Non-asymptotic Convergence of Training Transformers for Next-token Prediction

Ruiquan Huang
Penn State University
State College, PA, 16801
rzh5514@psu.edu

Yingbin Liang
Ohio State University
Columbus, OH, 43210
liang.889@osu.edu

Jing Yang
Penn State Univeristy
State College, PA, 16801
yangjing@psu.edu

Abstract

Transformers have achieved extraordinary success in modern machine learning due to their excellent ability to handle sequential data, especially in next-token prediction (NTP) tasks. However, the theoretical understanding of their performance in NTP is limited, with existing studies focusing mainly on asymptotic performance. This paper provides a fine-grained non-asymptotic analysis of the training dynamics of a one-layer transformer consisting of a self-attention module followed by a feed-forward layer. We first characterize the essential structural properties of training datasets for NTP using a mathematical framework based on partial orders. Then, we design a two-stage training algorithm, where the pre-processing stage for training the feed-forward layer and the main stage for training the attention layer exhibit fast convergence performance. Specifically, both layers converge *sub-linearly* to the direction of their corresponding max-margin solutions. We also show that the cross-entropy loss enjoys a *linear* convergence rate. Furthermore, we show that the trained transformer presents non-trivial prediction ability with dataset shift, which sheds light on the remarkable generalization performance of transformers. Our analysis technique involves the development of novel properties on the attention gradient and further in-depth analysis of how these properties contribute to the convergence of the training process. Our experiments further validate our theoretical findings.

1 Introduction

The transformer architecture (Vaswani et al., 2017) has revolutionized the field of machine learning, establishing itself as a foundation model for numerous applications, including natural language processing (NLP) (Devlin et al., 2018), computer vision (Dosovitskiy et al., 2020), and multi-modal signal processing (Tsai et al., 2019). In particular, transformers achieve tremendous empirical success in large language models (LLMs) such as GPT-3 (Brown et al., 2020). Despite the empirical success, limited theoretical understanding of transformers have caused a series of critical concerns about their robustness, interpretability, and bias issues (Bommasani et al., 2021; Belkin, 2024).

To overcome these issues, recent advances in transformer theory have investigated the convergence of training transformers under theoretically amenable setting such as linear regression (Mahankali et al., 2023; Zhang et al., 2023; Huang et al., 2023) and binary classification (Tarzanagh et al., 2023b,a; Vasudeva et al., 2024; Li et al., 2023). Nevertheless, one of the fundamental task in LLMs and other generative models is next-token prediction (NTP), which involves predicting the next word or token in a sequence, given the previous tokens. In NTP, a few recent theoretical studies have started to investigate the training dynamics of transformers (Tian et al., 2023a; Li et al., 2024). However, those works lack of fine-grained *non-asymptotic* convergence analysis of the training process, posing the following open questions for further investigation:

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

How fast does the training of a transformer converge in NTP?

In addition, a pre-trained transformer empirically exhibits non-trivial generalization ability. A follow-up question from a theoretical point of view is that

Can we show the generalization capability of a trained transformer on unseen data?

In this paper, we take a first step towards addressing the aforementioned questions by studying the training dynamics of a single layer transformer consisting of a self-attention layer and a feed-forward layer for NTP. We summarize our contribution as follows.

- We develop a mathematical framework based on partial order to formally characterize the essential structural properties of the training dataset for next-token prediction. In particular, we introduce a realizable setting for training datasets where the loss can be minimized to near zero, which admits a *collocation* and *query-dependent partial orders*. A collocation is a set of token pairs where each token is directly paired with its subsequent token. Query-dependent partial orders is a set of partial orders where each partial order classifies tokens into three categories: optimal tokens, non-optimal tokens and non-comparable tokens. These structural properties define favorable max-margin problems on both the feed-forward layer and the self-attention layer.
- Second, we design a two-stage training algorithm based on normalized gradient descent. In stage 1 of pre-processing, we use the collocation to train the feed-forward layer. In stage 2, we use the entire dataset to train the self-attention layer. We show that the feed-forward layer and the query-key attention matrix converge sublinearly in direction respectively to the max-margin solution for classifying next token from all other tokens in the preprocessing dataset, and to the max-margin solution for classifying the optimal from non-optimal tokens. In addition, the norm of the transformer parameters grows linearly, which further yields a *linear* convergence rate of the cross-entropy loss. Our two-stage algorithm decouples the training of the feed-forward and attention layers without losing optimality, as stage 1's max-margin solution is judiciously designed to facilitate stage 2's fine-grained classification for optimal token prediction.
- Third, we show that the trained transformer has generalization ability for making non-trivial prediction on unseen data. In particular, the transformer is trained to learn an extended query-dependent partial order, where the non-comparable tokens are inserted in between the optimal tokens and non-optimal tokens. Thus, the trained transformer will attend to non-comparable tokens if optimal tokens are not in a new sentence and further make desirable prediction.

