boldsymbol{W}_V^{(2)}}_{\text{Unique Semantic Projection}}. \tag{13}$$

The first term aggregates the residual connection and attention to other weak semantic tokens. As shown in Lemma 4.5, due to the counterbalancing gradients, this component is isotropic, contributing only a uniform bias to the logits across all classes.

The second term aggregates the unique semantic tokens. By Lemma 4.5, $\boldsymbol{W}_V^{(2)}$ maps these tokens discriminatively towards the target class y_i. Consequently, the y_i-th entry dominates the logits, enabling the weak semantic token to function as a *context carrier* that drives correct prediction.

In this sense, this mechanism reconciles the apparent contradiction between the *attention sink* phenomenon and prediction accuracy. While the static embedding of a weak semantic token is natively class-agnostic (as shown in Lemma 4.5), the causal masking mechanism fundamentally alters its role in the intermediate layers. By aggregating the preceding context, the weak token transforms from a neutral symbol into a composite representation that encodes the discriminative features of the unique semantic tokens. Consequently, the "contextualized" weak token becomes higher correlated with the target label. Thus, rather than implying an abandonment of semantic information, the sink phenomenon can be interpreted as the model utilizing these tokens as carriers to recover the necessary signals, a behavior that remains consistent with optimizing the prediction performance.

5 Proof Sketch

This section outline the proof logic, characterizing how gradient dynamics driven by specific structural and distributional biases induce the emergence of attention sinks. Detailed proofs are provided in Appendix.

5.1 Proof Sketch for First Token Sink

Decomposition of Dynamics and Orthogonality. A key technical challenge in analyzing Transformer dynamics is the coupling of parameters across different sequences. However, in dataset $\mathcal{D}_1$, every token is unique and represented as an orthogonal one-hot vector. This global uniqueness ensures that the gradients generated by sequence i only affect the rows of weight matrices $\boldsymbol{W}_{QK}^{(2)}$ and $\boldsymbol{W}_V^{(2)}$ corresponding to its specific tokens. Consequently, the training processes of different sequences are effectively decoupled, allowing us to analyze a single representative sequence $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ with target y without loss of generality.

Based on the identity initialization of the first block, the input to the second block $\boldsymbol{h}_i$ for each token i is a concatenation of its original embedding and the mean of its preceding context:

$$\boldsymbol{h}_i = \begin{bmatrix} \boldsymbol{x}_i \\ \frac{1}{i}\sum_{j=1}^{i} \boldsymbol{x}_j \end{bmatrix} \in \mathbb{R}^{2d}. \tag{14}$$

To track the evolution of the second block, we decompose the value projection $\boldsymbol{W}_V^{(2)}$ into two scalar projection scores, $\alpha_i(t)$ and $\beta_i(t)$, which correspond to the *residual pathway* and the *attention pathway* respectively:

$$\alpha_i(t) = [\boldsymbol{x}_i^\top, \boldsymbol{0}]\boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_y, \tag{15}$$

$$\beta_i(t) = [\boldsymbol{0}, \boldsymbol{x}_i^\top]\boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_y. \tag{16}$$

Here, α_i captures the direct contribution of the i-th token, while β_i captures its contribution through the aggregated context from the first block.

Origin of Structural Bias. At the start of training ($t = 0$), the attention weights are uniformly distributed ($a_i(0) = 1/T$) due to zero initialization of $\boldsymbol{W}_{QK}^{(2)}$. The discrete-time gradient updates for these scores are:

$$\alpha_i(t+1) = \alpha_i(t) + \eta a_i(t)(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y), \tag{17}$$

$$\beta_i(t+1) = \beta_i(t) + \eta \left(\sum_{j=i}^{T} \frac{1}{j} a_j(t) \right) (1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y). \tag{18}$$

While the updates for $\alpha_i(1)$ are identical for all i, the summation term in the β_i update creates a significant disparity. Because the first token ($i = 1$) is visible to every subsequent position in the causal mask, it accumulates gradient signals from the entire sequence. As i increases, the number of summed terms decreases, leading to a strictly decreasing magnitude:

$$\beta_1(1) > \beta_2(1) > \cdots > \beta_T(1). \tag{19}$$

This initial "value dominance" of the first token acts as the primary catalyst. In the following step ($t = 1$), the gradient of the loss with respect to $\boldsymbol{W}_{QK}^{(2)}$ is proportional to these value projections, directly boosting the attention weight a_1.

Positive Feedback and Sink Locking. We prove that with a learning rate $\eta = \Theta(T \ln d)$, the structural bias is strong enough to push the attention mass of the first token to near-saturation after only two steps:

$$a_1(2) = \frac{\exp(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_1)}{\sum_{j=1}^{T} \exp(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_j)} \geq 1 - \frac{1}{d}. \tag{20}$$

Once the model enters this regime, a self-reinforcing loop is established: the high attention $a_1(t)$ ensures that $\beta_1(t)$ grows faster than any other $\beta_j(t)$, which in turn provides a larger gradient signal to keep the attention anchored on the first token. This feedback loop "locks" the system into a stable sink state.

5.2 Proof Sketch for Weak Semantic Token Sink

The emergence of sinks on uninformative tokens in $\mathcal{D}_2$ follows a parallel logic but is driven by *frequency disparity* rather than positional visibility. In $\mathcal{D}_2$, weak semantic tokens (at indices $\mathcal{C}$) appear ubiquitously across all N sequences in the dataset, whereas unique semantic tokens appear only once.

We define $\alpha_{i,j,k}(t) = [\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}]\boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_k$ to track the learning of individual token embeddings. The core mechanism is revealed by comparing the update rules for the target class y_i:

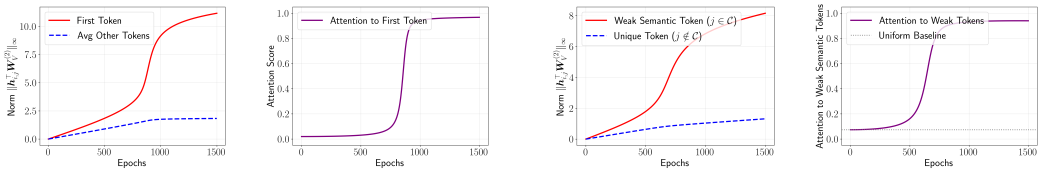
1. **Unique Tokens ($j \notin \mathcal{C}$):** The update is purely local, depending only on the attention weight within its own sequence:

$$\alpha_{i,j,y_i}(t+1) = \alpha_{i,j,y_i}(t) + \frac{\eta}{N} a_{i,j}(t) \left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i}\right). \tag{21}$$

2. **Weak Tokens ($j \in \mathcal{C}$):** The update is global, aggregating gradients from every sequence in the entire dataset where this common token appears:

$$\alpha_{i,j,y_{i'}}(t+1) = \alpha_{i,j,y_{i'}}(t) + \frac{\eta}{N} \Big(\sum_{j' \in \mathcal{C}} a_{i',j'}(t) - \sum_{m=1}^{N} \sum_{j' \in \mathcal{C}} a_{m,j'}(t) \mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_{y_{i'}} \Big). \tag{22}$$

Since the weak token w is present in every sequence, it effectively "collects" gradients N times more frequently than any unique token. Because the softmax probability $\mathcal{S}(\cdot)$ remains small during the early training phase, the negative "noise" terms (interference from other classes) do not counteract this growth. Consequently, the projection norm for weak tokens grows at a rate that significantly outpaces unique tokens ($\alpha_{\text{weak}} \gg \alpha_{\text{unique}}$). This dominant signal in the value matrix creates an overwhelming gradient for the query-key mechanism, forcing the model to progressively anchor its attention on these high-frequency, class-agnostic tokens.



(a) First Token Sink Dynamics

(b) Weak Semantic Token Sink

Figure 1: Evolution of the value projection norm $\|\boldsymbol{h}_{i,j}^\top \boldsymbol{W}_V^{(2)}\|_\infty$ (left) and the corresponding attention trajectory (right) for: (a) the first token sink in dataset $\mathcal{D}_1$ and (b) the weak semantic tokens sink in $\mathcal{D}_2$. In each subplot, the red solid curves denote the designated sink tokens (the initial token or weak semantic tokens), while the blue dashed curves represent the average of all other informative tokens. This empirical progression aligns with our theoretical results.

6 Experiments

We verify our theoretical results on synthetic datasets constructed according to the definition of $\mathcal{D}_1$ and $\mathcal{D}_2$. We train a two-block Transformer using Gradient Descent with learning rate $\eta = 1$. Specifically, $\mathcal{D}_1$ (*first token sink*) consists of $N = 32$ sequences with context length $T = 50$ and embedding dimension $d = 1,632$. $\mathcal{D}_2$ (*weak semantic token sink*) consists of $N = 32$ sequences, each containing $C = 4$ weak semantic tokens and semantic segments of length $T = 10$, with embedding dimension $d = 1,888$. More experimental results can be found in Appendix A.

Dynamics of First Token Sink. Figure 1(a) illustrates the formation process of the first token sink. Consistent with our theoretical predictions, the magnitude of the value projection associated with the first token grows significantly faster than that of other tokens (left of Figure 1(a)). This dominance in magnitude directly drives the attention mechanism: as shown in the right of Figure 1(a), the attention weight allocated to the first token monotonically increases, eventually approaching 1. Notably, we observe a self-reinforcing acceleration phase (e.g., around epoch 800), where the increasing sparsity of attention further amplifies the value projection growth of the first token.

Dynamics of Weak Semantic Token Sink. We observe a parallel phenomenon in the weak semantic token setting, as depicted in Figure 1(b). The weak semantic tokens rapidly establish dominant value projections compared to unique semantic tokens (left of Figure 1(b)). Consequently, the model learns to concentrate the majority of its attention mass on these specific tokens (right of Figure 1(b)), effectively verifying that the weak semantic tokens with high frequency naturally evolve into attention sinks during training.

7 Conclusion

In this work, we provide the first rigorous theoretical analysis of the training dynamics underlying attention sinks in Transformers. By investigating a two-block architecture, we identify the distinct optimization incentives that drive this phenomenon: the First Token Sink emerges as a structural necessity due to the gradient bias inherent in causal masking, while the Weak Semantic Token Sink is driven by the frequency-dependent accumulation of gradients on common tokens. Crucially, our analysis resolves the paradox between the concentration of attention on uninformative tokens and the model's predictive performance. We demonstrate that in deep networks, these sink tokens are not passive artifacts but are actively repurposed as context carriers. By aggregating features from preceding tokens in the first block, they allow subsequent layers to retrieve necessary semantic signals while maintaining stable attention distributions. These insights not only deepen our mechanistic understanding of Transformer training but also provide a theoretical foundation for designing new architectures. One limitation of our current framework is the focus on a simplified two-block setting. Future directions include extending this theoretical analysis to multi-layer Transformers and general natural language distributions to further bridge the gap between theory and practice.

REFERENCES

Yongqi An, Xu Zhao, Tao Yu, Ming Tang, and Jinqiao Wang. Systematic outliers in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Michael Bronstein, Petar Veličković, and Razvan Pascanu. Why do llms attend to the first token? *arXiv preprint arXiv:2504.02732*, 2025.

Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 1877–1901, 2020.

Nicola Cancedda. Spectral filters, dark signals, and attention sinks. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4792–4808, 2024.

Guoxuan Chen, Han Shi, Jiawei Li, Yihang Gao, Xiaozhe Ren, Yimeng Chen, Xin Jiang, Zhenguo Li, Weiyang Liu, and Chao Huang. SepLLM: Accelerate large language models by compressing one segment into one separator. In *Forty-second International Conference on Machine Learning*, 2025.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *The Twelfth International Conference on Learning Representations*, 2024.

Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025.

Wei Huang, Haotong Qin, Yangdong Liu, Yawei Li, Qinshuo Liu, Xianglong Liu, Luca Benini, Michele Magno, Shiming Zhang, and XIAOJUAN QI. Slim-LLM: Salience-driven mixed-precision quantization for large language models. In *Forty-second International Conference on Machine Learning*, 2025.

Samy Jelassi, Michael Eli Sander, and Yuanzhi Li. Vision transformers provably learn spatial structure. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022.

Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. How do nonlinear transformers learn and generalize in in-context learning? In *Forty-first International Conference on Machine Learning*, 2024.

Yuchen Li, Yuanzhi Li, and Andrej Risteski. How do transformers learn topic structure: Towards a mechanistic understanding. In *International Conference on Machine Learning*, pp. 19689–19729, 2023.

Haokun Lin, Haobo Xu, Yichen Wu, Jingzhi Cui, Yingtao Zhang, Linzhan Mou, Linqi Song, Zhenan Sun, and Ying Wei. Duquant: Distributing outliers via dual transformation makes stronger quantized LLMs. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024.

Ruikang Liu, Haoli Bai, Haokun Lin, Yuening Li, Han Gao, Zhengzhuo Xu, Lu Hou, Jun Yao, and Chun Yuan. Intactkv: Improving large language model quantization by keeping pivot tokens intact. In *62nd Annual Meeting of the Association for Computational Linguistics (ACL 2024)*, pp. 7716–7741. Association for Computational Linguistics, 2024.

Aniruddha Nrusimha, Mayank Mishra, Naigang Wang, Dan Alistarh, Rameswar Panda, and Yoon Kim. Mitigating the impact of outlier channels for language model quantization with activation regularization. *arXiv preprint arXiv:2404.03605*, 2024.

OpenAI. Gpt-4 technical report, 2023.

Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri. What are you sinking? a geometric approach on attention sink. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Chen Siyu, Sheen Heejune, Wang Tianhao, and Yang Zhuoran. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality (extended abstract). In *Proceedings of Thirty Seventh Conference on Learning Theory*, pp. 4573–4573, 2024.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024a.

Mingjie Sun, Zhuang Liu, Anna Bair, and J Zico Kolter. A simple and effective pruning approach for large language models. In *The Twelfth International Conference on Learning Representations*, 2024b.

Yuandong Tian, Yiping Wang, Beidi Chen, and Simon Shaolei Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Yuandong Tian, Yiping Wang, Zhenyu Zhang, Beidi Chen, and Simon Shaolei Du. JoMA: Demystifying multilayer transformers via joint dynamics of MLP and attention. In *The Twelfth International Conference on Learning Representations*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pp. 6000–6010, 2017.

Zhongwei Wan, Ziang Wu, Che Liu, Jinfa Huang, Zhihong Zhu, Peng Jin, Longyue Wang, and Li Yuan. Look-m: Look-once optimization in kv cache for efficient multimodal long-context inference. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 4065–4078, 2024.

Xiuying Wei, Yunchen Zhang, Xiangguo Zhang, Ruihao Gong, Shanghang Zhang, Qi Zhang, Fengwei Yu, and Xianglong Liu. Outlier suppression: pushing the limit of low-bit transformer language models. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, pp. 17402–17414, 2022.

Haoyi Wu and Kewei Tu. Layer-condensed kv cache for efficient inference of large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11175–11188, 2024.

Xinyi Wu, Yifei Wang, Stefanie Jegelka, and Ali Jadbabaie. On the emergence of position bias in transformers. In *ICML 2025 Workshop on Methods and Opportunities at Small Scale*, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *The Twelfth International Conference on Learning Representations*, 2024.

Lu Yin, You Wu, Zhenyu Zhang, Cheng-Yu Hsieh, Yaqing Wang, Yiling Jia, Gen Li, AJAY KUMAR JAISWAL, Mykola Pechenizkiy, Yi Liang, Michael Bendersky, Zhangyang Wang, and Shiwei Liu. Outlier weighed layerwise sparsity (OWL): A missing secret sauce for pruning LLMs to high sparsity. In *Forty-first International Conference on Machine Learning*, 2024.

Zhongzhi Yu, Zheng Wang, Yonggan Fu, Huihong Shi, Khalid Shaikh, and Yingyan Celine Lin. Unveiling and harnessing hidden attention sinks: Enhancing large language models without training through attention calibration. In *Forty-first International Conference on Machine Learning*, 2024.

Jingzhao Zhang, Sai Praneeth Karimireddy, Andreas Veit, Seungyeon Kim, Sashank Reddi, Sanjiv Kumar, and Suvrit Sra. Why are adaptive methods good for attention models? In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 15383–15393, 2020.

Stephen Zhang, Mustafa Khan, and Vardan Papyan. Attention sinks: A 'catch, tag, release' mechanism for embeddings. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Zhenyu Zhang, Shiwei Liu, Runjin Chen, Bhavya Kailkhura, Beidi Chen, and Zhangyang Wang. Q-hitter: A better token oracle for efficient llm inference via sparse-quantized kv cache. *Proceedings of Machine Learning and Systems*, 6:381–394, 2024.

AI USE STATEMENT

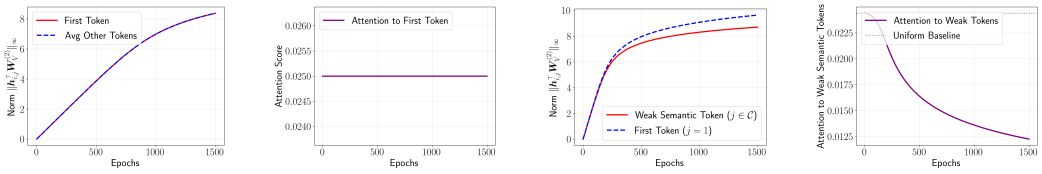
In this work, we used generative AI tools only for language editing and polishing of the manuscript. We have not used generative AI tools for research ideation, methodology development, theoretical derivations or proofs, experimental design or execution, data analysis, or code generation; other disclosure categories are not applicable to this work. We reviewed all AI-assisted edits and verified that they preserve the intended technical meaning. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

APPENDIX OVERVIEW

- **Appendix A**: **Additional Experiments.** We conduct ablation studies regarding architectural components and dataset configurations, and provide further empirical validation on natural language using the real-world dataset.
- **Appendix B**: **Basic Calculation.** We formally derive the internal representations and the associated gradient formulas for the Transformer.
- **Appendix C**: **First Token Sink.** We provide the theoretical analysis regarding the training dynamics of the *first token sink*.
- **Appendix D**: **Weak Semantic Token Sink.** We provide the theoretical analysis regarding the training dynamics of the *weak semantic tokens*.
- **Appendix E**: **Complete Proof.** We present detailed, step-by-step proofs for all lemmas and theorems introduced in the paper.

A ADDITIONAL EXPERIMENTS

To verify the mechanistic origins of attention sinks, we conduct two ablation studies targeting the causal mask and token frequency, respectively. The results are visualized in Figure 2.



(a) Val. Proj. (No Mask) (b) Attn. (No Mask) (c) Val. Proj. ($C = 1$) (d) Attn. ($C = 1$)

Figure 2: **Ablation Studies on Sink Formation. (a)-(b) Ablation on Causal Masking:** When training on $\mathcal{D}_1$ without a causal mask, the first token's value projection (Red) does not diverge from others (Blue), and attention does not collapse. **(c)-(d) Ablation on Frequency:** When training on $\mathcal{D}_2$ with a single weak semantic token ($C = 1$), the weak token (Red) fails to establish value dominance or attract significant attention, unlike the high-frequency setting.

Ablation on Causal Masking (Dataset $\mathcal{D}_1$). We verify the necessity of the causal mask for the First Token Sink by removing it during training on $\mathcal{D}_1$ ($N = 32, T = 40$). As shown in Figure 2 (a) and (b), in the absence of the causal constraint, the gradient bias vanishes. Specifically, the first token's value projection norm remains comparable to other tokens, and the attention on the first token remains at baseline levels throughout training.

Ablation on Token Frequency (Dataset $\mathcal{D}_2$). We verify the role of token frequency for the Weak Semantic Token Sink by reducing the weak semantic token's occurrence to once per sequence ($C = 1$) in $\mathcal{D}_2$ ($N = 8, T = 20$). Figures 2 (c) and (d) demonstrate that without the cumulative gradient advantage provided by high frequency, the single weak semantic token fails to evolve into a sink. Contrary to the sink behavior, its value projection fails to gain dominance and its attention scores degrade over training, confirming that high recurrence is an essential driving force enabling weak semantic tokens to aggregate attention.

Real-world Experiments on WikiText-2 Moving beyond synthetic settings, we conducted experiments using a two-block multi-head attention architecture trained from scratch on the WikiText-2 dataset to verify if the sink phenomenon persists in natural language. To monitor the attention dynamics during the optimization process, we tracked the attention distribution over a representative narrative sentence:

> *"scientists discovered an ancient , mysterious artifact hidden deep in the dense , dark forest , unlocking secrets of the past ."*

Specifically, we track the trajectories of the attention mass for the top-3 attended tokens at different training stages. The results are summarized in Table 1.

Table 1: **Trajectories of Attention Mass on WikiText-2.** The table shows the evolution of attention weights allocated to the first token and two weak semantic tokens. By 15k steps, these three tokens collectively capture over 52% of the total attention mass.

Step	**0**	**3k**	**6k**	**9k**	**12k**	**15k**
"scientists" (First token)	0.0565	0.0074	0.0188	0.0751	0.1040	0.1624
the last "the" (Weak token)	0.0343	0.0667	0.0944	0.0975	0.1062	0.1047
the last "." (Weak token)	0.0545	0.2488	0.2550	0.2192	0.2985	0.2611
Sum of Top-3	**0.1453**	**0.3229**	**0.3682**	**0.3918**	**0.5087**	**0.5282**

As shown in Table 1, the first token and specific weak semantic tokens (such as punctuation and common function words) exhibit a clear trend of accumulating attention mass as training progresses. Although their individual weights fluctuate slightly, they collectively monopolize over 52% of the attention mass by the end of training. This observation confirms our theoretical premise that these tokens function synergistically as distributed attention sinks in models trained on real-world data distributions.

B Basic Calculation

In this section, we formally derive the internal representations, and the associated gradients of the Transformer model.

Lemma B.1 (Internal Representation)**.** *Under Assumption 4.1, given an input sequence $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, the internal representation for the i-th token ($\boldsymbol{x}_i$) is as follows*

$$\boldsymbol{h}_i = \begin{bmatrix} \boldsymbol{x}_i \\ \sum_{j=1}^{i} \frac{1}{i} \boldsymbol{x}_j \end{bmatrix}$$

Based on Lemma B.1, we further derive the gradients of the Transformer's second block as follows.

Lemma B.2 (Gradients Calculation)**.** *Assume that the Transformer is trained on dataset $\mathcal{D}$ with N samples $\{\boldsymbol{x}_{i,1}, \cdots, \boldsymbol{x}_{i,T}; \boldsymbol{x}_{i,T+1}\}_{i=1}^{N}$, where $\boldsymbol{x}_{i,j}$ denotes the j-th token of the i-th sequence. The gradients for the parameters of the Transformer's second block are as follows.*

$$\nabla_{\boldsymbol{W}_{QK}^{(2)}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{i=1}^{N} \begin{bmatrix} \boldsymbol{x}_{i,T} \\ \frac{1}{T} \sum_{j=1}^{T} \boldsymbol{x}_{i,j} \end{bmatrix} \boldsymbol{g}_i^\top \boldsymbol{W}_V^{(2)\top} \boldsymbol{H}_i^{(1)\top} \left(diag(\boldsymbol{a}_i) - \boldsymbol{a}_i \boldsymbol{a}_i^\top \right) \boldsymbol{H}_i^{(1)}$$

$$\nabla_{\boldsymbol{W}_{V}^{(2)}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = -\frac{1}{N} \sum_{i=1}^{N} \boldsymbol{H}_i^{(1)\top} \boldsymbol{a}_i \boldsymbol{g}_i^\top$$

where $\boldsymbol{a}_i = \mathcal{S}\left(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)} \boldsymbol{H}_i^{(1)\top} \right)^\top$ denotes the attention vector for the i-th sequence, $\boldsymbol{h}_{i,j}^\top$ denotes the internal representations for the j-th token of the i-th sample. $\boldsymbol{g}_i = \boldsymbol{x}_{i,T+1} - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T+1})$ denotes the loss derivative for the i-th sequence.

C First Token Sink

In this section, we present the detailed theoretical analysis of the First Token Sink phenomenon when the two-block Transformer is trained on dataset $\mathcal{D}_1$. We first define the dynamics and introduce the key lemmas governing the evolution of value projections and attention scores. The complete proofs for these lemmas are provided in Appendix E.

By the definition of dataset $\mathcal{D}_1$, each token only appear in one sequence. By Lemma B.2 and the *one-hot* properties of the tokens' embedding, each token can only affect specific rows in $\boldsymbol{W}_{QK}^{(2)}$ and $\boldsymbol{W}_{V}^{(2)}$. Therefore, the training dynamics related to a sequence will not affect the other sequences if the model is trained on dataset $\mathcal{D}_1$. We can simplify the analysis by assuming that the Transformer is trained on only one sequence $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$ with the next token to be predicted denoted as $\boldsymbol{x}_{T+1}$.

To begin with, the input embedding is as follows

$$\boldsymbol{H}^{(0)} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]^\top \in \mathbb{R}^{T\times d}.$$

By Lemma B.1, the internal representations is as follows

$$\boldsymbol{h}_i = \begin{bmatrix} \boldsymbol{x}_i \\ \sum_{j=1}^{i} \frac{1}{i}\boldsymbol{x}_j^\top \end{bmatrix},$$

$$\boldsymbol{H}^{(1)} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_i, \ldots, \boldsymbol{h}_T]^\top = \begin{bmatrix} \boldsymbol{x}_1 & \ldots & \boldsymbol{x}_i & \ldots & \boldsymbol{x}_T \\ \boldsymbol{x}_1 & \ldots & \sum_{j=1}^{i} \frac{1}{i}\boldsymbol{x}_j & \ldots & \sum_{j=1}^{T} \frac{1}{T}\boldsymbol{x}_j \end{bmatrix}^\top.$$

Then the output equation $\hat{\boldsymbol{x}}_{T+1}^\top = \mathcal{S}\left(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)} \boldsymbol{H}^{(1)\top}\right) \boldsymbol{H}^{(1)} \boldsymbol{W}_V^{(2)}$ can be expanded into

$$\begin{aligned}
\hat{\boldsymbol{x}}_{T+1}^\top &= \mathcal{S}\left(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)} \boldsymbol{H}^{(1)\top}\right) \boldsymbol{H}^{(1)} \boldsymbol{W}_V^{(2)} \\
&= \mathcal{S}\left(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)} [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_i, \ldots, \boldsymbol{h}_T]\right) [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_i, \ldots, \boldsymbol{h}_T]^\top \boldsymbol{W}_V^{(2)} \\
&= \mathcal{S}\left(\left[\boldsymbol{x}_T^\top, \sum_{j=1}^{T} \frac{1}{T}\boldsymbol{x}_j^\top\right] \boldsymbol{W}_{QK}^{(2)} \begin{bmatrix} \boldsymbol{x}_1 & \ldots & \boldsymbol{x}_T \\ \boldsymbol{x}_1 & \ldots & \sum_{j=1}^{T} \frac{1}{T}\boldsymbol{x}_j \end{bmatrix}\right) \begin{bmatrix} \boldsymbol{x}_1 & \ldots & \boldsymbol{x}_T \\ \boldsymbol{x}_1 & \ldots & \sum_{j=1}^{T} \frac{1}{T}\boldsymbol{x}_j \end{bmatrix}^\top \boldsymbol{W}_V^{(2)}
\end{aligned}$$

Proof idea sketch. To analyze the training dynamics of the Transformer, we decompose the output equation $\hat{\boldsymbol{x}}_{T+1}^\top = \mathcal{S}\left(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)} \boldsymbol{H}^{(1)\top}\right) \boldsymbol{H}^{(1)} \boldsymbol{W}_V^{(2)}$ into two components: the attention scores, determined by $\mathcal{S}\left(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)} \boldsymbol{H}^{(1)\top}\right)$ and the value projection $\boldsymbol{H}^{(1)} \boldsymbol{W}_V^{(2)}$.

We further decompose the internal representations $\boldsymbol{H}^{(1)}$ into two orthogonal parts: $[\boldsymbol{x}_i^\top, \boldsymbol{0}_{1\times d}]$ (corresponding to the residual path) and $[\boldsymbol{0}_{1\times d}, \boldsymbol{x}_i^\top]$ (corresponding to the output of the first block), and then analyze the dynamics of $\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)} \begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{0}_{d\times 1} \end{bmatrix}$, $\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)} \begin{bmatrix} \boldsymbol{0}_{d\times 1} \\ \boldsymbol{x}_i \end{bmatrix}$, $[\boldsymbol{x}_i^\top, \boldsymbol{0}_{1\times d}] \boldsymbol{W}_V^{(2)}$ and $[\boldsymbol{0}_{1\times d}, \boldsymbol{x}_i^\top] \boldsymbol{W}_V^{(2)}$. Finally we combine these results to derive the full dynamics of the attention block.

Lemma C.1 (Update Rules)**.** *We denote* $\alpha_i(t) = [\boldsymbol{x}_i^\top, \boldsymbol{0}_{1\times d}] \boldsymbol{W}_V^{(2)}(t) \boldsymbol{e}_y$, $\beta_i(t) = [\boldsymbol{0}_{1\times d}, \boldsymbol{x}_i^\top] \boldsymbol{W}_V^{(2)}(t) \boldsymbol{e}_y$, $\gamma_i(t) = \boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(t) \begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{0}_{d\times 1} \end{bmatrix}$ *and* $\rho_i(t) = \boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(t) \begin{bmatrix} \boldsymbol{0}_{d\times 1} \\ \boldsymbol{x}_i \end{bmatrix}$. *Then their update rules are as follows*

$$\alpha_i(t+1) = \alpha_i(t) + \eta a_i(t) \left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right)$$

$$\beta_i(t+1) = \beta_i(t) + \eta \sum_{j=i}^{T} \frac{1}{j} a_j(t) \left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right)$$

$$\gamma_i(t+1) = \gamma_i(t) + \eta \|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1} \left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right) \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} a_i(t) (\boldsymbol{e}_i - \boldsymbol{a}(t))$$

$$\rho_i(t+1) = \rho_i(t) + \eta\|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1}\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right) \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} \sum_{j=i}^{T} \frac{1}{j} a_j(t)(\boldsymbol{e}_j - \boldsymbol{a}(t))$$

where $\boldsymbol{a} = [a_1, \ldots, a_T]^\top$ and a_i is the attention score on the i-th token. Assuming that the category for this sequence is y, and thus $\boldsymbol{x}_{T+1} = \boldsymbol{e}_y$. $\mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y$ denotes the y-th element of $\mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))$.

Then we analyze the first two step of GD.

Lemma C.2. *Under the Assumption 4.1, with training rate $\eta = \Theta(T \ln d)$, the attention on the first token at the second step satisfies:*

$$\frac{\exp(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_1)}{\sum_{i=1}^{T} \exp(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_i)} \geq 1 - \frac{1}{d}.$$

Moreover, we have

$$\alpha_1(t) \geq \alpha_i(t), \beta_1(t) \geq \beta_i(t), \gamma_1(t) \geq \gamma_i(t), \rho_1(t) \geq \rho_i(t)$$

for $t \in \{0, 1, 2\}, i \in [2, T]$.

Now the attention on the first token is large enough, we then prove that the attention on it will keep growing.

Lemma C.3. *If for some t', $a_1(t) \geq \frac{1}{2}, \alpha_1(t) \geq \alpha_i(t), \beta_1(t) \geq \beta_i(t), \gamma_1(t) \geq \gamma_i(t), \rho_1(t) \geq \rho_i(t)$ for $i \in [2, T]$ hold at $t = t'$, then these inequalities hold for $t \geq t'$. Moreover, $a_1(t+1) \geq a_1(t)$ for $t \geq t'$.*

We next present the main results of the first token sink.

Theorem C.4 (Restatement of Theorem 4.2)**.** *Under Assumption 4.1, if the Transformer is trained on dataset $\mathcal{D}_1$ with training rate $\eta = \Theta(NT \ln d)$, then the attention on the first token at the second step satisfies:*

$$\frac{\exp(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,1})}{\sum_{j=1}^{T} \exp(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j})} \geq 1 - \frac{1}{d}$$

for $t \geq 2$.

Theorem C.4 is derived from Lemma C.2 and Lemma C.3. By Lemma B.2 and the *one-hot* properties of the tokens' embedding, each token can only affect specific rows in $\boldsymbol{W}_{QK}^{(2)}$ and $\boldsymbol{W}_V^{(2)}$, so we can easily extend the one-sequence results to N sequences, and we need to scale up the training rate η by N.

Lemma C.5 (Restatement of Lemma 4.3)**.** *Under the same assumption as Theorem C.4, for $i \in [N], t \geq 0$, the following inequalities hold:*

1. ***Dominance in Value:*** *$\boldsymbol{h}_{i,1}^\top \boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_y \geq \boldsymbol{h}_{i,j}^\top \boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_y$ for all $j \in [2, T]$.*
2. ***Non-decreasing Attention:*** *$\frac{\exp(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)}(t+1)\boldsymbol{h}_{i,1})}{\sum_{j=1}^{T} \exp(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)}(t+1)\boldsymbol{h}_{i,j})} \geq \frac{\exp(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,1})}{\sum_{j=1}^{T} \exp(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j})}$.*

The dominance in value of Lemma C.5 is derived from $\alpha_1(t) \geq \alpha_i(t), \beta_1(t) \geq \beta_i(t)$, and by the definition $\boldsymbol{h}_i^\top \boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_y = \alpha_i(t) + \sum_{j=1}^{i} \frac{1}{i}\beta_j(t)$, thus $\boldsymbol{h}_1^\top \boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_y \geq \boldsymbol{h}_i^\top \boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_y$. By extending the results to N sequences, we have $\boldsymbol{h}_{i,1}^\top \boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_y \geq \boldsymbol{h}_{i,j}^\top \boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_y$.

By Lemma C.2 and Lemma C.3, we have $a_1(t+1) \geq a_1(t)$ for $t \geq 0$, then we have the *Non-decreasing Attention* result.

D Weak Semantic Token Sink

In this section, we focus on the theoretical analysis of the Weak Semantic Token Sink phenomenon when the model is trained on dataset $\mathcal{D}_2$. We explicitly derive the gradient update rules to demonstrate

how the high frequency of weak semantic tokens drives their dominance in the value projection. Furthermore, we characterize the divergent optimization dynamics between these common tokens and the unique semantic tokens. The complete proofs for all lemmas and theorems presented here are provided in Appendix E.

Following Lemma B.2, the gradients for $\boldsymbol{W}_{QK}^{(2)}$ and $\boldsymbol{W}_{V}^{(2)}$ when trained on dataset $\mathcal{D}_2$ should be:

$$\nabla_{\boldsymbol{W}_{QK}^{(2)}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} \boldsymbol{x}_{i,T'} \\ \frac{1}{T'}\sum_{j=1}^{T'} \boldsymbol{x}_{i,j} \end{bmatrix} \boldsymbol{g}_i^\top \boldsymbol{W}_V^{(2)\top} \boldsymbol{H}_i^{(1)\top} \left(diag(\boldsymbol{a}_i) - \boldsymbol{a}_i \boldsymbol{a}_i^\top\right) \boldsymbol{H}_i^{(1)}$$

$$\nabla_{\boldsymbol{W}_{V}^{(2)}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{H}_i^{(1)\top} \boldsymbol{a}_i \boldsymbol{g}_i^\top$$

Then we have the following lemma.

Lemma D.1 (Update Rules)**.** *We denote* $\alpha_{i,j,k}(t) = [\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_k$ $\beta_{i,j,k}(t) = [\boldsymbol{0}_{1\times d}, \boldsymbol{x}_{i,j}^\top]\boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_k$, $\gamma_{i,j}(t) = \boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t) \begin{bmatrix} \boldsymbol{x}_{i,j} \\ \boldsymbol{0}_{d\times 1} \end{bmatrix}$ *and* $\rho_{i,j}(t) = \boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t) \begin{bmatrix} \boldsymbol{0}_{d\times 1} \\ \boldsymbol{x}_{i,j} \end{bmatrix}$. *Then their update rules are as follows*

$$\alpha_{i,j,k}(t+1) = \alpha_{i,j,k}(t) + \frac{\eta}{N} a_{i,j}(t)\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i}\right)$$

$$\beta_{i,j,k}(t+1) = \beta_{i,j,k}(t) + \frac{\eta}{N}\sum_{l=j}^{T'} \frac{1}{l} a_{i,l}(t)\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i}\right)$$

for $i \in [N], j \notin \mathcal{C}, k = y_i$.

$$\alpha_{i,j,k}(t+1) = \alpha_{i,j,k}(t) - \frac{\eta}{N} a_{i,j}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k$$

$$\beta_{i,j,k}(t+1) = \beta_{i,j,k}(t) - \frac{\eta}{N}\sum_{l=j}^{T'} \frac{1}{l} a_{i,l}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k$$

for $i \in [N], j \notin \mathcal{C}, k \neq y_i$.

$$\alpha_{i,j,k}(t+1) = \alpha_{i,j,k}(t) + \frac{\eta}{N}\left(\sum_{j'\in\mathcal{C}} a_{i',j'}(t) - \sum_{m=1}^{N}\sum_{j'\in\mathcal{C}} a_{m,j'}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_{y_{i'}}\right)$$

$$\beta_{i,j,k}(t+1) = \beta_{i,j,k}(t) + \frac{\eta}{N}\left(\sum_{j'\in\mathcal{C}}\sum_{l=j'}^{T'} \frac{1}{l} a_{i',l}(t) - \sum_{m=1}^{N}\sum_{j'\in\mathcal{C}}\sum_{l=j'}^{T'} \frac{1}{l} a_{m,l}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_{y_{i'}}\right)$$

for $i \in [N], j \in \mathcal{C}, k = y_{i'}$ *for some* $i' \in [N]$.

$$\alpha_{i,j,k}(t+1) = \alpha_{i,j,k}(t) - \frac{\eta}{N}\sum_{m=1}^{N}\sum_{j'\in\mathcal{C}} a_{m,j'}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_k$$

$$\beta_{i,j,k}(t+1) = \beta_{i,j,k}(t) - \frac{\eta}{N}\sum_{m=1}^{N}\sum_{j'\in\mathcal{C}}\sum_{l=j'}^{T'} \frac{1}{l} a_{m,l}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_k$$

for $i \in [N], j \in \mathcal{C}, k \neq y_{i'}$ *for* $\forall i' \in [N]$.

$$\gamma_{i,j}(t+1) = \gamma_{i,j}(t) + \frac{\eta}{N}\|\boldsymbol{h}_{i,T'}\|_2^2 \boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top} a_{i,j}(t)(\boldsymbol{e}_j - \boldsymbol{a}_i(t))$$

$$\rho_{i,j}(t+1) = \rho_{i,j}(t) + \frac{\eta}{N}\|\boldsymbol{h}_{i,T'}\|_2^2 \boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{k=j}^{T'} \frac{1}{k} a_{i,k}(t)(\boldsymbol{e}_k - \boldsymbol{a}_i(t))$$

for $i \in [N], j \notin \mathcal{C}$.

$$\gamma_{i,j}(t+1) = \gamma_{i,j}(t) + \frac{\eta}{N}\left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2}\right) \boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}} a_{i,j'}(t)(\boldsymbol{e}_{j'} - \boldsymbol{a}_i(t))$$

$$\rho_{i,j}(t+1) = \rho_{i,j}(t) + \frac{\eta}{N}\left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2}\right) \boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}}\sum_{k=j'}^{T'} \frac{1}{k} a_{i,k}(t)(\boldsymbol{e}_k - \boldsymbol{a}_i(t))$$

for $i \in [N], j \in \mathcal{C}$.

where $\boldsymbol{a}_i = [a_{i,1}, \ldots, a_{i,T'}]^\top$ *and* $a_{i,j}$ *is the attention score on the j-th token in the i-th sequence. Assuming that the category for the i-th sequence is* y_i*, and thus* $\boldsymbol{x}_{i,T'+1} = \boldsymbol{e}_{y_i}$*.* $\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_i$ *denotes the i-th element of* $\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))$.