2 Related Work

Inspired by Brown et al. (2020), who demonstrated that pre-trained transformers can learn in-context - i.e., learn new tasks during inference with only a few samples - a series of works focus on the expressiveness power of transformers (Akyürek et al., 2022; Bai et al., 2023; Von Oswald et al., 2023; Fu et al., 2023; Giannou et al., 2023; Lin et al., 2023). These studies have shown that there exist parameter configurations such that transformers can perform various algorithms such as gradient descent. Additionally, Edelman et al. (2022) showed that transformers can represent a sparse function.

Regarding the training dynamics and optimization of transformers under in-context learning, Ahn et al. (2024); Mahankali et al. (2023); Zhang et al. (2023); Huang et al. (2023) studied the dynamics of a single attention layer, single-head transformer for the in-context learning of linear regression tasks. Cui et al. (2024) proved that multi-head attention outperforms single-head attention. Cheng et al. (2023) showed that local optimal solutions in transformers can perform gradient descent in-context for non-linear functions. Kim and Suzuki (2024) studied the nonconvex mean-field dynamics of transformers, and Nichani et al. (2024) established a convergence rate of $\tilde{O}(1/t)$ for the training loss in learning a causal graph. Additionally, Chen et al. (2024) investigated the gradient flow in training multi-head attention. Chen and Li (2024) proposed a supervised training algorithm for multi-head transformers.

Another line of research focuses on the training dynamics of transformers for binary classification problems. Tarzanagh et al. (2023b,a) demonstrated an equivalence between the optimization dynamics of a single attention layer and a certain SVM problem. While Tarzanagh et al. (2023b,a) only proved an asymptotic convergence result, Vasudeva et al. (2024) improved the convergence rate to $t^{-3/4}$. Li et al. (2023) studied the training dynamics of vision transformers and showed that the generalization error

can approach zero given sufficient training samples. Additionally, Deora et al. (2023) investigated the training and generalization error under the neural tangent kernel (NTK) regime.

For transformers trained on next-token prediction (NTP), Tian et al. (2023a) analyzed the training dynamics of a single-layer transformer, while Tian et al. (2023b) studied the joint training dynamics of multi-layer transformers. Li et al. (2024) demonstrated the asymptotic convergence of transformers trained with a logarithmic loss function for NTP. Although these works provided valuable insights into the training dynamics of transformers for NTP, they did not provide the finite-time convergence analysis, which is the focus of this paper. We remark that Thrampoulidis (2024) studied NTP without transformer structure.

Our work is also related to the classical implicit bias framework for training neural networks (NNs). In particular Soudry et al. (2018); Nacson et al. (2019); Ji and Telgarsky (2021); Ji et al. (2021) established convergence rate of gradient descent-based optimization. Phuong and Lampert (2020); Frei et al. (2022); Kou et al. (2024) studied the implicit bias of ReLU/Leaky-ReLU networks on orthogonal data. A comprehensive survey is provided in Vardi (2023). However, these works focused on classical neural networks, whereas we investigate the implicit bias of transformers for NTP.

3 Problem Setup

Notations. All vectors considered in this paper are column vectors. We use $\mathbb{1}\{A\}$ to denote the indicator function of A, i.e., $\mathbb{1}\{A\} = 1$ if A holds, and $\mathbb{1}\{A\} = 0$ otherwise. $\|W\|$ represents the Frobenious norm of the matrix W. For a vector v, we use $[v]_i$ to denote the i-th coordinate of v. We use $\phi(v)$ to denote the softmax function, i.e., $[\phi(v)]_i = \exp(v_i)/\sum_j \exp(e_j^\top v)$, which can be applied to any vector with arbitrary dimension. We use $\{e_i\}_{i\in[|\mathcal{V}|]}$ to denote the canonical basis of $\mathbb{R}^{|\mathcal{V}|}$, i.e., $[e_i]_j = \mathbb{1}\{i = j\}$. The inner product $\langle A, B\rangle$ of two matrices A, B equals to $\mathrm{Trace}(AB^\top)$.

Next-token prediction. We consider the task of next-token prediction, which aims to predict the subsequent token in a token sequence given its preceding tokens. Formally, suppose that there exists a finite vocabulary set $\mathcal{V} \subset \mathbb{R}^d$ that consists of all possible tokens, where d is the dimension of the embedding. Each token $x \in \mathcal{V}$ is associated with a unique index $\mathrm{I}(x) \in \{1, 2, \ldots, |\mathcal{V}|\}$, where I is the index function. An L-length sentence $X = [x_1, \ldots, x_L] \in \mathcal{V}^L \subset \mathbb{R}^{d\times L}$ is a sequence of L tokens, where L is an integer. We assume that the maximum length of sentences is $L_{\max}$. The subsequent tokens in sentences are generated from a set of ground-truth model $\{p_L^* : \mathcal{V}^L \to \mathcal{V}\}_{L<L_{\max}}$, where p_L^* generates the next token x_{L+1} given the sentence X for any $1 \le L < L_{\max}$. The task of next-token prediction requires us to learn all models $\{p_L^*\}_{L<L_{\max}}$ given a training dataset $\mathcal{D}_0 = \{(X, x_{L+1}) | L < L_{\max}, X \in \mathcal{V}^L, x_{L+1} \in \mathcal{V}\}$. Notably, if $X = [x_1, \ldots, x_L] \in \mathcal{D}_0$, then for any $\ell < L$, $([x_1, \ldots, x_\ell], x_{\ell+1})$ is also a training sample, since it follows p_ℓ^* as well.