We next bound the magnitude of the output.

Lemma D.2. *For* $t \le O(\frac{N}{\eta})$*, we have*

$$\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k \le \frac{2}{d}$$

for $k \in [d]$.

This bound ensures that the gradients backpropagated from the loss remain stable, thereby facilitating a tractable analysis of the dynamics for $\boldsymbol{W}_{QK}$ and $\boldsymbol{W}_V$.

We next establish the dominance of $\gamma_{i,j}(t)$ for $j \in \mathcal{C}$, which contribute to weak semantic token sink.

Lemma D.3. *For* $i \in [N], j \in \mathcal{C}, j' \notin \mathcal{C}, t \le \frac{N}{4\eta}$*, if* $\left(1 - \sum_{l\in\mathcal{C}} a_{i,l}(t)\right) \ge \frac{1}{2T}$*, we have*

$$\gamma_{i,j}(t) \ge T|\gamma_{i,j'}(t)|$$

$$\gamma_{i,j}(t) \ge \frac{1}{16}|\rho_{i,j}(t)|$$

$$\gamma_{i,j}(t) \ge T|\rho_{i,j'}(t)|$$

$$\gamma_{i,j}(t+1) \ge \gamma_{i,j}(t) + \frac{\eta}{5N}\left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2}\right)\sum_{j'\in\mathcal{C}} a_{i,j'}(t)\left(1 - \sum_{l\in\mathcal{C}} a_{i,l}(t)\right)\alpha_{i,j,y_i}(t)$$

Lemma D.3 establishes the dominance of weak semantic tokens in the optimization process. Roughly speaking, it proves that the gradient signal driving the attention towards weak semantic tokens is significantly larger than both the signal for unique tokens and the interference noise terms. This overwhelming magnitude ensures that the attention distribution is primarily updated to focus on these high-frequency tokens.

Theorem D.4 (Restatement of Theorem 4.4)**.** *Under Assumption 4.1, if the Transformer is trained on dataset* $\mathcal{D}_2$ *with training rate* $\eta = O(1)$*, then there exist a* $t \le \frac{N}{4\eta}$*, such that*

$$\frac{\sum_{l\in\mathcal{C}} \exp(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,l})}{\sum_{j=1}^{T} \exp(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j})} \ge 1 - \frac{1}{T}$$

Lemma D.5 (Restatement of Lemma 4.5)**.** *Let* $\alpha_{i,j,k}(t) = [\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_k$ *and* $\beta_{i,j,k}(t) = [\boldsymbol{0}_{1\times d}, \boldsymbol{x}_{i,j}^\top]\boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_k$ *denote the projection scores associated with the residual and attention pathways, respectively. We distinguish their behavior based on the token type of* $\boldsymbol{x}_{i,j}$*:*

1. ***Weak Semantic Token Uniformity:*** *If $\boldsymbol{x}_{i,j}$ is a weak semantic token ($j \in \mathcal{C}$), the projection magnitudes are consistent across all classes:*

$$\alpha_{i,j,y_k}(t) = \alpha_{i,j,y_{k'}}(t),\ \beta_{i,j,y_k}(t) = \beta_{i,j,y_{k'}}(t) \tag{23}$$

   $\forall k, k' \in [N]$. This indicates that the embedding $\boldsymbol{w}$ does not favor any specific class label.

2. ***Unique Semantic Token Dominance:*** *For $t \geq 1$, if $\boldsymbol{x}_{i,j}$ is a unique semantic token associated with class y_i ($j \notin \mathcal{C}$), the projection exhibits strong dominance for the target class:*

$$\alpha_{i,j,y_i}(t) > \max_{k \neq y_i} \alpha_{i,j,k}(t),\ \beta_{i,j,y_i}(t) > \max_{k \neq y_i} \beta_{i,j,k}(t). \tag{24}$$

Lemma D.5 is directly derived from Lemma D.1.

E Complete Proof

This section provides a detailed proof of the theorems and lemmas mentioned above.

E.1 Proof of Lemma B.1

Lemma E.1 (Restatement of Lemma B.1)**.** *Under Assumption 4.1, given an input sequence $\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T\}$, the internal representation for the i-th token ($\boldsymbol{x}_i$) is as follows*

$$\boldsymbol{h}_i = \begin{bmatrix} \boldsymbol{x}_i \\ \sum_{j=1}^{i} \frac{1}{i} \boldsymbol{x}_j \end{bmatrix}$$

Proof. Recall that

$$\boldsymbol{H}^{(0)} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]^\top \in \mathbb{R}^{T \times d}.$$

$$\boldsymbol{O}^{(1)} = \mathcal{S}\left(\boldsymbol{H}^{(0)} \boldsymbol{W}_{QK}^{(1)} \boldsymbol{H}^{(0)\top} + \boldsymbol{M}\right) \boldsymbol{H}^{(0)} \boldsymbol{W}_{V}^{(1)} \in \mathbb{R}^{T \times d}$$

where $M_{ij} = 0$ if $i \geq j$ and $-\infty$ otherwise. By the Assumption 4.1 where $\boldsymbol{W}_{QK}^{(1)} = \boldsymbol{0}$ and the properties of the softmax function, we have:

$$\mathcal{S}\left(\boldsymbol{H}^{(0)} \boldsymbol{W}_{QK}^{(1)} \boldsymbol{H}^{(0)\top} + \boldsymbol{M}\right) = \mathcal{S}\left(\boldsymbol{0} + \boldsymbol{M}\right) = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{T} & \frac{1}{T} & \frac{1}{T} & \ldots & \frac{1}{T} \end{bmatrix} \in \mathbb{R}^{T \times T},$$

that is, $\mathcal{S}\left(\boldsymbol{H}^{(0)} \boldsymbol{W}_{QK}^{(1)} \boldsymbol{H}^{(0)\top} + \boldsymbol{M}\right)_{ij} = \frac{1}{i}$ if $i \geq j$ and 0 otherwise. By the Assumption 4.1 where $\boldsymbol{W}_{V}^{(1)} = \boldsymbol{I}_d$ and $\boldsymbol{H}^{(0)} = [\boldsymbol{x}_1, \ldots, \boldsymbol{x}_T]^\top$, we have

$$\boldsymbol{O}^{(1)} = \begin{bmatrix} 1 & 0 & 0 & \ldots & 0 \\ \frac{1}{2} & \frac{1}{2} & 0 & \ldots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{T} & \frac{1}{T} & \frac{1}{T} & \ldots & \frac{1}{T} \end{bmatrix} \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \boldsymbol{x}_T^\top \end{bmatrix} = \begin{bmatrix} \boldsymbol{x}_1^\top \\ \vdots \\ \sum_{j=1}^{i} \frac{1}{i} \boldsymbol{x}_j^\top \\ \vdots \\ \sum_{j=1}^{T} \frac{1}{T} \boldsymbol{x}_j^\top \end{bmatrix}$$

that is, the i-th row of $\boldsymbol{O}^{(1)}$ is $\sum_{j=1}^{i} \frac{1}{i} \boldsymbol{x}_j^\top$. By the definition that $\boldsymbol{H}^{(1)} = [\boldsymbol{h}_1, \ldots, \boldsymbol{h}_T]^\top = \operatorname{concat}\left([\boldsymbol{H}^{(0)}, \boldsymbol{O}^{(1)}]\right)$, the internal representation for the i-th token is $\boldsymbol{h}_j = \begin{bmatrix} \boldsymbol{x}_i \\ \sum_{j=1}^{i} \frac{1}{i} \boldsymbol{x}_j \end{bmatrix}$. □

E.2 Proof of Lemma B.2

Lemma E.2 (Restatement of Lemma B.2)**.** *Assume that the Transformer is trained on dataset $\mathcal{D}$ with N samples $\{\boldsymbol{x}_{i,1}, \cdots, \boldsymbol{x}_{i,T}; \boldsymbol{x}_{i,T+1}\}_{i=1}^N$, where $\boldsymbol{x}_{i,j}$ denotes the j-th token of the i-th sequence. The gradients for the parameters of the Transformer's second block are as follows.*

$$\nabla_{\boldsymbol{W}_{QK}^{(2)}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N} \begin{bmatrix} \boldsymbol{x}_{i,T} \\ \frac{1}{T}\sum_{j=1}^{T} \boldsymbol{x}_{i,j} \end{bmatrix} \boldsymbol{g}_i^\top \boldsymbol{W}_V^{(2)\top} \boldsymbol{H}_i^{(1)\top} \left(diag(\boldsymbol{a}_i) - \boldsymbol{a}_i \boldsymbol{a}_i^\top\right) \boldsymbol{H}_i^{(1)}$$

$$\nabla_{\boldsymbol{W}_{V}^{(2)}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = -\frac{1}{N}\sum_{i=1}^{N} \boldsymbol{H}_i^{(1)\top} \boldsymbol{a}_i \boldsymbol{g}_i^\top$$

where $\boldsymbol{a}_i = \mathcal{S}\left(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)} \boldsymbol{H}_i^{(1)\top}\right)^\top$ denotes the attention vector for the i-th sequence, $\boldsymbol{h}_{i,j}^\top$ denotes the internal representations for the j-th token of the i-th sample. $\boldsymbol{g}_i = \boldsymbol{x}_{i,T+1} - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T+1})$ denotes the loss derivative for the i-th sequence.

Proof. Denote a vector $\boldsymbol{v} \in \mathbb{R}^d$ and an one-hot vector $\boldsymbol{e}_y$ with the y-th entry is 1, and the others are 0. We have

$$\ell(\boldsymbol{v}, \boldsymbol{e}_y) = -\log\left(\mathcal{S}(\boldsymbol{v}^\top)\boldsymbol{e}_y\right) = -\log\left(\frac{\exp(v_y)}{\sum_{i=1}^d \exp(v_i)}\right) = \log\left(\sum_{i=1}^d \exp(v_i)\right) - v_y$$

Then we have

$$\frac{\partial \ell(\boldsymbol{v}, \boldsymbol{e}_y)}{\partial v_y} = \frac{\exp(v_y)}{\sum_{i=1}^d \exp(v_i)} - 1$$
$$\frac{\partial \ell(\boldsymbol{v}, \boldsymbol{e}_y)}{\partial v_j} = \frac{\exp(v_j)}{\sum_{i=1}^d \exp(v_i)}$$

for $j \neq y$. Thus $\nabla_{\boldsymbol{v}} \ell(\boldsymbol{v}, \boldsymbol{e}_y) = \mathcal{S}(\boldsymbol{v}) - \boldsymbol{e}_y$.

For the loss of i-th sequence $\ell(\hat{\boldsymbol{x}}_{i,T+1}, \boldsymbol{x}_{i,T+1})$, we have $\nabla_{\hat{\boldsymbol{x}}_{i,T+1}} \ell(\hat{\boldsymbol{x}}_{i,T+1}, \boldsymbol{x}_{i,T+1}) = \mathcal{S}(\hat{\boldsymbol{x}}_{i,T+1}) - \boldsymbol{e}_y = \mathcal{S}(\hat{\boldsymbol{x}}_{i,T+1}) - \boldsymbol{x}_{i,T+1} = -\boldsymbol{g}_i$.

We can also derive that $\frac{\partial \mathcal{S}(\boldsymbol{v})}{\partial \boldsymbol{v}} = diag(\boldsymbol{v}) - \boldsymbol{v}\boldsymbol{v}^\top$.

Recall the output equation $\hat{\boldsymbol{x}}_{T+1}^\top = \mathcal{S}\left(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)} \boldsymbol{H}^{(1)\top}\right) \boldsymbol{H}^{(1)} \boldsymbol{W}_V^{(2)}$, and by the chain rule, we can easily calulate the gradient for $\boldsymbol{W}_{QK}^{(2)}$ and $\boldsymbol{W}_V^{(2)}$. □

E.3 Proof of Lemma C.1

Lemma E.3 (Restatement of Lemma C.1)**.** *We denote $\alpha_i(t) = [\boldsymbol{x}_i^\top, \boldsymbol{0}_{1\times d}] \boldsymbol{W}_V^{(2)}(t) \boldsymbol{e}_y$, $\beta_i(t) = [\boldsymbol{0}_{1\times d}, \boldsymbol{x}_i^\top] \boldsymbol{W}_V^{(2)}(t) \boldsymbol{e}_y$, $\gamma_i(t) = \boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(t) \begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{0}_{d\times 1} \end{bmatrix}$ and $\rho_i(t) = \boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(t) \begin{bmatrix} \boldsymbol{0}_{d\times 1} \\ \boldsymbol{x}_i \end{bmatrix}$. Then their update rules are as follows*

$$\alpha_i(t+1) = \alpha_i(t) + \eta a_i(t) \left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right)$$

$$\beta_i(t+1) = \beta_i(t) + \eta \sum_{j=i}^{T} \frac{1}{j} a_j(t) \left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right)$$

$$\gamma_i(t+1) = \gamma_i(t) + \eta \|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1} \left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right) \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} a_i(t) (\boldsymbol{e}_i - \boldsymbol{a}(t))$$

$$\rho_i(t+1) = \rho_i(t) + \eta \|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1} \left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right) \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} \sum_{j=i}^{T} \frac{1}{j} a_j(t) (\boldsymbol{e}_j - \boldsymbol{a}(t))$$

where $\boldsymbol{a} = [a_1, \ldots, a_T]^\top$ and a_i is the attention score on the i-th token. Assuming that the category for this sequence is y, and thus $\boldsymbol{x}_{T+1} = \boldsymbol{e}_y$. $\mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y$ denotes the y-th element of $\mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))$.

Proof. By Lemma B.2, and consider only one sequence, we have

$$\nabla_{\boldsymbol{W}_V^{(2)}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = -\boldsymbol{H}^{(1)\top} \boldsymbol{a} \boldsymbol{g}^\top$$

where $\boldsymbol{H}^{(1)} = \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_i & \dots & \boldsymbol{x}_T \\ \boldsymbol{x}_1 & \dots & \sum_{j=1}^{i} \frac{1}{i} \boldsymbol{x}_j & \dots & \sum_{j=1}^{T} \frac{1}{T} \boldsymbol{x}_j \end{bmatrix}^\top$, $\boldsymbol{g} = \boldsymbol{x}_{T+1} - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1})$, and $\boldsymbol{a} = [a_1, \dots, a_T]^\top$ where a_i is the attention score on the i-th token.

Assuming that the category for this sequence is y, and thus $\boldsymbol{x}_{T+1} = \boldsymbol{e}_y$. By the properties of softmax function (bounded by $(0, 1)$, and sum up to 1), we notice that

- The y-th entry of $\boldsymbol{g} = \boldsymbol{e}_y - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1})$ is positive, and the others are negative.
- The entries of $\boldsymbol{g}$ sum up to 0.

Moreover, if all elements of $\hat{\boldsymbol{x}}_{T+1}$ excluding the y-th element are equal, it follows that all elements of $\boldsymbol{g}$ excluding the y-th element are also equal. Because the gradient of $\boldsymbol{W}_V^{(2)}$ ends with $\boldsymbol{g}$, this symmetry is preserved in subsequent iterations. Given that $\boldsymbol{W}_V^{(2)}$ is initialized as a zero matrix, we can conclude by induction that this property holds throughout the entire training procedure.

We denote $\alpha_i(t) = [\boldsymbol{x}_i^\top, \boldsymbol{0}_{1\times d}] \boldsymbol{W}_V^{(2)}(t) \boldsymbol{e}_y$ and $\beta_i(t) = [\boldsymbol{0}_{1\times d}, \boldsymbol{x}_i^\top] \boldsymbol{W}_V^{(2)}(t) \boldsymbol{e}_y$, then we have $[\boldsymbol{x}_i^\top, \boldsymbol{0}_{1\times d}] \boldsymbol{W}_V^{(2)}(t) \boldsymbol{e}_j = -\frac{\alpha_i(t)}{d-1}$ and $[\boldsymbol{0}_{1\times d}, \boldsymbol{x}_i^\top] \boldsymbol{W}_V^{(2)}(t) \boldsymbol{e}_j = -\frac{\beta_i(t)}{d-1}$ for $j \neq y$. Moreover, we have $\hat{\boldsymbol{x}}_{T+1}^\top(t) \boldsymbol{e}_i = -\frac{1}{d-1} \hat{\boldsymbol{x}}_{T+1}^\top(t) \boldsymbol{e}_y$ for $i \neq y$.