Decoder-only transformer. A decoder-only transformer is a stack of blocks consisting of a self-attention layer and a feed-forward layer. For simplicity, we consider one-layer transformer, where the self-attention layer is determined by three matrices: $W_{\mathrm{k}} \in \mathbb{R}^{d\times d_1}$, $W_{\mathrm{q}} \in \mathbb{R}^{d_1\times d}$ and $W_{\mathrm{v}} \in \mathbb{R}^{d_2\times d}$, namely key, query, and value matrices, and the feed-forward layer is determined by $W_{\mathrm{o}} \in \mathbb{R}^{|\mathcal{V}|\times d_2}$. Here d_1, d_2 are hidden dimensions. Mathematically, given the input $X = [x_1 \ldots, x_L]$, we write the one-layer transformer as $\mathrm{T}_\theta(X) := \phi(W_{\mathrm{o}}W_{\mathrm{v}}X\phi(X^\top W_{\mathrm{k}}W_{\mathrm{q}}x_L)) \in [0, 1]^{|\mathcal{V}|}$, where $\theta := (W_{\mathrm{o}}, W_{\mathrm{v}}, W_{\mathrm{k}}, W_{\mathrm{q}})$, and ϕ is the softmax function. We note that the inner softmax function ϕ is part of the attention model, and the outer softmax function ϕ is the decoder that generates a probability distribution over $\mathcal{V}$ for token prediction.

Reparameterization. We reparameterize the transformer architecture by consolidating the key and query matrices into a unified matrix W_{kq}, such that $W_{\mathrm{kq}} = W_{\mathrm{k}}W_{\mathrm{q}}$. Similarly, we reparameterize the product of the feed-forward (W_{o}) and value (W_{v}) matrices as a single matrix W_{ov}, defined as $W_{\mathrm{ov}} = W_{\mathrm{o}}W_{\mathrm{v}}$. Such a reparameterization is commonly adopted in transformer theory works (Huang et al., 2023; Tian et al., 2023a; Li et al., 2024; Nichani et al., 2024). Thus, the transformer under those reparameterization is given by $\mathrm{T}_\theta(X) := \phi(W_{\mathrm{ov}}X\phi(X^\top W_{\mathrm{kq}}x_L)) \in [0, 1]^{|\mathcal{V}|}$.

Cross-entropy loss. Given the training dataset $\mathcal{D}_0$ and the transformer model, we seek to learn p_* by minimizing (training) the *cross-entropy* loss $\mathcal{L}(\theta)$ defined as follows:

$$\mathcal{L}(\theta) = -\frac{1}{|\mathcal{D}_0|}\sum_{(X,x_{L+1})\in\mathcal{D}_0} \log e_{\mathrm{I}(x_{L+1})}^\top \mathrm{T}_\theta(X),$$

where $\mathrm{I}(x_{L+1})$ is the index of x_{L+1} in $\mathcal{V}$.

4 Realizable Training Dataset and Two-Stage Algorithm

In this section, we first provide a mathematical framework based on partial order to formally characterize a realizable training dataset for next-token prediction. We will then describe a two-stage algorithm for next-token prediction that we study.

4.1 Realizable Training Dataset

We characterize a realizable training dataset via two structural properties, where the training loss can be made arbitrarily close to zero. We first provide some intuitions about those two properties.

Existence of "collocation". First, we note that if a sentence $X = [x_1, \ldots, x_L]$ is a legal training sample, $([x_1], x_2)$ is also in the training dataset. In addition, the output of a transformer given one single input token only depends on W_{ov}, i.e. the feed-forward layer. Since training loss can be arbitrarily close to 0, there exists a sequence $\{W_t\}$ such that $\lim_{t\to\infty} -\sum_{x\in\mathcal{D}_0} \log e_\iota(x)^\top \phi(W_t x) = 0$, where $\iota(x)$ is the index of next token of x, and the summation is over the case when x is the first token. Due to that $\phi(W_t x)$ is a probability distribution, the equality holds only when ι is injective, since otherwise it is an entropy of some distribution which is strictly greater than 0. Therefore, there exists an injective map $\mathrm{n} : \mathcal{V} \to \mathcal{V}$ such that every sentence starts with x, must have a unique next token $\mathrm{n}(x)$. We call the set of pairs $\{x, \mathrm{n}(x)\}_{x\in\mathcal{V}}$ a *collocation*. We remark that $p_1^* = \mathrm{n}$.