By the definition $\alpha_i(t) = [\boldsymbol{x}_i^\top, \boldsymbol{0}_{1\times d}] \boldsymbol{W}_V^{(2)}(t) \boldsymbol{e}_y$, we have

$$\begin{aligned}
\alpha_i(t+1) &= \alpha_i(t) - \eta [\boldsymbol{x}_i^\top, \boldsymbol{0}_{1\times d}] \nabla_{\boldsymbol{W}_V^{(2)}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}(t)) \boldsymbol{e}_y \\
&= \alpha_i(t) + \eta [\boldsymbol{x}_i^\top, \boldsymbol{0}_{1\times d}] \boldsymbol{H}^{(1)\top} \boldsymbol{a}(t) \boldsymbol{g}^\top(t) \boldsymbol{e}_y \\
&= \alpha_i(t) + \eta [\boldsymbol{x}_i^\top, \boldsymbol{0}_{1\times d}] \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_i & \dots & \boldsymbol{x}_T \\ \boldsymbol{x}_1 & \dots & \sum_{j=1}^{i} \frac{1}{i} \boldsymbol{x}_j & \dots & \sum_{j=1}^{T} \frac{1}{T} \boldsymbol{x}_j \end{bmatrix} \\
&\quad \boldsymbol{a}(t) \left(\boldsymbol{e}_y - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t)) \right)^\top \boldsymbol{e}_y \\
&= \alpha_i(t) + \eta a_i(t) \left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y \right).
\end{aligned}$$

Similarly, we have

$$\begin{aligned}
\beta_i(t+1) &= \beta_i(t) - \eta [\boldsymbol{0}_{1\times d}, \boldsymbol{x}_i^\top] \nabla_{\boldsymbol{W}_V^{(2)}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}(t)) \boldsymbol{e}_y \\
&= \beta_i(t) + \eta [\boldsymbol{0}_{1\times d}, \boldsymbol{x}_i^\top] \boldsymbol{H}^{(1)\top} \boldsymbol{a}(t) \boldsymbol{g}^\top(t) \boldsymbol{e}_y \\
&= \beta_i(t) + \eta [\boldsymbol{0}_{1\times d}, \boldsymbol{x}_i^\top] \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_i & \dots & \boldsymbol{x}_T \\ \boldsymbol{x}_1 & \dots & \sum_{j=1}^{i} \frac{1}{i} \boldsymbol{x}_j & \dots & \sum_{j=1}^{T} \frac{1}{T} \boldsymbol{x}_j \end{bmatrix} \\
&\quad \boldsymbol{a}(t) \left(\boldsymbol{e}_y - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t)) \right)^\top \boldsymbol{e}_y \\
&= \beta_i(t) + \eta \sum_{j=i}^{T} \frac{1}{j} a_j(t) \left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y \right).
\end{aligned} \tag{25}$$

By Lemma B.2, and consider only one sequence, we have

$$\nabla_{\boldsymbol{W}_{QK}^{(2)}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = \begin{bmatrix} \boldsymbol{x}_T \\ \frac{1}{T} \sum_{j=1}^{T} \boldsymbol{x}_j \end{bmatrix} \boldsymbol{g}^\top \boldsymbol{W}_V^{(2)\top} \boldsymbol{H}^{(1)\top} \left(diag(\boldsymbol{a}) - \boldsymbol{a}\boldsymbol{a}^\top \right) \boldsymbol{H}^{(1)}$$

By the definition $\gamma_i(t) = \boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(t) \begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{0}_{d\times 1} \end{bmatrix}$, $\rho_i(t) = \boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(t) \begin{bmatrix} \boldsymbol{0}_{d\times 1} \\ \boldsymbol{x}_i \end{bmatrix}$, we have

$$\begin{aligned}
\gamma_i(t+1) &= \gamma_i(t) - \eta \boldsymbol{h}_T^\top \nabla_{\boldsymbol{W}_{QK}^{(2)}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}(t)) \begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{0}_{d\times 1} \end{bmatrix} \\
&= \gamma_i(t) + \eta \|\boldsymbol{h}_T\|_2^2 \boldsymbol{g}^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} \left(diag(\boldsymbol{a}(t)) - \boldsymbol{a}(t)\boldsymbol{a}^\top(t) \right) \boldsymbol{H}^{(1)} \begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{0}_{d\times 1} \end{bmatrix} \\
&= \gamma_i(t) + \eta \|\boldsymbol{h}_T\|_2^2 \boldsymbol{g}^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} \left(diag(\boldsymbol{a}(t)) - \boldsymbol{a}(t)\boldsymbol{a}^\top(t) \right) \boldsymbol{e}_i \\
&= \gamma_i(t) + \eta \|\boldsymbol{h}_T\|_2^2 \left(\boldsymbol{e}_y - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t)) \right)^\top \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} a_i(t) (\boldsymbol{e}_i - \boldsymbol{a}(t)).
\end{aligned} \tag{26}$$

By the properties that $[\boldsymbol{x}_i^\top, \boldsymbol{0}_{1\times d}] \boldsymbol{W}_V^{(2)}(t) \boldsymbol{e}_j = -\frac{\alpha_i(t)}{d-1}$, $[\boldsymbol{0}_{1\times d}, \boldsymbol{x}_i^\top] \boldsymbol{W}_V^{(2)}(t) \boldsymbol{e}_j = -\frac{\beta_i(t)}{d-1}$ for $j \neq y$. and $\hat{\boldsymbol{x}}_{T+1}^\top(t) \boldsymbol{e}_i = -\frac{1}{d-1} \hat{\boldsymbol{x}}_{T+1}^\top(t) \boldsymbol{e}_y$ for $i \neq y$. We have

$$\begin{aligned}
&\left(\boldsymbol{e}_y - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t)) \right)^\top \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} \\
&= (1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y) \left(1 + (d-1) \cdot (\frac{1}{d-1})^2 \right) \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} \\
&= \frac{d}{d-1} (1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y) \, \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top}
\end{aligned}$$

Therefore, we have

$$\begin{aligned}
\gamma_i(t+1) &= \gamma_i(t) + \eta \|\boldsymbol{h}_T\|_2^2 \left(\boldsymbol{e}_y - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t)) \right)^\top \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} a_i(t) (\boldsymbol{e}_i - \boldsymbol{a}(t)) \\
&= \gamma_i(t) + \eta \|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1} (1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y) \, \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} a_i(t) (\boldsymbol{e}_i - \boldsymbol{a}(t))
\end{aligned} \tag{27}$$

Similarly,

$$\begin{aligned}
\rho_i(t+1) &= \rho_i(t) - \eta \boldsymbol{h}_T^\top \nabla_{\boldsymbol{W}_{QK}^{(2)}} \mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}(t)) \begin{bmatrix} \boldsymbol{0}_{d\times 1} \\ \boldsymbol{x}_i \end{bmatrix} \\
&= \rho_i(t) + \eta \|\boldsymbol{h}_T\|_2^2 \boldsymbol{g}^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} \left(diag(\boldsymbol{a}(t)) - \boldsymbol{a}(t)\boldsymbol{a}^\top(t) \right) \boldsymbol{H}^{(1)} \begin{bmatrix} \boldsymbol{0}_{d\times 1} \\ \boldsymbol{x}_i \end{bmatrix} \\
&= \rho_i(t) + \eta \|\boldsymbol{h}_T\|_2^2 \boldsymbol{g}^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} \left(diag(\boldsymbol{a}(t)) - \boldsymbol{a}(t)\boldsymbol{a}^\top(t) \right) \sum_{j=i}^{T} \frac{1}{j} \boldsymbol{e}_j \\
&= \rho_i(t) + \eta \|\boldsymbol{h}_T\|_2^2 \boldsymbol{g}^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} \sum_{j=i}^{T} \frac{1}{j} a_j(t) (\boldsymbol{e}_j - \boldsymbol{a}(t)) \\
&= \rho_i(t) + \eta \|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1} (1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y) \, \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}^{(1)\top} \sum_{j=i}^{T} \frac{1}{j} a_j(t) (\boldsymbol{e}_j - \boldsymbol{a}(t))
\end{aligned} \tag{28}$$

□

E.4 Proof of Lemma C.2

Lemma E.4 (Restatement of Lemma C.2)**.** *Under the Assumption 4.1, with training rate $\eta = \Theta(T \ln d)$, the attention on the first token at the second step satisfies:*

$$\frac{\exp(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2) \boldsymbol{h}_1)}{\sum_{i=1}^{T} \exp(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2) \boldsymbol{h}_i)} \geq 1 - \frac{1}{d}.$$

Moreover, we have

$$\alpha_1(t) \geq \alpha_i(t), \beta_1(t) \geq \beta_i(t), \gamma_1(t) \geq \gamma_i(t), \rho_1(t) \geq \rho_i(t)$$

for $t \in [2], i \in [2, T]$.

Proof. **First Step.**

By the zero initialization assumption $\boldsymbol{W}_V^{(2)}(0) = \mathbf{0}$ and $\boldsymbol{W}_{QK}^{(2)}(0) = \mathbf{0}$, we have:

- all the tokens share the same attention score ($a_i = \frac{1}{T}$ for $i \in [T]$)
- the output is zero, thus $\mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y = \frac{1}{d}$
- the first step of gradient for QK is zero ($\boldsymbol{W}_{QK}^{(2)}(1) = \mathbf{0}$, $\gamma_i(1) = \rho_i(1) = 0$ for $i \in [T]$)

For V, we have

$$\alpha_i(1) = \alpha_i(0) + \eta a_i(0)\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right) = \frac{\eta}{T} \cdot \left(1 - \frac{1}{d}\right),$$

$$\beta_i(1) = \beta_i(0) + \eta \sum_{j=i}^{T} \frac{1}{j} a_j(t)\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right) = \frac{\eta}{T}\left(1 - \frac{1}{d}\right) \sum_{j=i}^{T} \frac{1}{j}$$

for $i \in [T]$. By $a_i(0) = a_i(1) = \frac{1}{T}$ for $i \in [T]$, it is clear that $\alpha_i(1) = \alpha_j(1)$, $\beta_i(1) \geq \beta_j(1)$, $\alpha_i(2) = \alpha_j(2)$, $\beta_i(2) \geq \beta_j(2)$ for $i \leq j$.

We use $\alpha_i(1)$ and $\beta_i(1)$ to bound the output:

$$\begin{aligned}
\hat{\boldsymbol{x}}_{T+1}^\top(1)\boldsymbol{e}_y &= \mathcal{S}\left(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(1)\boldsymbol{H}^{(1)\top}\right)\boldsymbol{H}^{(1)}\boldsymbol{W}_V^{(2)}(1)\boldsymbol{e}_y \\
&= \boldsymbol{a}^\top(1)\begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_i & \dots & \boldsymbol{x}_T \\ \boldsymbol{x}_1 & \dots & \sum_{j=1}^{i} \frac{1}{i}\boldsymbol{x}_j & \dots & \sum_{j=1}^{T} \frac{1}{T}\boldsymbol{x}_j \end{bmatrix}^\top \boldsymbol{W}_V^{(2)}(1)\boldsymbol{e}_y \\
&= \begin{bmatrix} a_1(1)\boldsymbol{x}_1 & \dots & a_i(1)\boldsymbol{x}_i & \dots & a_T(1)\boldsymbol{x}_T \\ a_1(1)\boldsymbol{x}_1 & \dots & \sum_{j=1}^{i} \frac{1}{i}a_i(1)\boldsymbol{x}_j & \dots & \sum_{j=1}^{T} \frac{1}{T}a_T(1)\boldsymbol{x}_j \end{bmatrix}^\top \boldsymbol{W}_V^{(2)}(1)\boldsymbol{e}_y \\
&= \sum_{i=1}^{T} a_i(1)\alpha_i(1) + \sum_{i=1}^{T} \beta_i(1) \sum_{j=i}^{T} \frac{1}{j} a_j(1) \\
&= \sum_{i=1}^{T} \frac{\eta}{T^2} \cdot \left(1 - \frac{1}{d}\right) + \sum_{i=1}^{T} \left(\sum_{j=i}^{T} \frac{1}{j}\right)^2 \cdot \frac{\eta}{T^2}\left(1 - \frac{1}{d}\right) \\
&\leq \frac{3\eta}{T}\left(1 - \frac{1}{d}\right),
\end{aligned} \tag{29}$$

where the last inequality is by $\sum_{i=1}^{T}\left(\sum_{j=i}^{T} \frac{1}{j}\right)^2 \leq 2T$.

we can upper bound $\mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(1))_y$ as follows

$$\begin{aligned}
\mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(1))_y &= \frac{\exp(\hat{\boldsymbol{x}}_{T+1}^\top(1)\boldsymbol{e}_y)}{\sum_{i=1}^{d}\exp(\hat{\boldsymbol{x}}_{T+1}^\top(1)\boldsymbol{e}_i)} \\
&= \frac{\exp(\hat{\boldsymbol{x}}_{T+1}^\top(1)\boldsymbol{e}_y)}{(d-1)\exp(-\frac{\hat{\boldsymbol{x}}_{T+1}^\top(t)\boldsymbol{e}_y}{d-1}) + \exp(\hat{\boldsymbol{x}}_{T+1}^\top(1)\boldsymbol{e}_y)} \\
&= \frac{\exp(\frac{d}{d-1}\hat{\boldsymbol{x}}_{T+1}^\top(1)\boldsymbol{e}_y)}{(d-1) + \exp(\frac{d}{d-1}\hat{\boldsymbol{x}}_{T+1}^\top(1)\boldsymbol{e}_y)} \\
&\leq \frac{\exp(\frac{d}{d-1}\frac{3\eta}{T}(1 - \frac{1}{d}))}{(d-1) + \exp(\frac{d}{d-1}\frac{3\eta}{T}(1 - \frac{1}{d}))} \\
&= \frac{\exp(\frac{3\eta}{T})}{(d-1) + \exp(\frac{3\eta}{T})}
\end{aligned} \tag{30}$$

where the inequality is by Eq.(29).

Second Step. For the second step of GD, we focus on analysing the dynamics of QK. Recall that

$$\gamma_i(t+1) = \gamma_i(t) + \eta\|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1}\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right) \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top} a_i(t)(\boldsymbol{e}_i - \boldsymbol{a}(t))$$

We have

$$\gamma_1(2) - \gamma_i(2) = \frac{\eta}{T}\|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1}\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(1))_y\right) \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(1)\top}(t)\boldsymbol{H}^{(1)\top}(\boldsymbol{e}_1 - \boldsymbol{e}_i)$$

To bound $\gamma_1(2) - \gamma_i(2)$, we next calculate $\boldsymbol{e}_y^\top \boldsymbol{W}_V^{(1)\top}(1)\boldsymbol{H}^{(1)\top}\boldsymbol{e}_i$.

$$\begin{aligned}
\boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(1)\boldsymbol{H}^{(1)\top}\boldsymbol{e}_i &= \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(1) \begin{bmatrix} \boldsymbol{x}_1 & \dots & \boldsymbol{x}_i & \dots & \boldsymbol{x}_T \\ \boldsymbol{x}_1 & \dots & \sum_{j=1}^i \frac{1}{i}\boldsymbol{x}_j & \dots & \sum_{j=1}^T \frac{1}{T}\boldsymbol{x}_j \end{bmatrix} \boldsymbol{e}_i \\
&= \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(1) \begin{bmatrix} \boldsymbol{x}_i \\ \sum_{j=1}^i \frac{1}{i}\boldsymbol{x}_j \end{bmatrix} \\
&= \alpha_i(1) + \sum_{j=1}^i \frac{1}{i}\beta_j(1)
\end{aligned} \tag{31}$$

Thus,

$$\begin{aligned}
\boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(1)\boldsymbol{H}^{(1)\top}(\boldsymbol{e}_1 - \boldsymbol{e}_i) &= \alpha_1(1) + \beta_1(1) - \alpha_i(1) - \sum_{j=1}^i \frac{1}{i}\beta_j(1) \\
&= (1 - \frac{1}{i})\beta_1(1) - \sum_{j=2}^i \frac{1}{i}\beta_j(1) \\
&\geq (1 - \frac{1}{i})\left(\beta_1(1) - \max_{j\in[2,i]} \beta_j(1)\right) \\
&\geq \frac{1}{2}\cdot\frac{\eta}{T}\left(1 - \frac{1}{d}\right),
\end{aligned} \tag{32}$$

where the last inequality is by $\beta_i(1) = \frac{\eta}{T}\left(1 - \frac{1}{d}\right)\sum_{j=i}^T \frac{1}{j}$ and thus $\beta_1(1) - \beta_i(1) \geq \frac{\eta}{T}\left(1 - \frac{1}{d}\right)$ for $i \in [2, T]$.

Therefore,

$$\begin{aligned}
\gamma_1(2) - \gamma_i(2) &= \frac{\eta}{T}\|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1}\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(1))_y\right) \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top}(\boldsymbol{e}_1 - \boldsymbol{e}_i) \\
&\geq \frac{\eta}{T}\|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1}\cdot\frac{d-1}{(d-1) + \exp(\frac{3\eta}{T})}\cdot\frac{1}{2}\cdot\frac{\eta}{T}\left(1 - \frac{1}{d}\right) \\
&= \frac{\eta^2\|\boldsymbol{h}_T\|_2^2}{T^2}\cdot\frac{d-1}{(d-1) + \exp(\frac{3\eta}{T})}
\end{aligned} \tag{33}$$

Recall that

$$\rho_i(t+1) = \rho_i(t) + \eta\|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1}\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right) \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top} \sum_{j=i}^T \frac{1}{j} a_j(t)(\boldsymbol{e}_j - \boldsymbol{a}(t))$$

we have

$$\begin{aligned}
\rho_1(2) - \rho_i(2) &= \eta\|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1}\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(1))_y\right) \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(1)\boldsymbol{H}^{(1)\top} \sum_{j=1}^{i-1} \frac{1}{j} a_j(1)(\boldsymbol{e}_j - \boldsymbol{a}(1)) \\
&= \frac{\eta}{T}\|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1}\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(1))_y\right) \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(1)\boldsymbol{H}^{(1)\top} \sum_{j=1}^{i-1} \frac{1}{j}(\boldsymbol{e}_j - \boldsymbol{a}(1))
\end{aligned} \tag{34}$$

Let $v_j(1) = \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(1)\boldsymbol{H}^{(1)\top}\boldsymbol{e}_j$. Recall that $\boldsymbol{e}_y^\top \boldsymbol{W}_V^{(1)\top}(1)\boldsymbol{H}^{(1)\top}\boldsymbol{e}_i = \alpha_i(1) + \sum_{j=1}^{i} \frac{1}{i}\beta_j(1)$ and $\alpha_i(1) = \alpha_j(1)$, $\beta_i(1) \geq \beta_j(1)$ for $i \leq j$, we established that $v_j(1)$ is monotonically decreasing with respect to j. Consider the target term $\mathcal{G} = \sum_{j=1}^{i-1} \frac{1}{j}(v_j(1) - \bar{v}(1))$. We know that $\sum_{j=1}^{T}(v_j(1) - \bar{v}(1)) = 0$. Let P be the set of indices where $v_j(1) > \bar{v}(1)$ (the "head") and N be the set where $v_j(1) < \bar{v}(1)$ (the "tail"). Due to monotonicity, indices in P are smaller than indices in N. The term $\mathcal{G}$ assigns larger weights $\frac{1}{j}$ to the positive deviations in the head, and smaller weights to the negative deviations in the tail.Specifically,

$$\mathcal{G} \geq \sum_{j \in P} \frac{1}{j}(v_j(1) - \bar{v}(1)) - \sum_{k \in N, k<i} \frac{1}{k}|v_k(1) - \bar{v}(1)|.$$

Since the total positive mass equals the total negative mass, and $\frac{1}{j} > \frac{1}{k}$ for all $j \in P, k \in N$, the weighted sum is strictly positive.Thus,

$$\boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(1)\boldsymbol{H}^{(1)\top} \sum_{j=1}^{i-1} \frac{1}{j}(\boldsymbol{e}_j - \boldsymbol{a}(1)) > 0.$$

This implying $\rho_1(2) \geq \rho_i(2)$ for $i \in [2, T]$.

Next, we establish the lower bound for the attention of the first token.