Existence of "order". Second, let us consider the output of a transformer T_θ given a legal sentence $X = [x_1, \ldots, x_L]$ with the next token $x_{L+1} = p_L^*(X)$. The transformer first calculates a convex combination of $x_1, \ldots, x_L$ with corresponding weight $\varphi_\ell \propto \exp(x_\ell^\top W_{\mathrm{kq}} x_L)$ for each $\ell \le L$. Then, the transformer outputs $\phi(\sum_\ell W_{\mathrm{ov}} x_\ell \varphi_\ell)$. Recall that the collocation forces x_ℓ to map to $\mathrm{n}(x_\ell)$, thus $\phi(W_{\mathrm{ov}} x_\ell)$ has a peak value at the coordinate equal to $\mathrm{I}(\mathrm{n}(x_\ell))$ (the index of $\mathrm{n}(x_\ell)$). Hence, $\mathrm{T}_\theta(X)$ can only have peak value at the coordinates within the set $\{\mathrm{I}(\mathrm{n}(x_\ell))\}_{\ell\le L}$. If the training loss can be arbitrarily close to 0, it is desirable to have $\mathrm{n}^{-1}(x_{L+1}) \in \{x_\ell\}_{\ell\le L}$. Therefore, for those x_ℓ with $\mathrm{n}(x_\ell) = x_{L+1}$, φ_ℓ must be larger than $\varphi_{\ell'}$ with $\mathrm{n}(x_{\ell'}) \neq x_{L+1}$. Finally, it worth noting that φ_ℓ depends on the final token x_L. This observation motivates us to define *query-dependent partial orders* on $\mathcal{V}$.

Definition 1 (x^q-partial order) *Fix a token x^q. An x^q-partial order assigns an ordering relationship $>_{x^q}$ for certain pairs of tokens in $\mathcal{V}$, and is created as follows. Let $\mathcal{D}_0^{x^q}$ be the set of all legal sentences in the training dataset that has the final token (query) x^q. Then, for any pair of tokens $x, x' \in \mathcal{V}$, we assign $x >_{x^q} x'$ if there exists a sentence $X = [x_1, \ldots, x_L] \in \mathcal{D}_0^{x^q}$ and x, x' are tokens in X such that $\mathrm{n}(x) = x_{L+1} \neq \mathrm{n}(x')$, where x_{L+1} is the next token of X.*

Note that Definition 1 is a "constructive definition" which might not be well-defined. However, as we are under the setting when the training loss can be arbitrarily close to 0, the aforementioned discussion shows that if $x >_{x^q} x'$, then $\varphi_\ell > \varphi_{\ell'}$, where $x = x_\ell$ and $x' = x_{\ell'}$ in some sentence. Thus, $\exp(x W_{\mathrm{kq}} x^q) > \exp(x' W_{\mathrm{kq}} x^q)$, which indeed need to be well-defined. Otherwise, we will have contradictions such as $\exp(x W_{\mathrm{kq}} x^q) > \exp(x' W_{\mathrm{kq}} x^q) < \exp(x W_{\mathrm{kq}} x^q)$. Mathematically, a well-defined (strict) partial order $>$ on a set $\mathcal{V}$ satisfies two axioms (Yannakakis, 1982): (i) there is no $x > x$; (ii) if $x > x'$ and $x' > x''$, then $x > x''$. Thus, x^q-partial order created by $\mathcal{D}_0$ is well-defined for every $x^q \in \mathcal{V}$.

Finally, let us discuss the impact of query-dependent partial orders on $\mathcal{D}_0$. For a given query x^q, the partial order $>_{x^q}$ divides tokens in $\mathcal{V}$ into four disjoint types.

- *(Strict) optimal tokens.* A token x is optimal, if there is no x' such that $x' >_{x^q} x$[1].
- *Confused tokens.* A token x is confused, if there exists x', x'' such that $x' >_{x^q} x >_{x^q} x''$.
- *(Strict) non-optimal tokens.* A token x is non-optimal if there is no x' such that $x >_{x^q} x'$[2].
- *Non-comparable tokens.* A token x is non-comparable if there is no x' such that $x >_{x^q} x'$ or $x' >_{x^q} x$.

[1] This is also related to the maximal element in a partially ordered set.

[2] This is also related to the minimal element in a partially ordered set.

In this work, we assume that there are no confused tokens. This assumption simplifies the problem, making it tractable to provide explicit convergence in direction for training a transformer in Section 5. In summary, we make the following structural assumption on the training dataset.

Assumption 1 (Realizable training dataset) *$\mathcal{D}_0$ admits (i) a collocation $\{x, \mathrm{n}(x)\}_{x\in\mathcal{V}}$; (ii) well-defined query-dependent partial orders, where every x^q-partial order has no confused tokens.*

We remark that combining the collocation and query-dependent partial orders, we can regenerate the training dataset as follows. For any sentence with only one token $X = [x]$, the next token is $\mathrm{n}(x)$. For other sentences $X = [x_1, \ldots, x_L]$, let x_ℓ be optimal under the partial order $>_{x_L}$, and then the next token of X is $\mathrm{n}(x_\ell)$. We next provide a simple example that justifies Assumption 1.

Example 1 *Consider a language system where the vocabulary consists of four tokens {S, V, O, P}, where S,V,O,P respectively stand for subject, verb, object, and punctuation mark. This system admits the commonly adopted word order (Dryer, 1991): S, V, O, P. Let the training dataset be {SVOP, VOP, OPP, PSV}.*

Let us create the corresponding collocation and the query-dependent partial orders from the dataset. The collocation is $\{(S, V), (V, O), (O, P), (P, S)\}$. That is, if a sentence starts with a subject, then the next token is a verb. Similarly, if a sentence starts with a verb, then the next token is an object, and so on. The query-dependent partial orders are created as follows:

Partial order under query S. $S>_S P$. *Partial order under query V. $V>_V S$.*
Partial order under query O. $O>_O S$, $O>_O V$. *Partial order under query P. $O>_P P$.*

Therefore, if a sentence starts with S (subject), the next token is V (verb) according to the collocation. Then, for the sentence SV, since the query is V and $V>_V S$, the next token of the sentence coincides with the next token of V, which is exactly O (object). Finally, for the sentence SVO, following similar argument, the next token is P (punctuation mark). This example satisfies Assumption 1 and aligns with real-world scenarios. An illustration is provided in Figure 1.

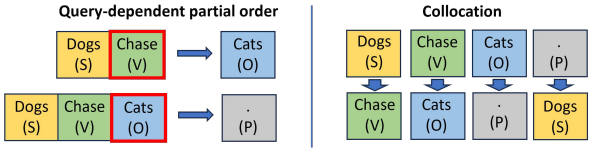

Figure 1: The left plot shows the mapping from sentence to the next token. The red rectangle indicates the optimal token in the corresponding sentence. The right plot shows the collocation relationship.

Additional notations of training data. It is worth noting that there are only finite number of distinct sentences. For ease of presentation, we introduce the following notations. Suppose there are N distinct sentences in the training dataset $\mathcal{D}_0$ indexed by $n \in \{1, \ldots, N\}$. For each distinct sentence $X^{(n)}$, we calculate its frequency $\pi^{(n)} \in [0, 1]$ in dataset $\mathcal{D}_0$ as $\pi^{(n)} = \frac{\sum_{(X, x_{L+1})\in\mathcal{D}_0} \mathbb{1}\{X=X^{(n)}\}}{|\mathcal{D}_0|}$.

Building upon this, with a little abuse of notation, we use $\mathrm{n}(X^{(n)}) \in \mathcal{V}$ to denote the subsequent token of the sentence $X^{(n)}$ and $\mathrm{In}(X^{(n)})$ to denote the index of $\mathrm{n}(X^{(n)})$.

We further denote $X_{-1}^{(n)}$ as the final token of $X^{(n)}$, and let $\bar{\mathrm{T}}_\theta(X) = W_{\mathrm{ov}} X \phi(X^\top W_{\mathrm{kq}} X_{-1}^{(n)})$. Then, the loss function $\mathcal{L}(\theta)$ can be rewritten as follows:

$$\mathcal{L}(\theta) = \sum_n \pi^{(n)} \left(\log \left(\sum_v \exp \left(e_v^\top \bar{\mathrm{T}}_\theta(X^{(n)}) \right) \right) - e_{\mathrm{In}(X^{(n)})}^\top \bar{\mathrm{T}}_\theta(X^{(n)}) \right). \tag{1}$$

4.2 Training Algorithm

For the realizable dataset satisfying Assumption 1, we propose a two-stage training algorithm using normalized gradient descent (NGD). The pseudo code of the algorithm is presented in Algorithm 1. In Section 5, we show that the two-stage algorithm decouples the training of the feed-forward and

attention layers without losing the optimality. This is because the training in stage 1 is designed to yield a suitable max-margin solution, which will enable the training of stage 2 to solve a fine-grained classifcation problem and identify the optimal token for prediction.

In the first stage of pre-processing, we use the collocation set to train the feed-forward layer $W_{\rm ov}$. For simplicity, we introduce the following notation for the training loss of the feed-forward layer. Given a collocation $\{x, \mathrm{n}(x)\}_{x\in\mathcal{V}}$, which can be obtained through extracting all length-2 sentences in the training dataset $\mathcal{D}_0$[3], we use normalized gradient descent to train $W_{\rm ov}$. Equivalently, the loss function can be written as

$$\mathcal{L}_0(W_{\rm ov}) = -\sum_{x\in\mathcal{V}} \log \frac{\exp(e_{\mathrm{In}(x)}^\top W_{\rm ov} x)}{\sum_{v\le|\mathcal{V}|}\exp(e_v^\top W_{\rm ov} x)},$$

where the self-attention elements are removed because the attention matrices are not trained here. Based on the above loss function, we initialize $W_{\rm ov}^{(0)} = 0 \in \mathbb{R}^{|\mathcal{V}|\times d}$, and subsequently take an update at each time t by NGD as in line 4 of Algorithm 1.