Recall that $\boldsymbol{h}_i = \begin{bmatrix} \boldsymbol{x}_i \\ \sum_{j=1}^{i} \frac{1}{i}\boldsymbol{x}_j \end{bmatrix}$ and $\gamma_i(t) = \boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(t) \begin{bmatrix} \boldsymbol{x}_i \\ \boldsymbol{0}_{d\times 1} \end{bmatrix}$, $\rho_i(t) = \boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(t) \begin{bmatrix} \boldsymbol{0}_{d\times 1} \\ \boldsymbol{x}_i \end{bmatrix}$. We have

$$\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_i = \gamma_i(t) + \sum_{j=1}^{i} \frac{1}{i}\rho_j(t).$$

Thus,

$$\begin{aligned} \boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_1 - \boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_i &= \gamma_1(2) + \rho_1(2) - \gamma_i(2) - \sum_{j=1}^{i} \frac{1}{i}\rho_j(2) \\ &\geq \gamma_1(t) - \gamma_i(t) \\ &\geq \frac{\eta^2 \|\boldsymbol{h}_T\|_2^2}{T^2} \cdot \frac{d-1}{(d-1) + \exp(\frac{3\eta}{T})} \end{aligned}$$

for $i \in [2, T]$. As we choose $\eta = \frac{T \ln d}{3}$, we have

$$\begin{aligned} \boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_1 - \boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_i &\geq \left(\frac{T \ln d}{3}\right)^2 \cdot \frac{\|\boldsymbol{h}_T\|_2^2}{T^2} \cdot \frac{d-1}{(d-1) + \exp(\frac{3}{T} \cdot \frac{T \ln d}{3})} \\ &= \frac{\|\boldsymbol{h}_T\|_2^2 (\ln d)^2}{9} \cdot \frac{d-1}{2d-1} \\ &\geq 2 \ln d \end{aligned}$$

for $i \in [2, T]$, where $\ln d \geq \frac{18(2d-1)}{\|\boldsymbol{h}_T\|_2^2 (d-1)}$ is enough for the last inequality.

Then we have

$$\begin{aligned}
a_1(2) &= \frac{\exp(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_1)}{\exp(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_1) + \sum_{i=2}^T \exp(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_i)} \\
&= \frac{\exp(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_1 - \max_{j\in[2,T]}\{\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_j\})}{\exp(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_1 - \max\limits_{j\in[2,T]}\{\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_j\})} \\
&\quad + \sum_{i=2}^T \exp(\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_i - \max_{j\in[2,T]}\{\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(2)\boldsymbol{h}_j\}) \\
&\geq \frac{\exp(2\ln d)}{\exp(2\ln d) + (d-1)} \\
&= \frac{d^2}{d^2+d-1} \\
&\geq 1 - \frac{1}{d}.
\end{aligned}$$

□

E.5 Proof of Lemma C.3

Lemma E.5 (Restatement of Lemma C.3)**.** *If for some t', $a_1(t) \geq \frac{1}{2}, \alpha_1(t) \geq \alpha_i(t), \beta_1(t) \geq \beta_i(t), \gamma_1(t) \geq \gamma_i(t), \rho_1(t) \geq \rho_i(t)$ for $i \in [2,T]$ hold at $t = t'$, then these inequalities hold for $t \geq t'$. Moreover, $a_1(t+1) \geq a_1(t)$.*

Proof. Let $a_1(t) \geq \frac{1}{2}, \alpha_1(t) \geq \alpha_i(t), \beta_1(t) \geq \beta_i(t), \gamma_1(t) \geq \gamma_i(t), \rho_1(t) \geq \rho_i(t)$ for $i \in [2,T]$ be the induction hypotheses. It is clear that they hold for $t = 2$, and we need to establish these properties for $t+1$.

Recall that

$$\alpha_i(t+1) = \alpha_i(t) + \eta a_i(t)\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right),$$

$$\beta_i(t+1) = \beta_i(t) + \eta \sum_{j=i}^T \frac{1}{j} a_j(t)\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right).$$

we have $\alpha_1(t+1) \geq \alpha_i(t+1)$ by $a_1(t) \geq \frac{1}{2}$, and $\beta_1(t+1) \geq \beta_i(t+1)$ because $\beta_i(t+1)$ is monotonically decreasing with respect to i. We next analyze γ, similar to Eq.(31), we have $\boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top}\boldsymbol{e}_i = \alpha_i(t) + \sum_{j=1}^i \frac{1}{i}\beta_j(t)$. Therefore,

$$\begin{aligned}
\boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top}\boldsymbol{e}_1 - \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top}\boldsymbol{e}_i &= \alpha_1(t) + \beta_1(t) - \alpha_i(t) - \sum_{j=1}^i \frac{1}{i}\beta_j(t) \\
&\geq \alpha_1(t) - \alpha_i(t) + (1-\frac{1}{i})\beta_1(t) - \sum_{j=2}^i \frac{1}{i}\beta_j(t) \\
&\geq (1-\frac{1}{i})\left(\beta_1(t) - \max_{j\in[2,i]}\beta_j(t)\right) \\
&\geq 0,
\end{aligned} \tag{35}$$

Recall that

$$\gamma_i(t+1) = \gamma_i(t) + \eta\|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1}\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y\right)\boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top} a_i(t)(\boldsymbol{e}_i - \boldsymbol{a}(t))$$

If $\boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top}(\boldsymbol{e}_i - \boldsymbol{a}(t)) \geq 0$:

$$
\begin{aligned}
&\gamma_1(t+1) - \gamma_i(t+1) \\
&= \gamma_1(t) - \gamma_i(t) + \eta\|\boldsymbol{h}_T\|_2^2 \\
&\quad (\boldsymbol{e}_y - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t)))^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top} a_1(t)(\boldsymbol{e}_1 - \boldsymbol{a}(t)) \\
&- \eta\|\boldsymbol{h}_T\|_2^2 (\boldsymbol{e}_y - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t)))^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top} a_i(t)(\boldsymbol{e}_i - \boldsymbol{a}(t)) \\
&\geq \gamma_1(t) - \gamma_i(t) + \eta\|\boldsymbol{h}_T\|_2^2 \\
&\quad (\boldsymbol{e}_y - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t)))^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top} a_1(t) ((\boldsymbol{e}_1 - \boldsymbol{a}(t)) - (\boldsymbol{e}_i - \boldsymbol{a}(t))) \\
&= \gamma_1(t) - \gamma_i(t) + \eta\|\boldsymbol{h}_T\|_2^2 \\
&\quad (\boldsymbol{e}_y - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t)))^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top} a_1(t) (\boldsymbol{e}_1 - \boldsymbol{e}_i) \\
&\geq \gamma_1(t) - \gamma_i(t) + \eta a_1(t)\|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1} (1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y) \\
&\quad \underbrace{\boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top} (\boldsymbol{e}_1 - \boldsymbol{e}_i)}_{\geq 0} \\
&\geq \gamma_1(t) - \gamma_i(t),
\end{aligned}
\tag{36}
$$

where the first inequality is by $a_1 \geq a_i$, the last inequality is by Eq.(32).

Similarly, if $\boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top}(\boldsymbol{e}_i - \boldsymbol{a}(t)) \leq 0$:

$$
\begin{aligned}
&\gamma_1(t+1) - \gamma_i(t+1) \\
&= \gamma_1(t) - \gamma_i(t) + \eta\|\boldsymbol{h}_T\|_2^2 \\
&\quad (\boldsymbol{e}_y - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t)))^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top} a_1(t)(\boldsymbol{e}_1 - \boldsymbol{a}(t)) \\
&- \eta\|\boldsymbol{h}_T\|_2^2 (\boldsymbol{e}_y - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t)))^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top} a_i(t)(\boldsymbol{e}_i - \boldsymbol{a}(t)) \\
&\geq \gamma_1(t) - \gamma_i(t) + \eta\|\boldsymbol{h}_T\|_2^2 \\
&\quad (\boldsymbol{e}_y - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t)))^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top} a_1(t)(\boldsymbol{e}_1 - \boldsymbol{a}(t)) \\
&= \gamma_1(t) - \gamma_i(t) + \eta a_1(t)\|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1} (1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y) \\
&\quad \underbrace{\boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top} \left(\boldsymbol{e}_1 - \sum_{j=1}^{T} a_j(t)\boldsymbol{e}_j\right)}_{\geq 0} \\
&\geq \gamma_1(t) - \gamma_i(t),
\end{aligned}
\tag{37}
$$

where the last inequality is by Eq.(32) and $\sum_{j=1}^{T} a_j(t) = 1$. Therefore, $\gamma_1(t+1) \geq \gamma_i(t+1)$.

Recall that

$$
\rho_i(t+1) = \rho_i(t) + \eta\|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1} (1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y)\, \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top} \sum_{j=i}^{T} \frac{1}{j} a_j(t)(\boldsymbol{e}_j - \boldsymbol{a}(t))
$$

We have

$$
\begin{aligned}
\rho_i(t+1) - \rho_i(t+1) = \rho_i(t) - \rho_i(t) + \eta\|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1} (1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y) \\
\boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top} \sum_{j=1}^{i-1} \frac{1}{j} a_j(t)(\boldsymbol{e}_j - \boldsymbol{a}(t))
\end{aligned}
$$

and define some notations as follows

- $C(t) = \eta\|\boldsymbol{h}_T\|_2^2 \frac{d}{d-1} (1 - \mathcal{S}(\hat{\boldsymbol{x}}_{T+1}(t))_y)$. Note that $C(t) \geq 0$.

- $v_j(t) = \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top}\boldsymbol{e}_j$, $\bar{v}(t) = \sum_{k=1}^T a_k(t)v_k(t) = \boldsymbol{e}_y^\top \boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}^{(1)\top}\boldsymbol{a}(t)$.
- $u_j(t) = a_j(t)(v_j(t) - \bar{v}(t))$. Note that $\sum_{k=1}^T u_k(t) = \sum_{k=1}^T a_k(t)(v_k(t) - \bar{v}(t)) = \sum_{k=1}^T a_k(t)v_k(t) - \sum_{k=1}^T a_k(t)\bar{v}(t) = \bar{v}(t) - \bar{v}(t) = 0$.
- $\mathcal{S}_{i-1}(t) = \sum_{j=1}^{i-1} \frac{1}{j} u_j(t)$

Then we have

$$\rho_1(t+1) - \rho_i(t+1) = \rho_1(t) - \rho_i(t) + C(t)\mathcal{S}_{i-1}(t)$$

Recall the update rule for the difference between the first token and the i-th token:where $C(t) \geq 0$ is a non-negative scalar, and the weighted gradient sum is defined as:

$$\mathcal{S}_{i-1}(t) = \sum_{j=1}^{i-1} \frac{1}{j} u_j(t)$$

Let $K = \{1, \ldots, T\}$. We partition the indices into a positive contribution set $\mathcal{P}$ and a negative contribution set $\mathcal{N}$ based on the sign of $u_k(t)$:

$$\mathcal{P} = \{k \in K \mid u_k(t) > 0\}, \quad \mathcal{N} = \{k \in K \mid u_k(t) < 0\}.$$

Since $v_1(t)$ is the maximum, $v_1(t) > \bar{v}(t)$, implying $u_1(t) > 0$, so $1 \in \mathcal{P}$. Let $\mathcal{P}_{rest} = \mathcal{P} \setminus \{1\}$.

The zero-mean property $\sum_{k=1}^T u_k(t) = 0$ implies mass conservation between positive and negative terms:

$$u_1(t) + \sum_{k \in \mathcal{P}_{rest}} u_k(t) = \sum_{k \in \mathcal{N}} |u_k(t)|. \tag{38}$$

We claim that under the condition $a_1(t) \geq 1/2$, the contribution of the first token $u_1(t)$ strictly exceeds the sum of all other positive contributions.

For any $k \in \mathcal{P}_{rest}$, we have $0 < v_k(t) - \bar{v}(t) < v_1(t) - \bar{v}(t)$ due to the maximality of $v_1(t)$. Thus:

$$\sum_{k \in \mathcal{P}_{rest}} u_k(t) = \sum_{k \in \mathcal{P}_{rest}} a_k(t)(v_k(t) - \bar{v}(t)) < (v_1(t) - \bar{v}(t)) \sum_{k \in \mathcal{P}_{rest}} a_k(t).$$

Using the condition $a_1(t) \geq 1/2$, the remaining probability mass is bounded:

$$\sum_{k \in \mathcal{P}_{rest}} a_k(t) \leq \sum_{k=2}^T a_k(t) = 1 - a_1(t) \leq a_1(t).$$

Substituting this back, we obtain:

$$\sum_{k \in \mathcal{P}_{rest}} u_k(t) < (v_1(t) - \bar{v}(t)) \cdot a_1(t) = u_1(t). \tag{39}$$

We now bound $\mathcal{S}_{i-1}(t)$. We separate the term for $j = 1$ and apply the coefficient inequality $\frac{1}{j} \leq \frac{1}{2}$ for all $j \geq 2$:

$$\begin{aligned}
\mathcal{S}_{i-1}(t) &= 1 \cdot u_1(t) + \sum_{j=2}^{i-1} \frac{1}{j} u_j(t) \\
&= u_1(t) + \sum_{j=2, j \in \mathcal{P}}^{i-1} \frac{1}{j} u_j(t) - \sum_{j=2, j \in \mathcal{N}}^{i-1} \frac{1}{j} |u_j(t)| \\
&\geq u_1(t) - \sum_{j=2, j \in \mathcal{N}}^{i-1} \frac{1}{j} |u_j(t)| \\
&\geq u_1(t) - \frac{1}{2} \sum_{k \in \mathcal{N}} |u_k(t)|.
\end{aligned}$$

Substituting Eq.(38) into the negative sum:

$$\mathcal{S}_{i-1}(t) \geq u_1(t) - \frac{1}{2}\left(u_1(t) + \sum_{k\in\mathcal{P}_{rest}} u_k(t)\right) = \frac{1}{2}\left(u_1(t) - \sum_{k\in\mathcal{P}_{rest}} u_k(t)\right).$$

Finally, by Eq.(39), we have

$$\mathcal{S}_{i-1}(t) > 0.$$

Therefore,

$$\rho_1(t+1) - \rho_i(t+1) = \rho_1(t) - \rho_i(t) + C(t)\mathcal{S}_{i-1}(t) \geq 0.$$

Moreover, we have

$$\boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_1 - \boldsymbol{h}_T^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_i = \gamma_1(t) + \rho_1(t) - \gamma_i(t) - \sum_{j=1}^{i}\frac{1}{i}\rho_j(t) \geq 0$$

By induction, we have $a_1(t) \geq \frac{1}{2}, \alpha_1(t) \geq \alpha_i(t), \beta_1(t) \geq \beta_i(t), \gamma_1(t) \geq \gamma_i(t), \rho_1(t) \geq \rho_i(t)$ for $i \in [2, T], t \geq 2$, and $a_1(t)$ is non-decreasing. Specifically, as $a_1(2) \geq 1 - \frac{1}{d}$, we have $a_1(t) \geq 1 - \frac{1}{d}$ for $t \geq 2$.

□

E.6 Proof of Lemma D.1

Lemma E.6 (Restatement of Lemma D.1)**.** *We denote* $\alpha_{i,j,k}(t) = [\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_k$ $\beta_{i,j,k}(t) = [\boldsymbol{0}_{1\times d}, \boldsymbol{x}_{i,j}^\top]\boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_k$, $\gamma_{i,j}(t) = \boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\begin{bmatrix}\boldsymbol{x}_{i,j}\\ \boldsymbol{0}_{d\times 1}\end{bmatrix}$ *and* $\rho_{i,j}(t) = \boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\begin{bmatrix}\boldsymbol{0}_{d\times 1}\\ \boldsymbol{x}_{i,j}\end{bmatrix}$. *Then their update rules are as follows*

$$\alpha_{i,j,k}(t+1) = \alpha_{i,j,k}(t) + \frac{\eta}{N}a_{i,j}(t)\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i}\right)$$

$$\beta_{i,j,k}(t+1) = \beta_{i,j,k}(t) + \frac{\eta}{N}\sum_{l=j}^{T'}\frac{1}{l}a_{i,l}(t)\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i}\right)$$

for $i \in [N], j \notin \mathcal{C}, k = y_i$.

$$\alpha_{i,j,k}(t+1) = \alpha_{i,j,k}(t) - \frac{\eta}{N}a_{i,j}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k$$

$$\beta_{i,j,k}(t+1) = \beta_{i,j,k}(t) - \frac{\eta}{N}\sum_{l=j}^{T'}\frac{1}{l}a_{i,l}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k$$

for $i \in [N], j \notin \mathcal{C}, k \neq y_i$.

$$\alpha_{i,j,k}(t+1) = \alpha_{i,j,k}(t) + \frac{\eta}{N}\left(\sum_{j'\in\mathcal{C}} a_{i',j'}(t) - \sum_{m=1}^{N}\sum_{j'\in\mathcal{C}} a_{m,j'}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_{y_{i'}}\right)$$

$$\beta_{i,j,k}(t+1) = \beta_{i,j,k}(t) + \frac{\eta}{N}\left(\sum_{j'\in\mathcal{C}}\sum_{l=j'}^{T'}\frac{1}{l}a_{i',l}(t) - \sum_{m=1}^{N}\sum_{j'\in\mathcal{C}}\sum_{l=j'}^{T'}\frac{1}{l}a_{m,l}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_{y_{i'}}\right)$$

for $i \in [N], j \in \mathcal{C}, k = y_{i'}$ *for some* $i' \in [N]$.

$$\alpha_{i,j,k}(t+1) = \alpha_{i,j,k}(t) - \frac{\eta}{N}\sum_{m=1}^{N}\sum_{j'\in\mathcal{C}} a_{m,j'}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_k$$

$$\beta_{i,j,k}(t+1) = \beta_{i,j,k}(t) - \frac{\eta}{N}\sum_{m=1}^{N}\sum_{j'\in\mathcal{C}}\sum_{l=j'}^{T'}\frac{1}{l}a_{m,l}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_k$$

for $i \in [N], j \in \mathcal{C}, k \neq y_{i'}$ *for* $\forall i' \in [N]$.

$$\gamma_{i,j}(t+1) = \gamma_{i,j}(t) + \frac{\eta}{N}\|\boldsymbol{h}_{i,T'}\|_2^2 \boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}a_{i,j}(t)(\boldsymbol{e}_j - \boldsymbol{a}_i(t))$$

$$\rho_{i,j}(t+1) = \rho_{i,j}(t) + \frac{\eta}{N}\|\boldsymbol{h}_{i,T'}\|_2^2 \boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{k=j}^{T'}\frac{1}{k}a_{i,k}(t)(\boldsymbol{e}_k - \boldsymbol{a}_i(t))$$

for $i \in [N], j \notin \mathcal{C}$.

$$\gamma_{i,j}(t+1) = \gamma_{i,j}(t) + \frac{\eta}{N}\left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2}\right) \boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)(\boldsymbol{e}_{j'} - \boldsymbol{a}_i(t))$$

$$\rho_{i,j}(t+1) = \rho_{i,j}(t) + \frac{\eta}{N}\left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2}\right) \boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}}\sum_{k=j'}^{T'}\frac{1}{k}a_{i,k}(t)(\boldsymbol{e}_k - \boldsymbol{a}_i(t))$$

for $i \in [N], j \in \mathcal{C}$.

where $\boldsymbol{a}_i = [a_{i,1}, \ldots, a_{i,T'}]^\top$ *and* $a_{i,j}$ *is the attention score on the j-th token in the i-th sequence. Assuming that the category for the i-th sequence is* y_i*, and thus* $\boldsymbol{x}_{i,T'+1} = \boldsymbol{e}_{y_i}$*.* $\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_i$ *denotes the i-th element of* $\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))$.