In the second stage, we fix the trained feed-forward layer and train the self-attention layer based on the loss function given in Equation (1) and using the entire dataset $\mathcal{D}_0$. Specifically, we initialize $W_{\rm kq} = 0 \in \mathbb{R}^{d\times d}$, and subsequently take an update at each time t by NGD as in line 7 of Algorithm 1.

Algorithm 1 Two-stage Normalized Gradient Descent

1: **Initialization:** $W_{\rm ov}^{(0)} = 0 \in \mathbb{R}^{|\mathcal{V}|\times d}$, $W_{\rm kq} = 0 \in \mathbb{R}^{d\times d}$.
2: **Input:** A collocation $\{x, \mathrm{n}(x)\}_{x\in\mathcal{V}}$, and a training dataset $\mathcal{D}_0$, learning rate η_0, η.
3: **for** $t \in \{0, 1, ..., T-1\}$ **do**
4: Update $W_{\rm ov}^{(t+1)}$ as $W_{\rm ov}^{(t+1)} = W_{\rm ov}^{(t)} - \eta_0 \frac{\nabla_{W_{\rm ov}}\mathcal{L}_0(W_{\rm ov}^{(t)})}{\|\nabla_{W_{\rm ov}}\mathcal{L}_0(W_{\rm ov}^{(t)})\|}$.
5: **end for**
6: **for** $t \in \{0, \ldots, T_1 - 1\}$ **do**
7: Update $W_{\rm kq}^{(t+1)}$ as $W_{\rm kq}^{(t+1)} = W_{\rm kq}^{(t)} - \eta \frac{\nabla_{W_{\rm kq}}\mathcal{L}(\theta^{(t)})}{\|\nabla_{W_{\rm kq}}\mathcal{L}(\theta^{(t)})\|}$, where $\theta^{(t)} = (W_{\rm ov}^{(T)}, W_{\rm kq}^{(t)})$.
8: **end for**

5 Training Dynamics of the Transformer

In this section, we present the convergence result for Algorithm 1. Before we proceed, we first introduce the following technical assumption, which has been commonly adopted in the previous theoretical studies of transformers (Huang et al., 2023; Li et al., 2024; Tian et al., 2023a).

Assumption 2 *The vocabulary set is orthornormal. Namely, the embedding has unit norm, i.e.,* $\|x\| = 1$*, and* $x^\top x' = 0$ *holds for any distinct tokens* x *and* x'*.*

5.1 Convergence of Training $W_{\rm ov}$

To characterize the training dynamics of $W_{\rm ov}$, we observe that the collocation $\{(x, \mathrm{n}(x))\}_{x\in\mathcal{V}}$ defines the following hard-margin problem:

$$W_{\rm ov}^* = \arg\min \|W\|, \quad \text{s.t.} \quad (e_{v^*} - e_v)Wx \ge 1, \quad \forall v^* = \mathrm{In}(x), v \ne \mathrm{In}(x). \tag{2}$$

It can be shown that $\lim_{B\to+\infty}\mathcal{L}_0(BW_{\rm ov}^*) = 0$. Thus, the loss function $\mathcal{L}_0$ trains $W_{\rm ov}$ to be the max-margin solution with $W_{\rm ov}x$ distinguishing the next token $\mathrm{n}(x)$ from all other tokens in $\mathcal{V}$.

Since $\mathcal{L}_0(\cdot)$ is convex, we have the following convergence result on the training of $W_{\rm ov}^{(t)}$.

Proposition 1 *Let* $W_{\rm ov}^*$ *be defined in Equation* (2)*. Under Assumptions 1-2, let* $W_{\rm ov}^{(t)}$ *be updated by Algorithm 1. Then, for any* $t \ge 2$*, we have* $\frac{t\eta_0}{2\|W_{\rm ov}^*\|} \le \|W_{\rm ov}^{(t)}\| \le t\eta_0$ *and the following bound holds:*

[3] A more general way is to use various standard techniques developed in linguistic analysis (Lehecka, 2015).

$$\left\langle \frac{W_{\mathrm{ov}}^{(t)}}{\|W_{\mathrm{ov}}^{(t)}\|}, \frac{W_{\mathrm{ov}}^{*}}{\|W_{\mathrm{ov}}^{*}\|} \right\rangle \geq 1 - \frac{5\|W_{\mathrm{ov}}^{*}\|^3 \log(2|\mathcal{V}|)\log t}{t\eta_0}.$$

Moreover, the loss function $\mathcal{L}_0$ satisfies that $\mathcal{L}_0(W_{\mathrm{ov}}^{(t)}) \leq O(\exp(-\eta_0 t/(4\|W_{\mathrm{ov}}^{*}\|)))$.

Proposition 1 states that during the training stage 1, the feed-forward layer $W_{\mathrm{ov}}^{(t)}$ converges in direction to $W_{\mathrm{ov}}^{*}/\|W_{\mathrm{ov}}^{*}\|$ at a rate of $O(\log t/t)$, which classifies the next token from all other tokens. In addition, since the norm of $W_{\mathrm{ov}}^{(t)}$ increases linearly, the loss $\mathcal{L}_0(W_{\mathrm{ov}}^{(t)})$ converges linearly to zero, i.e., $\mathcal{L}_0(W_{\mathrm{ov}}^{(t)}) = O(\exp(-C_0 t))$ for some constant C_0.

5.2 Convergence of Training W_{kq}

Recall that after the training stage 1 with T steps, we obtain a trained feed-forward layer $W_{\mathrm{ov}}^{(T)}$. Then, we fix $W_{\mathrm{ov}}^{(T)}$ and use normalized gradient descent to train W_{kq}. To characterize the training dynamics of the key-query matrix W_{kq}, we note that each query-dependent partial order also defines a hard-margin problem. Let $l(n) \subset \{1, \ldots, L^{(n)}\}$ be the set of indices of the optimal tokens of $X^{(n)}$. Recall that x_ℓ is optimal if there is no $x_{\ell'}$ such that $x_{\ell'} >_{X_{-1}^{(n)}} x_\ell$ and $\mathrm{In}(x_\ell) = \mathrm{In}(X^{(n)})$. That is, $W_{\mathrm{kq}} X_{-1}^{(n)}$ should correctly classify optimal tokens x_ℓ and non-optimal tokens $x_{\ell'}$. This is formalized in the following problem:

$$W_{\mathrm{kq}}^{*} = \arg\min \|W\|, \quad \text{s.t.} \quad (x_{\ell_*}^{(n)} - x_{\ell}^{(n)}) W X_{-1}^{(n)} \geq 1, \quad \forall \ell_* \in l(n), \ell \notin l(n), \forall n. \tag{3}$$

We will show that the loss function in Equation (1) given the well trained $W_{\mathrm{ov}}^{(T)}$ will train W_{kq} towards the max-margin solution W_{kq}^{*} in direction for classifying between the optimal and non-optimal token. We further make the following technical assumption.

Assumption 3 *For any sample $X^{(n)}$, the number of optimal tokens is not less than the number of non-optimal tokens. Formally, for any non-optimal token x in $X^{(n)}$, we have $|l(n)| \geq \sum_\ell \mathbb{1}\{x_\ell = x\}$.*

Assumption 3 is consistent with practical and empirical observations, where optimal tokens often demonstrate higher relevance, making them more frequent in subsequent outcomes.

We now present the convergence result for the training of the key-query matrix in stage 2.

Theorem 1 *Let Assumptions 1-3 hold. Let W_{kq}^{*} be the solution of Equation* (3)*. Let $\eta < O(1)$ and $W_{\mathrm{kq}}^{(t)}$ be updated by Algorithm 1. Then, for any $t \geq 2$, we have that $\frac{t\eta}{2\|W_{\mathrm{kq}}^{*}\|} \leq \|W_{\mathrm{kq}}^{(t)}\| \leq t\eta$. In addition, the following inequality holds.*

$$\left\langle \frac{W_{\mathrm{kq}}^{(t)}}{\|W_{\mathrm{kq}}^{(t)}\|}, \frac{W_{\mathrm{kq}}^{*}}{\|W_{\mathrm{kq}}^{*}\|} \right\rangle \geq 1 - \frac{54 N L_{\max}^4 \|W_{\mathrm{kq}}^{*}\|^4 \log^2 t}{t\eta}.$$

Theorem 1 states that the key-query matrix W_{kq} converges in direction to the max-margin solution $W_{\mathrm{kq}}^{*}/\|W_{\mathrm{kq}}^{*}\|$ at a convergence rate of $O(\log^2 t/t)$. We further show that the norm of W_{kq} also grows linearly in t, i.e., $\|W_{\mathrm{kq}}\| = \Omega(t)$. Combining these results, we have the following theorem on the convergence of the loss function and the training accuracy.

Theorem 2 (Loss Convergence) *For any training sentence $X^{(n)} = [x_1^{(n)}, \ldots, x_L^{(n)}]$, let $\varphi_\ell^{(n,t)} \propto \exp(x_\ell^{(n)} W_{\mathrm{kq}}^{(t)} x_L^{(n)})$ be the attention weight. Under the conditions in Theorem 1, there is an absolute constant c_0 such that when $T \geq c_0 \|W_{\mathrm{ov}}^{*}\|^5 \log(|\mathcal{V}|)\log T/\eta_0$ and $t \geq c_0 N L_{\max}^4 \|W_{\mathrm{kq}}^{*}\|^6 \log^2 t/\eta$, the optimal token weight satisfies $\min_n \sum_{\ell_* \in l(n)} \varphi_{\ell_*}^{(n,t)} \geq (1 + L_{\max}\exp(-tC_1)))^{-1}$. In addition, the loss function $\mathcal{L}$ converges linearly[4] to its minimal value:*

$$\mathcal{L}(\theta^{(t)}) = \mathcal{L}(W_{\mathrm{ov}}^{(T)}, W_{\mathrm{kq}}^{(t)}) \leq |\mathcal{V}| \exp\left(-TC_0\left(1 - \frac{2L_{\max}}{L_{\max} + \exp(C_1 t)}\right)\right), \tag{4}$$