Proof. Recall that $\nabla_{\boldsymbol{W}_V^{(2)}}\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}) = -\sum_{i=1}^N \boldsymbol{H}_i^{(1)\top}\boldsymbol{a}_i\boldsymbol{g}_i^\top$, we have

$$\begin{aligned}
&[\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t+1)\\
&= [\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t) - \eta[\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\nabla_{\boldsymbol{W}_V^{(2)}}\mathcal{L}_{\mathcal{D}}(\boldsymbol{\theta}(t))\\
&= [\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t) + \frac{\eta}{N}[\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\sum_{i'=1}^N \boldsymbol{H}_{i'}^{(1)\top}\boldsymbol{a}_{i'}(\boldsymbol{e}_{y_{i'}} - \mathcal{S}(\hat{\boldsymbol{x}}_{i',T'+1}(t)))^\top
\end{aligned}$$

For $j \notin \mathcal{C}$, note that $\boldsymbol{x}_{i,j}$ only appears in $\boldsymbol{H}_i^{(1)}$, we have

$$\begin{aligned}
[\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t+1) &= [\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t) + \frac{\eta}{N}[\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\boldsymbol{H}_i^{(1)\top}\boldsymbol{a}_i(\boldsymbol{e}_{y_i} - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t)))^\top\\
&= [\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t) + \frac{\eta}{N}a_{i,j}(\boldsymbol{e}_{y_i} - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t)))^\top
\end{aligned}$$

Then we have

$$\begin{aligned}
\alpha_{i,j,k}(t+1) &= [\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t+1)\boldsymbol{e}_k\\
&= [\boldsymbol{x}_{i,j}^\top, \boldsymbol{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_k + \frac{\eta}{N}a_{i,j}(t)\left(\boldsymbol{e}_{y_i} - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))\right)^\top\boldsymbol{e}_k\\
&= \begin{cases}\alpha_{i,j,k}(t) + \frac{\eta}{N}a_{i,j}(t)\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i}\right) & \text{for } k = y_i,\\ \alpha_{i,j,k}(t) - \frac{\eta}{N}a_{i,j}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k & \text{for } k \neq y_i.\end{cases}
\end{aligned}$$

Similarly,

$$\begin{aligned}
\beta_{i,j,k}(t+1) &= [\boldsymbol{0}_{1\times d}, \boldsymbol{x}_{i,j}^\top]\boldsymbol{W}_V^{(2)}(t+1)\boldsymbol{e}_k\\
&= [\boldsymbol{0}_{1\times d}, \boldsymbol{x}_{i,j}^\top]\boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_k + \frac{\eta}{N}\sum_{l=j}^{T'}\frac{1}{l}a_{i,l}(t)\left(\boldsymbol{e}_{y_i} - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))\right)^\top\boldsymbol{e}_k\\
&= \begin{cases}\beta_{i,j,k}(t) + \frac{\eta}{N}\sum_{l=j}^{T'}\frac{1}{l}a_{i,l}(t)\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i}\right) & \text{for } k = y_i,\\ \beta_{i,j,k}(t) - \frac{\eta}{N}\sum_{l=j}^{T'}\frac{1}{l}a_{i,l}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k & \text{for } k \neq y_i.\end{cases}
\end{aligned}$$

For $j \in \mathcal{C}$, note that $\boldsymbol{x}_{i,j}$ appears in every $\boldsymbol{H}_{i'}^{(1)}$ for many times, we have

$$
\begin{aligned}
&[\boldsymbol{x}_{i,j}^\top, \mathbf{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t+1) \\
&= [\boldsymbol{x}_{i,j}^\top, \mathbf{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t) + \frac{\eta}{N}[\boldsymbol{x}_{i,j}^\top, \mathbf{0}_{1\times d}] \sum_{i'=1}^{N} \boldsymbol{H}_{i'}^{(1)\top} \boldsymbol{a}_{i'}(\boldsymbol{e}_{y_{i'}} - \mathcal{S}(\hat{\boldsymbol{x}}_{i',T'+1}(t)))^\top \\
&= [\boldsymbol{x}_{i,j}^\top, \mathbf{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t) + \frac{\eta}{N} \sum_{m=1}^{N} \sum_{j'\in\mathcal{C}} a_{m,j'}(\boldsymbol{e}_{y_m} - \mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t)))^\top
\end{aligned}
$$

Then we have

$$
\begin{aligned}
\alpha_{i,j,k}(t+1) &= [\boldsymbol{x}_{i,j}^\top, \mathbf{0}_{1\times d}]\boldsymbol{W}_V^{(2)}(t+1)\boldsymbol{e}_k \\
&= \alpha_{i,j,k}(t) + \frac{\eta}{N} \sum_{m=1}^{N} \sum_{j'\in\mathcal{C}} a_{m,j'}(\boldsymbol{e}_{y_m} - \mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t)))^\top \boldsymbol{e}_k \\
&= \begin{cases}
\alpha_{i,j,k}(t) + \dfrac{\eta}{N} \left(\begin{array}{l} \displaystyle\sum_{j'\in\mathcal{C}} a_{i',j'}(t) \\ \displaystyle -\sum_{m=1}^{N} \sum_{j'\in\mathcal{C}} a_{m,j'}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_{y_{i'}} \end{array} \right) \\
\qquad k = y_{i'} \text{ for some } i' \in [N], \\
\alpha_{i,j,k}(t) - \dfrac{\eta}{N} \displaystyle\sum_{m=1}^{N} \sum_{j'\in\mathcal{C}} a_{m,j'}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_k \\
\qquad k \neq y_{i'} \text{ for } \forall i' \in [N].
\end{cases}
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\beta_{i,j,k}(t+1) &= [\mathbf{0}_{1\times d}, \boldsymbol{x}_{i,j}^\top]\boldsymbol{W}_V^{(2)}(t+1)\boldsymbol{e}_k \\
&= [\mathbf{0}_{1\times d}, \boldsymbol{x}_{i,j}^\top]\boldsymbol{W}_V^{(2)}(t)\boldsymbol{e}_k \\
&\quad + \frac{\eta}{N} \sum_{m=1}^{N} \sum_{j'\in\mathcal{C}} \sum_{l=j'}^{T'} \frac{1}{l} a_{m,l}(t) \left(\boldsymbol{e}_{y_i} - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))\right)^\top \boldsymbol{e}_k \\
&= \begin{cases}
\beta_{i,j,k}(t) + \dfrac{\eta}{N} \left(\begin{array}{l} \displaystyle\sum_{j'\in\mathcal{C}} \sum_{l=j'}^{T'} \frac{1}{l} a_{i',l}(t) \\ \displaystyle -\sum_{m=1}^{N} \sum_{j'\in\mathcal{C}} \sum_{l=j'}^{T'} \frac{1}{l} a_{m,l}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_{y_{i'}} \end{array} \right) \\
\qquad k = y_{i'} \text{ for some } i' \in [N], \\
\beta_{i,j,k}(t) - \dfrac{\eta}{N} \displaystyle\sum_{m=1}^{N} \sum_{j'\in\mathcal{C}} \sum_{l=j'}^{T'} \frac{1}{l} a_{m,l}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_k \\
\qquad k \neq y_{i'} \text{ for } \forall i' \in [N].
\end{cases}
\end{aligned}
$$

For $i \in [N], j \notin \mathcal{C}$, the caluculation is similar to that of distinct-only dataset ($\mathcal{D}_{\text{pure}}$):

$$
\begin{aligned}
\gamma_{i,j}(t+1) &= \gamma_{i,j}(t) + \frac{\eta}{N}\|\boldsymbol{h}_{i,T'}\|_2^2 \\
&\qquad \boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top} a_{i,j}(t)(\boldsymbol{e}_j - \boldsymbol{a}_i(t)) \\
\rho_{i,j}(t+1) &= \rho_{i,j}(t) + \frac{\eta}{N}\|\boldsymbol{h}_{i,T'}\|_2^2 \\
&\qquad \boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top} \sum_{k=j}^{T} \frac{1}{k} a_{i,k}(t)(\boldsymbol{e}_k - \boldsymbol{a}_i(t))
\end{aligned}
$$

For $i \in [N], j \in \mathcal{C}$, note that each $\boldsymbol{h}_{i,T'}$ share a same direction of $[\mathbf{0}_{1\times d}, \frac{C}{T'}\boldsymbol{x}_{i,j}^\top]^\top$ because each sequence has C weak semantic tokens.

$$
\begin{aligned}
\gamma_{i,j}(t+1) &= \gamma_{i,j}(t) \\
&\quad + \left[\boldsymbol{x}_{i,T'}^\top, \frac{1}{T'}\sum_{j'=1}^{T'}\boldsymbol{x}_{i,j'}^\top\right]\sum_{i'=1}^{N}\begin{bmatrix}\boldsymbol{x}_{i',T'} \\ \frac{1}{T'}\sum_{j'=1}^{T'}\boldsymbol{x}_{i',j'}\end{bmatrix} \\
&\qquad \boldsymbol{g}_{i'}^\top \boldsymbol{W}_V^{(2)\top}\boldsymbol{H}_{i'}^{(1)\top}\left(diag(\boldsymbol{a}_{i'}) - \boldsymbol{a}_{i'}\boldsymbol{a}_{i'}^\top\right)\boldsymbol{H}_{i'}^{(1)}\begin{bmatrix}\boldsymbol{x}_{i,j} \\ \mathbf{0}_{d\times 1}\end{bmatrix} \\
&= \gamma_{i,j}(t) + \frac{\eta}{N}\|\boldsymbol{h}_{i,T'}\|_2^2 \boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}} a_{i,j'}(t)(\boldsymbol{e}_{j'} - \boldsymbol{a}_i(t)) \\
&\quad + \frac{\eta}{N}\left[\mathbf{0}_{d\times 1}, \frac{C}{T'}\boldsymbol{x}_{i,j}^\top\right]\sum_{i'\neq i}^{N}\begin{bmatrix}\boldsymbol{x}_{i',T'} \\ \frac{1}{T'}\sum_{j'=1}^{T'}\boldsymbol{x}_{i',j'}\end{bmatrix} \\
&\qquad \boldsymbol{g}_{i'}^\top \boldsymbol{W}_V^{(2)\top}\boldsymbol{H}_{i'}^{(1)\top}\sum_{j'\in\mathcal{C}} a_{i',j'}(t)(\boldsymbol{e}_{j'} - \boldsymbol{a}_{i'}(t)) \\
&= \gamma_{i,j}(t) + \frac{\eta}{N}\left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2}\right) \\
&\qquad \boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}} a_{i,j'}(t)(\boldsymbol{e}_{j'} - \boldsymbol{a}_i(t))
\end{aligned}
$$

the last equality is due to the symmetry: $\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}} a_{i,j'}(t)(\boldsymbol{e}_{j'} - \boldsymbol{a}_i(t)) = \boldsymbol{g}_{i'}^\top \boldsymbol{W}_V^{(2)\top}\boldsymbol{H}_{i'}^{(1)\top}\sum_{j'\in\mathcal{C}} a_{i',j'}(t)(\boldsymbol{e}_{j'} - \boldsymbol{a}_{i'}(t))$. Similarly, we have

$$
\begin{aligned}
\rho_{i,j}(t+1) = \rho_{i,j}(t) + \frac{\eta}{N}\left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2}\right) \\
\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}}\sum_{k=j'}^{T}\frac{1}{k}a_{i,k}(t)(\boldsymbol{e}_k - \boldsymbol{a}_i(t))
\end{aligned}
$$

□

E.7 Proof of Lemma D.2

Lemma E.7 (Restatement of Lemma D.2)**.** *For $t \le O(\frac{N}{\eta})$, we have*

$$\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k \le \frac{2}{d}$$

for $k \in [d]$.

Proof. the output for the i-th sequence is

$$\hat{\boldsymbol{x}}_{i,T'+1}^\top = \mathcal{S}\left(\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}\boldsymbol{H}_i^{(1)\top}\right)\boldsymbol{H}_i^{(1)}\boldsymbol{W}_V^{(2)}.$$

Note that the softmax $\mathcal{S}\left(\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}\boldsymbol{H}_i^{(1)\top}\right)$ sum up to 1, and $\boldsymbol{H}_i^{(1)}\boldsymbol{W}_V^{(2)}\boldsymbol{e}_k$ is a linear combination of $\alpha_{i,j,k}$ and $\beta_{i,j,k}$ with coefficients add up to 1. So the magnitude of $\hat{\boldsymbol{x}}_{i,T'+1}^\top$ is bounded by the maximum of $|\alpha_{i,j,k}| + |\beta_{i,j,k}|$.

Under the induction hypotheses of $\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k \le \frac{2}{d}$, and thus $\sum_{m=1}^{N}\sum_{j'\in\mathcal{C}} a_{m,j'}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_k \le \frac{2N}{d} \le 1$. Then we can choose the following two equation in Lemma D.1 to derive the upper bound for α and β.

$$\alpha_{i,j,k}(t+1) = \alpha_{i,j,k}(t) + \frac{\eta}{N}\left(\sum_{j'\in\mathcal{C}} a_{i',j'}(t) - \sum_{m=1}^{N}\sum_{j'\in\mathcal{C}} a_{m,j'}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_{y_{i'}}\right)$$

for $i \in [N], j \in \mathcal{C}, k = y_{i'}$ for some $i' \in [N]$.

$$\beta_{i,j,k}(t+1) = \beta_{i,j,k}(t) + \frac{\eta}{N} \sum_{l=j}^{T'} \frac{1}{l} a_{i,l}(t) \left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i}\right)$$

for $i \in [N], j \notin \mathcal{C}, k = y_i$.

By $\sum_{j' \in \mathcal{C}} a_{i',j'}(t) \leq 1$, we have

$$|\alpha_{i,j,k}(t+1)| \leq |\alpha_{i,j,k}(t)| + \frac{\eta}{N}$$

$$|\beta_{i,j,k}(t+1)| \leq |\beta_{i,j,k}(t)| + \frac{\eta}{N}$$

Therefore,

$$|\alpha_{i,j,k}(t)|, |\beta_{i,j,k}(t)| \leq \frac{\eta t}{N}$$

Assume that each element of $\hat{\boldsymbol{x}}_{i,T'+1}^\top$ is bounded by M, i.e. $|\hat{\boldsymbol{x}}_{i,T'+1}^\top \boldsymbol{e}_k| \leq M$ for $k \in [d]$. and the softmax output can be bounded:

$$\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k \leq \frac{\exp(M)}{\exp(M) + (d-1)\exp(-M)} = \frac{1}{1 + (d-1)\exp(-2M)}$$

To ensure $\frac{1}{1+(d-1)\exp(-2M)} \leq \frac{2}{d}$, we must have $M \leq \frac{1}{2} \ln\left(\frac{2(d-1)}{d-2}\right) \leq \frac{1}{2}$, which implying $|\alpha_{i,j,k}(t)| + |\beta_{i,j,k}(t)| \leq \frac{2\eta t}{N} \leq \frac{1}{2}$. Therefore, $t \leq \frac{N}{4\eta}$ is sufficient to ensure $\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k \leq \frac{2}{d}$. □

We next bound $\boldsymbol{g}_i^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}_i^{(1)\top} \boldsymbol{e}_j$

Lemma E.8. *For $i \in [N], j \in [T'], t \leq \frac{N}{4\eta}$, we have*

$$\boldsymbol{g}_i^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}_i^{(1)\top} \boldsymbol{e}_j \leq 2\alpha_{i,j,y_i}(t) + \frac{2}{j} \sum_{j'=1}^{j} \beta_{i,j',y_i}(t)$$

$$\boldsymbol{g}_i^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}_i^{(1)\top} \boldsymbol{e}_j \geq \frac{1}{2}\alpha_{i,j,y_i}(t) + \frac{1}{2j} \sum_{j'=1}^{j} \beta_{i,j',y_i}(t)$$

Proof.

$$\begin{aligned}
&\boldsymbol{g}_i^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}_i^{(1)\top} \boldsymbol{e}_j \\
&= \left(\boldsymbol{e}_{y_i} - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))\right) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{h}_{i,j} \\
&= (1 - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i}) \alpha_{i,j,y_i}(t) - \sum_{k \neq y_i}^{d} \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k \alpha_{i,j,k}(t) \\
&+ \frac{1}{j} \sum_{j'=1}^{j} \left((1 - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i}) \beta_{i,j',y_i}(t) - \sum_{k \neq y_i}^{d} \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k \beta_{i,j',k}(t) \right)
\end{aligned}$$

for $j \notin \mathcal{C}$. comparing the following two equation in Lemma D.1:

$$\alpha_{i,j,k}(t+1) = \alpha_{i,j,k}(t) + \frac{\eta}{N} a_{i,j}(t) \left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i}\right)$$

for $k = y_i$,

$$\alpha_{i,j,k}(t+1) = \alpha_{i,j,k}(t) - \frac{\eta}{N} a_{i,j}(t) \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k$$

for $k \neq y_i$.

By $\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k \leq \frac{2}{d}$, we have $\alpha_{i,j,y_i}(t) \geq \frac{1-\frac{2}{d}}{\frac{2}{d}}|\alpha_{i,j,k}(t)| = \frac{d-2}{2}|\alpha_{i,j,k}(t)|$ for $k \neq y_i$.

Thus,

$$
\begin{aligned}
&(1-\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i})\alpha_{i,j,y_i}(t) - \sum_{k\neq y_i}^{d} \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k \alpha_{i,j,k}(t) \\
&\geq (1-\frac{2}{d})\alpha_{i,j,y_i}(t) \\
&\geq \frac{1}{2}\alpha_{i,j,y_i}(t)
\end{aligned}
$$

and

$$
\begin{aligned}
&(1-\mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i})\alpha_{i,j,y_i}(t) - \sum_{k\neq y_i}^{d} \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_k \alpha_{i,j,k}(t) \\
&\leq \alpha_{i,j,y_i}(t) + d\cdot\frac{2}{d}\cdot\frac{2}{d-2}\alpha_{i,j,y_i}(t) \\
&\leq 2\alpha_{i,j,y_i}(t)
\end{aligned}
$$

The similar properties also apply to other α and β, then we have

$$
\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\boldsymbol{e}_j \leq 2\alpha_{i,j,y_i}(t) + \frac{2}{j}\sum_{j'=1}^{j}\beta_{i,j',y_i}(t)
$$

$$
\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\boldsymbol{e}_j \geq \frac{1}{2}\alpha_{i,j,y_i}(t) + \frac{1}{2j}\sum_{j'=1}^{j}\beta_{i,j',y_i}(t)
$$

□

E.8 Proof of Lemma D.3

Lemma E.9 (Restatement of Lemma D.3)**.** *For $i \in [N], j \in \mathcal{C}, j' \notin \mathcal{C}, t \leq \frac{N}{4\eta}$, if $\big(1-\sum_{l\in\mathcal{C}} a_{i,l}(t)\big) \geq \frac{1}{2T}$, we have*

$$
\begin{aligned}
\gamma_{i,j}(t) &\geq T|\gamma_{i,j'}(t)| \\
\gamma_{i,j}(t) &\geq \frac{1}{16}|\rho_{i,j}(t)| \\
\gamma_{i,j}(t) &\geq T|\rho_{i,j'}(t)| \\
\gamma_{i,j}(t+1) &\geq \gamma_{i,j}(t) + \frac{\eta}{5N}\left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2}\right)\sum_{j'\in\mathcal{C}} a_{i,j'}(t)\left(1-\sum_{l\in\mathcal{C}} a_{i,l}(t)\right)\alpha_{i,j,y_i}(t)
\end{aligned}
$$

Proof. We begin with the induction hypothesis that $a_{i,j} \geq a_{i',j'}$ for $i, i' \in [N], j \in \mathcal{C}, j' \notin \mathcal{C}$. It is clear that this hypothesis holds at $t = 0$. We then bound $\alpha_{i,j,y_i}(t)$ and $\beta_{i,j,y_i}(t)$. By Lemma D.1, for $j \in \mathcal{C}$, we have

$$
\begin{aligned}
\alpha_{i,j,y_i}(t+1) &= \alpha_{i,j,y_i}(t) + \frac{\eta}{N}\left(\sum_{j'\in\mathcal{C}} a_{i,j'}(t) - \sum_{m=1}^{N}\sum_{j'\in\mathcal{C}} a_{m,j'}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_{y_i}\right) \\
&\geq \alpha_{i,j,y_i}(t) + \frac{\eta}{N}\sum_{j'\in\mathcal{C}} a_{i,j'}(t)\left(1-\sum_{m=1}^{N}\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_{y_i}\right) \\
&\geq \alpha_{i,j,y_i}(t) + \frac{\eta}{2N}\sum_{j'\in\mathcal{C}} a_{i,j'}(t)
\end{aligned} \tag{40}
$$

where the first inequality is by the symmetry that $a_{i,j}(t) = a_{m,j}(t)$ for $j \in \mathcal{C}$, the last inequality is by $\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_{y_i} \leq \frac{2}{d}$ and thus $N \cdot \mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_{y_i} \leq \frac{2N}{d} = O(\frac{1}{T'}) \leq \frac{1}{2}$.

We also have

$$\begin{aligned}\beta_{i,j,y_i}(t+1) &= \beta_{i,j,y_i}(t) + \frac{\eta}{N}\left(\sum_{j'\in\mathcal{C}}\sum_{l=j}^{T'}\frac{1}{l}a_{i,l}(t) - \sum_{m=1}^{N}\sum_{j'\in\mathcal{C}}\sum_{l=j}^{T'}\frac{1}{l}a_{m,l}(t)\mathcal{S}(\hat{\boldsymbol{x}}_{m,T'+1}(t))_{y_i}\right)\\ &\leq \beta_{i,j,y_i}(t) + \frac{\eta}{N}\sum_{j'\in\mathcal{C}}\sum_{l=j}^{T'}\frac{1}{l}a_{i,l}(t)\end{aligned}$$

for $j \in \mathcal{C}$. By $\mathcal{C} = \{T+1, 2(T+1), \cdots, C(T+1)\}$, we have

$$\begin{aligned}\sum_{j\in\mathcal{C}}\sum_{l=j}^{T'}\frac{1}{l}a_{i,l}(t) &= \sum_{j=1}^{C}\sum_{l=j(T+1)}^{C}\frac{1}{l}a_{i,l}(t)\\ &= \sum_{l=T+1}^{T'}\left(\sum_{j=1}^{\lfloor\frac{l}{T+1}\rfloor}1\right)\frac{1}{l}a_{i,l}(t)\\ &= \sum_{l=T+1}^{T'}\lfloor\frac{l}{T+1}\rfloor\frac{1}{l}a_{i,l}(t)\\ &\leq \sum_{l=T+1}^{T'}\left(\frac{l}{T+1}\right)\frac{1}{l}a_{i,l}(t)\\ &= \frac{1}{T+1}\sum_{l=T+1}^{T'}a_{i,l}(t)\\ &< \frac{1}{T+1}\end{aligned}$$

Therefore,

$$\begin{aligned}\beta_{i,j,y_i}(t+1) &\leq \beta_{i,j,y_i}(t) + \frac{\eta}{N}\cdot\frac{1}{T+1}\\ &\leq \beta_{i,j,y_i}(t) + \frac{\eta}{N}\frac{T'}{C(T+1)}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)\\ &\leq \beta_{i,j,y_i}(t) + \frac{2\eta}{N}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)\end{aligned} \tag{41}$$

for $j \in \mathcal{C}$. The second inequality is by the induction hypothesis that $a_{i,k} \geq a_{i',k'}$ for $i, i' \in [N], k \in \mathcal{C}, k' \notin \mathcal{C}$, and thus $\sum_{j'\in\mathcal{C}} a_{i,j'}(t) \geq \frac{C}{T'}$. The last inequality is due to $T' = C(T+1) + T$ and thus $\frac{T'}{C(T+1)} \leq 2$.

For $j \notin \mathcal{C}$, by Lemma D.1, we have

$$\begin{aligned}\alpha_{i,j,y_i}(t+1) &= \alpha_{i,j,y_i}(t) + \frac{\eta}{N}a_{i,j}(t)\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i}\right)\\ &\leq \alpha_{i,j,y_i}(t) + \frac{\eta}{N}a_{i,j}(t)\\ &\leq \alpha_{i,j,y_i}(t) + \frac{\eta}{N}\cdot\frac{\sum_{j'\in\mathcal{C}}a_{i,j'}(t)}{C}\end{aligned} \tag{42}$$

where the last inequality is by the induction hypothesis that $a_{i,k} \geq a_{i',k'}$ for $i, i' \in [N], k \in \mathcal{C}, k' \notin \mathcal{C}$.

$$\begin{aligned}
\beta_{i,j,y_i}(t+1) &= \beta_{i,j,y_i}(t) + \frac{\eta}{N}\sum_{l=j}^{T'}\frac{1}{l}a_{i,l}(t)\left(1 - \mathcal{S}(\hat{\boldsymbol{x}}_{i,T'+1}(t))_{y_i}\right) \\
&\leq \beta_{i,j,y_i}(t) + \frac{\eta}{N}\sum_{l=j}^{T'}\frac{1}{l}a_{i,l}(t) \\
&= \beta_{i,j,y_i}(t) + \frac{\eta}{N}\left(\sum_{l\geq j, l\notin\mathcal{C}}^{T'}\frac{1}{l}a_{i,l}(t) + \sum_{l\geq j, l\in\mathcal{C}}^{T'}\frac{1}{l}a_{i,l}(t)\right) \\
&\leq \beta_{i,j,y_i}(t) + \frac{\eta}{N}\left(\frac{1+\ln T'}{C}\sum_{j'\in\mathcal{C}}a_{i,j'}(t) + \frac{1}{T+1}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)\right) \\
&\leq \beta_{i,j,y_i}(t) + \frac{\eta}{16N}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)
\end{aligned} \tag{43}$$

where the second inequality is by $\sum_{l=1}^{T'}\frac{1}{l} \leq 1 + \ln T'$ and $l \geq \frac{1}{T+1}$ for $l \in \mathcal{C}$. The last inequality is by choosing sufficiently large $C = \Omega(\ln T)$ and $T = \Omega(1)$.

To summarize Eq.(40), (41), (42), (43), we have

$$\alpha_{i,j,y_i}(t) \geq \frac{1}{4}\beta_{i,j,y_i}(t)$$

$$\alpha_{i,j,y_i}(t) \geq \frac{C}{2}\alpha_{i,j',y_i}(t)$$

$$\alpha_{i,j,y_i}(t) \geq 8\beta_{i,j',y_i}(t)$$

for $i \in [N], j \in \mathcal{C}, j' \notin \mathcal{C}$.

Recall Lemma E.8

$$\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\boldsymbol{e}_j \leq 2\alpha_{i,j,y_i}(t) + \frac{2}{j}\sum_{j'=1}^{j}\beta_{i,j',y_i}(t)$$

$$\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\boldsymbol{e}_j \geq \frac{1}{2}\alpha_{i,j,y_i}(t) + \frac{1}{2j}\sum_{j'=1}^{j}\beta_{i,j',y_i}(t)$$

We have

$$\begin{aligned}
&\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}(\boldsymbol{e}_j - \boldsymbol{e}_{j'}) \\
&\geq \frac{1}{2}\alpha_{i,j,y_i}(t) + \frac{1}{2j}\sum_{l=1}^{j}\beta_{i,l,y_i}(t) - 2\alpha_{i,j',y_i}(t) - \frac{2}{j'}\sum_{l=1}^{j'}\beta_{i,l,y_i}(t) \\
&\geq \frac{1}{2}\alpha_{i,j,y_i}(t) - \frac{2}{C}\cdot 2\alpha_{i,j,y_i}(t) - \frac{2}{j'}\cdot j'\cdot\left(\frac{4}{T}+\frac{1}{8}\right)\cdot\alpha_{i,j,y_i}(t) \\
&\geq \frac{1}{5}\alpha_{i,j,y_i}(t)
\end{aligned} \tag{44}$$

for $j \in \mathcal{C}, j' \notin \mathcal{C}$. The second inequality is by $\alpha_{i,j,y_i}(t) \geq \frac{1}{4}\beta_{i,j,y_i}(t)$, $\alpha_{i,j,y_i}(t) \geq \frac{C}{2}\alpha_{i,j',y_i}(t)$ and $\alpha_{i,j,y_i}(t) \geq 8\beta_{i,j',y_i}(t)$. Specifically, the ratio of $\beta_{i,j,y_i}(t), j \in \mathcal{C}$ appear in $\sum_{l=1}^{j'}\beta_{i,l,y_i}(t)$ is bounded by $\frac{1}{T+1}$, so we have $\frac{2}{j'}\sum_{l=1}^{j'}\beta_{i,l,y_i}(t) \leq \frac{2}{j'}\cdot j'\cdot\frac{4}{T}$ by $\alpha_{i,j,y_i}(t) \geq \frac{1}{4}\beta_{i,j,y_i}(t)$. The last inequality is by sufficiently large $C = \Omega(1)$ and $T = \Omega(1)$.

Similarly, for $j \in [T']$, we can upper bound $\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\boldsymbol{e}_j$ as follows

$$\begin{aligned}|\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\boldsymbol{e}_j| &\le 2\alpha_{i,j,y_i}(t) + \frac{2}{j}\sum_{j'=1}^{j}\beta_{i,j',y_i}(t)\\ &\le 2\alpha_{i,k,y_i}(t) + \frac{2}{j'}\cdot j'\cdot\left(\frac{4}{T}+\frac{1}{8}\right)\cdot\alpha_{i,k,y_i}(t)\\ &\le 3\alpha_{i,k,y_i}(t)\end{aligned} \tag{45}$$

for any $k \in \mathcal{C}$.

We will use Eq.(44), (45) to bound γ and ρ.

$$\begin{aligned}&\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)(\boldsymbol{e}_{j'}-\boldsymbol{a}_i(t))\\ &=\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)(\boldsymbol{e}_{j'}-\sum_{l=1}^{T'}a_{i,l}(t)\boldsymbol{e}_l)\\ &=\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)\left(\boldsymbol{e}_{j'}-\sum_{l\in\mathcal{C}}a_{i,l}(t)\boldsymbol{e}_l\right)\\ &\quad-\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)\sum_{l\notin\mathcal{C}}a_{i,l}(t)\boldsymbol{e}_l\\ &=\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)\left(1-\sum_{l\in\mathcal{C}}a_{i,l}(t)\right)\boldsymbol{e}_{j'}\\ &\quad-\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)\left(1-\sum_{l\in\mathcal{C}}a_{i,l}(t)\right)\boldsymbol{e}_l\\ &\ge\frac{1}{5}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)\left(1-\sum_{l\in\mathcal{C}}a_{i,l}(t)\right)\alpha_{i,j,y_i}(t)\end{aligned} \tag{46}$$

for $j \in \mathcal{C}$, where the inequality is by Eq.(44). Therefore,

$$\gamma_{i,j}(t+1) \ge \gamma_{i,j}(t) + \frac{\eta}{5N}\left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2}\right)\sum_{j'\in\mathcal{C}}a_{i,j'}(t)\left(1-\sum_{l\in\mathcal{C}}a_{i,l}(t)\right)\alpha_{i,j,y_i}(t) \tag{47}$$

for $j \in \mathcal{C}$. Similarly, we have

$$
\begin{aligned}
&\left| \boldsymbol{g}_i^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}_i^{(1)\top} \sum_{j' \in \mathcal{C}} \sum_{k=j'}^{T'} \frac{1}{k} a_{i,k}(t)(\boldsymbol{e}_k - \boldsymbol{a}_i(t)) \right| \\
&= \left| \boldsymbol{g}_i^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}_i^{(1)\top} \sum_{j' \in \mathcal{C}} \left(\begin{array}{l} \displaystyle\sum_{k \geq j', k \in \mathcal{C}}^{T'} \frac{1}{k} a_{i,k}(t)(\boldsymbol{e}_k - \boldsymbol{a}_i(t)) \\ + \displaystyle\sum_{k \geq j', k \notin \mathcal{C}}^{T'} \frac{1}{k} a_{i,k}(t)(\boldsymbol{e}_k - \boldsymbol{a}_i(t)) \end{array} \right) \right| \\
&\leq \frac{1}{T+1} \left| \boldsymbol{g}_i^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}_i^{(1)\top} \sum_{j' \in \mathcal{C}} a_{i,j'}(t)(\boldsymbol{e}_{j'} - \boldsymbol{a}_i(t)) \right| \\
&+ \frac{1}{T+1} \left| \boldsymbol{g}_i^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}_i^{(1)\top} \sum_{j' \in \mathcal{C}} \sum_{k \geq j', k \notin \mathcal{C}}^{T'} a_{i,k}(t)(\boldsymbol{e}_k - \boldsymbol{a}_i(t)) \right| \\
&\leq \frac{1}{T+1} \left| \boldsymbol{g}_i^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}_i^{(1)\top} \sum_{j' \in \mathcal{C}} a_{i,j'}(t)(\boldsymbol{e}_{j'} - \boldsymbol{a}_i(t)) \right| \\
&+ \frac{3}{T+1} \left(1 - \sum_{l \in \mathcal{C}} a_{i,l}(t) \right) \alpha_{i,j,y_i}(t)
\end{aligned} \tag{48}
$$

where the first inequality is due to $\sum_{j' \in \mathcal{C}} \frac{1}{k} \mathbb{1}(k \geq j') = \sum_{i=1}^{\lfloor \frac{k}{T+1} \rfloor} \frac{1}{k} \leq \frac{1}{T+1}$. The last inequality is by Eq.(45).

Recall that

$$
\begin{aligned}
\gamma_{i,j}(t+1) = \gamma_{i,j}(t) + \frac{\eta}{N} \left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2} \right) \\
\boldsymbol{g}_i^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}_i^{(1)\top} \sum_{j' \in \mathcal{C}} a_{i,j'}(t)(\boldsymbol{e}_{j'} - \boldsymbol{a}_i(t))
\end{aligned}
$$

$$
\begin{aligned}
\rho_{i,j}(t+1) = \rho_{i,j}(t) + \frac{\eta}{N} \left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2} \right) \\
\boldsymbol{g}_i^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}_i^{(1)\top} \sum_{j' \in \mathcal{C}} \sum_{k=j'}^{T'} \frac{1}{k} a_{i,k}(t)(\boldsymbol{e}_k - \boldsymbol{a}_i(t))
\end{aligned}
$$

for $i \in [N], j \in \mathcal{C}$. We have

$$
\begin{aligned}
&16\left(\gamma_{i,j}(t+1)-\gamma_{i,j}(t)\right)-\left|\left(\rho_{i,j}(t+1)-\rho_{i,j}(t)\right)\right| \\
&\geq \frac{16\eta}{N}\left(\|\boldsymbol{h}_{i,T'}\|_2^2+\frac{C^2(N-1)}{T'^2}\right)\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)(\boldsymbol{e}_{j'}-\boldsymbol{a}_i(t)) \\
&-\frac{1}{T+1}\cdot\frac{\eta}{N}\cdot\left(\|\boldsymbol{h}_{i,T'}\|_2^2+\frac{C^2(N-1)}{T'^2}\right)\left|\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)(\boldsymbol{e}_{j'}-\boldsymbol{a}_i(t))\right| \\
&-\frac{3}{T+1}\cdot\frac{\eta}{N}\cdot\left(\|\boldsymbol{h}_{i,T'}\|_2^2+\frac{C^2(N-1)}{T'^2}\right)\left(1-\sum_{l\in\mathcal{C}}a_{i,l}(t)\right)\alpha_{i,j,y_i}(t) \\
&\geq \frac{15(T+1)+T}{5(T+1)}\frac{\eta}{N}\left(\|\boldsymbol{h}_{i,T'}\|_2^2+\frac{C^2(N-1)}{T'^2}\right)\sum_{j'\in\mathcal{C}}a_{i,j'}(t)\left(1-\sum_{l\in\mathcal{C}}a_{i,l}(t)\right)\alpha_{i,j,y_i}(t) \\
&-\frac{3}{T+1}\cdot\frac{\eta}{N}\cdot\left(\|\boldsymbol{h}_{i,T'}\|_2^2+\frac{C^2(N-1)}{T'^2}\right)\left(1-\sum_{l\in\mathcal{C}}a_{i,l}(t)\right)\alpha_{i,j,y_i}(t) \\
&\geq 0
\end{aligned}
$$

for $i \in [N], j \in \mathcal{C}$. The first inequality is by Eq.(48), the second inequality is by (46), the last inequality is by $\frac{15(T+1)+T}{5(T+1)}\sum_{j'\in\mathcal{C}}a_{i,j'}(t) \geq \frac{15(T+1)+T}{5(T+1)}\cdot\frac{C}{C(T+1)+T} \geq \frac{3}{T+1}$ as we choose a sufficiently large $C = \Omega(1)$.

Therefore, we have

$$\gamma_{i,j}(t) \geq \frac{1}{16}\left|\rho_{i,j}(t)\right|$$

for $i \in [N], j \in \mathcal{C}$.

We next bound the following two equations:

$$
\begin{aligned}
\gamma_{i,j}(t+1) &= \gamma_{i,j}(t)+\frac{\eta}{N}\|\boldsymbol{h}_{i,T'}\|_2^2 \\
&\qquad \boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}a_{i,j}(t)(\boldsymbol{e}_j-\boldsymbol{a}_i(t)) \\
\rho_{i,j}(t+1) &= \rho_{i,j}(t)+\frac{\eta}{N}\|\boldsymbol{h}_{i,T'}\|_2^2 \\
&\qquad \boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}\sum_{k=j}^{T'}\frac{1}{k}a_{i,k}(t)(\boldsymbol{e}_k-\boldsymbol{a}_i(t))
\end{aligned}
$$

for $i \in [N], j \notin \mathcal{C}$.

Because $a_{i,j}(t) \leq \frac{\sum_{j'\in\mathcal{C}}a_{i,j'}(t)}{C}$ and $a_{i,j}(t) \leq \left(1-\sum_{l\in\mathcal{C}}a_{i,l}(t)\right)$,

if $\sum_{l\in\mathcal{C}}a_{i,l}(t) \leq \frac{1}{2}$:

$$a_{i,j}(t) \leq \frac{\sum_{j'\in\mathcal{C}}a_{i,j'}(t)}{C} \leq \frac{2}{C}\sum_{j'\in\mathcal{C}}a_{i,j'}(t)\left(1-\sum_{l\in\mathcal{C}}a_{i,l}(t)\right)$$

if $\sum_{l\in\mathcal{C}}a_{i,l}(t) \geq \frac{1}{2}$:

$$a_{i,j}(t) \leq \left(1-\sum_{l\in\mathcal{C}}a_{i,l}(t)\right) \leq 2\sum_{j'\in\mathcal{C}}a_{i,j'}(t)\left(1-\sum_{l\in\mathcal{C}}a_{i,l}(t)\right)$$

Thus, $a_{i,j}(t) \leq 2\sum_{j'\in\mathcal{C}}a_{i,j'}(t)\left(1-\sum_{l\in\mathcal{C}}a_{i,l}(t)\right)$. Then we have

$$
\begin{aligned}
|\gamma_{i,j}(t+1)| &= |\gamma_{i,j}(t)|+\left|\frac{\eta}{N}\|\boldsymbol{h}_{i,T'}\|_2^2\boldsymbol{g}_i^\top(t)\boldsymbol{W}_V^{(2)\top}(t)\boldsymbol{H}_i^{(1)\top}a_{i,j}(t)(\boldsymbol{e}_j-\boldsymbol{a}_i(t))\right| \\
&\leq |\gamma_{i,j}(t)|+\left|\frac{6\eta}{N}\|\boldsymbol{h}_{i,T'}\|_2^2\sum_{j'\in\mathcal{C}}a_{i,j'}(t)\left(1-\sum_{l\in\mathcal{C}}a_{i,l}(t)\right)\alpha_{i,k,y_i}(t)\right|
\end{aligned}
\tag{49}
$$

for any $k \in \mathcal{C}, j \notin \mathcal{C}$, where the inequality is by $a_{i,j}(t) \leq 2\sum_{j' \in \mathcal{C}} a_{i,j'}(t)\left(1 - \sum_{l \in \mathcal{C}} a_{i,l}(t)\right)$ and Eq.(45). Therefore, by Eq.(47) and $N \geq \widetilde{\Omega}(T^4)$, we have

$$\gamma_{i,j}(t) \geq T\gamma_{i,j'}(t)$$

for $j \in \mathcal{C}, j' \notin \mathcal{C}$.

Similarly, if $\left(1 - \sum_{l \in \mathcal{C}} a_{i,l}(t)\right) \geq \frac{1}{2T}$, we have

$$\begin{aligned}
\sum_{k=j}^{T'} \frac{1}{k} a_{i,k}(t) &= \sum_{k \geq j, k \in \mathcal{C}}^{T'} \frac{1}{k} a_{i,k}(t) + \sum_{k \geq j, k \notin \mathcal{C}}^{T'} \frac{1}{k} a_{i,k}(t) \\
&\leq \frac{1}{T+1} \sum_{j' \in \mathcal{C}} a_{i,j'}(t) + \left(1 - \sum_{l \in \mathcal{C}} a_{i,l}(t)\right) \\
&\leq \frac{2T}{T+1} \sum_{j' \in \mathcal{C}} a_{i,j'}(t) \left(1 - \sum_{l \in \mathcal{C}} a_{i,l}(t)\right) + 2T \sum_{j' \in \mathcal{C}} a_{i,j'}(t) \left(1 - \sum_{l \in \mathcal{C}} a_{i,l}(t)\right) \\
&\leq 3T \sum_{j' \in \mathcal{C}} a_{i,j'}(t) \left(1 - \sum_{l \in \mathcal{C}} a_{i,l}(t)\right)
\end{aligned}$$

where the first inequality is due to the index of $k \in \mathcal{C} \geq T+1$, the second inequality is by $1 \leq 2T\left(1 - \sum_{l \in \mathcal{C}} a_{i,l}(t)\right)$ and $1 \leq 2T \sum_{j' \in \mathcal{C}} a_{i,j'}(t) = \frac{2CT}{C(T+1)+T}$. Then we have

$$\begin{aligned}
|\rho_{i,j}(t+1)| &= |\rho_{i,j}(t)| + \left| \frac{\eta}{N} \|\boldsymbol{h}_{i,T'}\|_2^2 \boldsymbol{g}_i^\top(t) \boldsymbol{W}_V^{(2)\top}(t) \boldsymbol{H}_i^{(1)\top} \sum_{k=j}^{T'} \frac{1}{k} a_{i,k}(t)(\boldsymbol{e}_k - \boldsymbol{a}_i(t)) \right| \\
&\leq |\rho_{i,j}(t)| + \left| \frac{9T\eta}{N} \|\boldsymbol{h}_{i,T'}\|_2^2 \sum_{j' \in \mathcal{C}} a_{i,j'}(t) \left(1 - \sum_{l \in \mathcal{C}} a_{i,l}(t)\right) \alpha_{i,k,y_i}(t) \right|
\end{aligned} \tag{50}$$

where the inequality is by Eq. (45) and $\sum_{k=j}^{T'} \frac{1}{k} a_{i,k}(t) \leq 3T \sum_{j' \in \mathcal{C}} a_{i,j'}(t)\left(1 - \sum_{l \in \mathcal{C}} a_{i,l}(t)\right)$. Then by $N \geq \widetilde{\Omega}(T^4)$, we have

$$\gamma_{i,j}(t) \geq T\,|\rho_{i,j'}(t)|$$

for $j \in \mathcal{C}, j' \notin \mathcal{C}$.

□

E.9 Proof of Theorem D.4

Theorem E.10 (Restatement of Theorem D.4)**.** *Under Assumption 4.1, if the Transformer is trained on dataset $\mathcal{D}_2$ with training rate $\eta = O(1)$, then there exist a $t \leq \frac{N}{4\eta}$, such that*

$$\frac{\sum_{l \in \mathcal{C}} \exp(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)}(t) \boldsymbol{h}_{i,l})}{\sum_{j=1}^{T} \exp(\boldsymbol{h}_{i,T}^\top \boldsymbol{W}_{QK}^{(2)}(t) \boldsymbol{h}_{i,j})} \geq 1 - \frac{1}{T}$$

Proof. If for some $t' \leq \frac{N}{4\eta}$, we have $\sum_{l \in \mathcal{C}} a_{i,l}(t) \leq 1 - \frac{1}{2T}$ for $t \leq t'$, then the results of Lemma D.3 hold:

$$\begin{aligned}
\gamma_{i,j}(t) &\geq T|\gamma_{i,j'}(t)| \\
\gamma_{i,j}(t) &\geq \frac{1}{16}|\rho_{i,j}(t)| \\
\gamma_{i,j}(t) &\geq T|\rho_{i,j'}(t)| \\
\gamma_{i,j}(t+1) &\geq \gamma_{i,j}(t) + \frac{\eta}{5N}\left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2}\right) \sum_{j' \in \mathcal{C}} a_{i,j'}(t) \left(1 - \sum_{l \in \mathcal{C}} a_{i,l}(t)\right) \alpha_{i,j,y_i}(t)
\end{aligned}$$

for $i \in [N], j \in \mathcal{C}, j' \notin \mathcal{C}$.

Recall the definitions $\boldsymbol{h}_{i,j} = \begin{bmatrix} \boldsymbol{x}_{i,j} \\ \sum_{l=1}^{j} \frac{1}{j} \boldsymbol{x}_{i,l} \end{bmatrix}$ and $\gamma_{i,j}(t) = \boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t) \begin{bmatrix} \boldsymbol{x}_{i,j} \\ \boldsymbol{0}_{d\times 1} \end{bmatrix}$, $\rho_{i,j}(t) = \boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t) \begin{bmatrix} \boldsymbol{0}_{d\times 1} \\ \sum_{l=1}^{j} \frac{1}{j} \boldsymbol{x}_{i,l} \end{bmatrix}$. We have

$$\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t) \boldsymbol{h}_{i,j} = \gamma_{i,j}(t) + \sum_{l=1}^{j} \frac{1}{j} \rho_{i,l}(t).$$

Then for $j \in \mathcal{C}, j' \notin \mathcal{C}$, we have

$$\begin{aligned} \boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t) \boldsymbol{h}_{i,j} - \boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t) \boldsymbol{h}_{i,j'} &= \gamma_{i,j}(t) - \gamma_{i,j'}(t) + \sum_{l=1}^{j} \frac{1}{j} \rho_{i,l}(t) - \sum_{l=1}^{j'} \frac{1}{j'} \rho_{i,l}(t) \\ &\geq \gamma_{i,j}(t) - \frac{1}{T}\gamma_{i,j}(t) - \frac{1}{T}\gamma_{i,j}(t) - \frac{32}{T}\gamma_{i,j}(t) \\ &\geq \frac{1}{2}\gamma_{i,j}(t) \end{aligned} \tag{51}$$

the first inequality is by $\gamma_{i,j}(t) \geq T|\gamma_{i,j'}(t)|, \gamma_{i,j}(t) \geq \frac{1}{16}|\rho_{i,j}(t)|, \gamma_{i,j}(t) \geq T|\rho_{i,j'}(t)|$. Specifically, the ratio of $\rho_{i,l}(t), l \in \mathcal{C}$ appear in $\sum_{l=1}^{j} \frac{1}{j} \rho_{i,l}(t)$ and $\sum_{l=1}^{j'} \frac{1}{j'} \rho_{i,l}(t)$ is bounded by $\frac{1}{T}$, so we have $\sum_{l=1}^{j} \frac{1}{j} |\rho_{i,l}(t)| + \sum_{l=1}^{j'} \frac{1}{j'} |\rho_{i,l}(t)| \leq \frac{32}{T}\gamma_{i,j}(t)$. The last inequality is by choosing a sufficiently large $T = \Omega(1)$.

We next proof that there exist a $t' \leq \frac{N}{8\eta}$, such that $\sum_{j\in\mathcal{C}} a_{i,j}(t') \geq \frac{1}{2}$.

Assume that for $t \leq \frac{N}{8\eta}$, $\sum_{l\in\mathcal{C}} a_{i,l}(t) < \frac{1}{2}$, then $\left(1 - \sum_{l\in\mathcal{C}} a_{i,l}(t)\right) \geq \frac{1}{2}$, and further we have

$$\begin{aligned} \gamma_{i,j}(t+1) &\geq \gamma_{i,j}(t) + \frac{\eta}{10N} \left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2} \right) \sum_{j'\in\mathcal{C}} a_{i,j'}(t) \left(1 - \sum_{l\in\mathcal{C}} a_{i,l}(t)\right) \alpha_{i,j,y_i}(t) \\ &\geq \gamma_{i,j}(t) + \frac{\eta}{10N} \left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2} \right) \frac{C}{T'} \alpha_{i,j,y_i}(t) \\ &\geq \gamma_{i,j}(t) + \frac{\eta C^3(N-1)}{10NT'^3} \alpha_{i,j,y_i}(t) \end{aligned}$$

for $t \leq t'$.

By Eq.(40) where $\alpha_{i,j,y_i}(t+1) \geq \frac{\eta}{2N} \sum_{j'\in\mathcal{C}} a_{i,j'}(t)$, we have

$$\alpha_{i,j,y_i}(t) \geq \frac{\eta}{2N} \sum_{s=1}^{t} \sum_{j'\in\mathcal{C}} a_{i,j'}(t) \geq \frac{\eta C}{2NT'} \cdot t$$

Therefore,

$$\begin{aligned} \gamma_{i,j}(t) &\geq \gamma_{i,j}(t-1) + \frac{\eta C^3(N-1)}{10NT'^3} \alpha_{i,j,y_i}(t-1) \\ &= \gamma_{i,j}(0) + \sum_{s=0}^{t-1} \frac{\eta C^3(N-1)}{10NT'^3} \alpha_{i,j,y_i}(s) \\ &= \gamma_{i,j}(0) + \sum_{s=0}^{t-1} \frac{\eta C^3(N-1)}{10NT'^3} \cdot \frac{\eta C}{2NT'} \cdot s \\ &= \frac{\eta^2 C^4(N-1)}{40N^2T'^4} \cdot t(t-1) \end{aligned}$$

At $t = \frac{N}{8\eta}$,

$$\gamma_{i,j}(t) \geq \frac{\eta^2 C^4(N-1)}{40N^2T'^4} \cdot \frac{N}{8\eta}(\frac{N}{8\eta} - 1) = \Theta\left(\frac{C^4(N-1)}{T'^4}\right) \geq 2\ln\left(\frac{T'-C}{C}\right)$$

where the last inequality is by $N = \widetilde{\Omega}(T^4)$. Then we have

$$\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j} - \boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j'} \geq \frac{1}{2}\gamma_{i,j}(t) \geq \ln\left(\frac{T'-C}{C}\right)$$

for $j \in \mathcal{C}, j' \notin \mathcal{C}$ by Eq.(51). Therefore,

$$\frac{\sum_{j\in\mathcal{C}} \exp(\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j})}{\sum_{j'\notin\mathcal{C}} \exp(\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j'})} \geq \frac{C}{T'-C}\exp\left(\ln\left(\frac{T'-C}{C}\right)\right) \geq 1.$$

Thus,

$$\sum_{j\in\mathcal{C}} a_{i,j}(t) = \frac{\sum_{j\in\mathcal{C}} \exp(\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j})}{\sum_{j\in\mathcal{C}} \exp(\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j}) + \sum_{j'\notin\mathcal{C}} \exp(\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j'})} \geq \frac{1}{2}$$

Conflict with the assumption that $\sum_{j\in\mathcal{C}} a_{i,j}(t) \leq \frac{1}{2}$ for $t \leq \frac{N}{8\eta}$. So there exist a $t' \leq \frac{N}{8\eta}$, such that $\sum_{j'\in\mathcal{C}} a_{i,j'}(t') \geq \frac{1}{2}$.

We next proof that there exist a $t' \leq \frac{N}{4\eta}$, such that $\sum_{j\in\mathcal{C}} a_{i,j}(t') \geq 1 - \frac{1}{T}$.

Assume that for $t \leq \frac{N}{4\eta}$, $\sum_{l\in\mathcal{C}} a_{i,l}(t) < 1 - \frac{1}{T}$, then $\left(1 - \sum_{l\in\mathcal{C}} a_{i,l}(t)\right) \geq \frac{1}{T}$.

For $\frac{N}{8\eta} \leq t \leq \frac{N}{4\eta}$, we have

$$\begin{aligned}
\gamma_{i,j}(t+1) &\geq \gamma_{i,j}(t) + \frac{\eta}{10N}\left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2}\right)\sum_{j'\in\mathcal{C}} a_{i,j'}(t)\left(1 - \sum_{l\in\mathcal{C}} a_{i,l}(t)\right)\alpha_{i,j,y_i}(t) \\
&\geq \gamma_{i,j}(t) + \frac{\eta}{10N}\left(\|\boldsymbol{h}_{i,T'}\|_2^2 + \frac{C^2(N-1)}{T'^2}\right)\cdot\frac{1}{2}\cdot\frac{1}{T}\cdot\alpha_{i,j,y_i}(t) \\
&\geq \gamma_{i,j}(t) + \frac{\eta C^2(N-1)}{20NTT'^2}\alpha_{i,j,y_i}(t)
\end{aligned}$$

and

$$\alpha_{i,j,y_i}(t) \geq \frac{\eta}{2N}\sum_{s=\frac{N}{8\eta}}^{t}\sum_{j'\in\mathcal{C}} a_{i,j'}(s) \geq \frac{\eta}{4N}\cdot(s - \frac{N}{8\eta})$$

The first inequality is by Eq.(40) where $\alpha_{i,j,y_i}(t+1) \geq \frac{\eta}{2N}\sum_{j'\in\mathcal{C}} a_{i,j'}(t)$, the second inequality is by $\sum_{j'\in\mathcal{C}} a_{i,j'}(s) \geq \frac{1}{2}$ for $s \geq \frac{N}{8\eta}$.

Therefore,

$$\begin{aligned}
\gamma_{i,j}(t) &\geq \gamma_{i,j}(t-1) + \frac{\eta C^2(N-1)}{20NTT'^2}\alpha_{i,j,y_i}(t-1) \\
&\geq \gamma_{i,j}(\frac{N}{8\eta}) + \sum_{s=\frac{N}{8\eta}}^{t-1}\frac{\eta C^2(N-1)}{20NTT'^2}\alpha_{i,j,y_i}(s) \\
&\geq \sum_{s=\frac{N}{8\eta}}^{t-1}\frac{\eta C^2(N-1)}{20NTT'^2}\cdot\frac{\eta}{4N}\cdot(s - \frac{N}{8\eta}) \\
&\geq \frac{\eta^2 C^2(N-1)}{160N^2TT'^2}(t - \frac{N}{8\eta})(t - \frac{N}{8\eta} - 1)
\end{aligned}$$

At $t = \frac{N}{4\eta}$,

$$\gamma_{i,j}(t) \geq \frac{\eta^2 C^2(N-1)}{160N^2TT'^2}(t - \frac{N}{8\eta})(t - \frac{N}{8\eta} - 1) = \Theta\left(\frac{C^2(N-1)}{TT'^2}\right) \geq 2T$$

where the last inequality is by $N = \widetilde{\Omega}(T^4)$.

Then we have

$$\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j} - \boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j'} \geq \frac{1}{2}\gamma_{i,j}(t) \geq T$$

for $j \in \mathcal{C}, j' \notin \mathcal{C}$ by Eq.(51). Therefore,

$$\frac{\sum_{j\in\mathcal{C}} \exp(\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j})}{\sum_{j'\notin\mathcal{C}} \exp(\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j'})} \geq \frac{C}{T' - C}\exp(T) \geq T - 1.$$

Because the exponential function grows fast, we can choose a sufficiently large $T = \Omega(1)$ to establish the last inequality.

Then we have

$$\begin{aligned}\sum_{j\in\mathcal{C}} a_{i,j}(t) &= \frac{\sum_{j\in\mathcal{C}} \exp(\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j})}{\sum_{j\in\mathcal{C}} \exp(\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j}) + \sum_{j'\notin\mathcal{C}} \exp(\boldsymbol{h}_{i,T'}^\top \boldsymbol{W}_{QK}^{(2)}(t)\boldsymbol{h}_{i,j'})} \\ &\geq \frac{T-1}{T} \geq 1 - \frac{1}{T}\end{aligned}$$

Conflict with the assumption that $\sum_{j\in\mathcal{C}} a_{i,j}(t) \leq 1 - \frac{1}{T}$ for $t \leq \frac{N}{4\eta}$. So there exist a $t' \leq \frac{N}{4\eta}$, such that $\sum_{j'\in\mathcal{C}} a_{i,j'}(t') \geq 1 - \frac{1}{T}$. □