[4]For fixed $W_{\mathrm{ov}}^{(T)}$, the minimum loss value $\mathcal{L}^* = |\mathcal{V}|\exp(-TC_0)$. Then Equation (4) implies $\mathcal{L}(\theta^{(t)}) - \mathcal{L}^* \leq \mathcal{L}^* TC_0 O(e^{-C_1 t})$ for sufficiently large t, which further implies the linear convergence in t.

where $C_0 = \frac{\eta_0}{4\|W_{\rm ov}^*\|^2}$ *and* $C_1 = \eta/(4L_{\max}\|W_{\rm kq}^*\|^2)$.

Theorem 2 shows that the training loss converges to its minimum value at a linear convergence rate. Furtherm, for $T = \Omega(\log(1/\epsilon_0))$ $t = \Omega(\log(1/\epsilon))$, the optimal token weight is given by $1/(1+\epsilon)$ for any $\epsilon > 0$, which is close to 1. This implies that the trained transformer attends to the optimal token and thus outputs the correct next token $\mathrm{n}(x_{\ell_*}^{(n)})$ with probability $1 - O(\epsilon_0)$.

5.3 Proof Sketch of Theorem 1

The proof consists of the following three main steps. The key proof step lies in carefully analyzing the projection of gradient $\nabla_{W_{\rm kq}}\mathcal{L}(\theta^{(t)})$ onto the token-query outer product $x_\ell^{(n)}(X_{-1}^{(n)})^\top$, max-margin attention weight matrix $W_{\rm kq}^*$, and the trained attention weight matrix $W_{\rm kq}^{(t)}$.

Step 1 (Lemma 5). By analyzing $\left\langle \nabla_{W_{\rm kq}}\mathcal{L}(\theta^{(t)}), x_\ell^{(n)}(X_{-1}^{(n)})^\top \right\rangle$, we characterize the dynamics of attention weights. Using mathematical induction, we show that the lower bound of optimal token weight is $1/L_{\max}$.

Step 2 (Lemma 6). Then, we show that the cosine similarity between the negative gradient and $W_{\rm kq}^*$ is strictly larger than the minimum optimal token weight. Utilizing step 1, due to the NGD update, the norm of the key-query matrix $W_{\rm kq}^{(t)}$ can be shown to grow linearly.

Step 3 (Lemma 7). Finally, we carefully compare the difference between the projections from gradient to the trained attention matrix and max-margin attention matrix. By separately evaluating the impact of the optimal and non-optimal tokens on those projections, we can show the following inequality for some constant C_0:

$$\left\langle \nabla_{W_{\rm kq}}\mathcal{L}(\theta^{(t)}), W_{\rm kq}^{(t)} \right\rangle \geq \left(1 + \frac{C_0 \log\|W_{\rm kq}^{(t)}\|}{\|W_{\rm kq}^{(t)}\|}\right) \left\langle \nabla_{W_{\rm kq}}\mathcal{L}(\theta^{(t)}), W_{\rm kq}^{*} \right\rangle \frac{\|W_{\rm kq}^{(t)}\|}{\|W_{\rm kq}^{*}\|}.$$

Utilizing step 2's result that $\|W_{\rm kq}^{(t)}\|$ grows linearly, the dynamics of the attention layer can be shown to converge in direction to the max-margin solution in Equation (3).

6 Generalization Ability

In this section, we prove the generalization ability of the trained transformers. Recall that Theorem 1 shows that $W_{\rm kq}^{(t)}$ converges to $W_{\rm kq}^*\|W_{\rm kq}^{(t)}\|/\|W_{\rm kq}^*\|$. To characterize the generalization ability, it is desirable to use the property of $W_{\rm kq}^*$, which is given in the following result.

Proposition 2 *Under Assumptions 1-2, fix a query token* x^q*, let* $\mathcal{O}_{x^q}, \mathcal{N}_{x^q}, \mathcal{M}_{x^q} \subset \mathcal{V}$ *be the set of optimal tokens, the set of non-optimal tokens, and the set of non-comparable tokens, under* x^q*-partial order, respectively. Then, the solution* $W_{\rm kq}^*$ *of Equation* (3) *satisfies* $x_0^\top W_{\rm kq}^* x^q = 0$ *for* $x_0 \in \mathcal{M}_{x^q}$*, and*

$$x_*^\top W_{\rm kq}^* x^q = \frac{|\mathcal{N}_{x^q}|}{|\mathcal{O}_{x^q}| + |\mathcal{N}_{x^q}|}, \quad x^\top W_{\rm kq}^* x^q = -\frac{|\mathcal{O}_{x^q}|}{|\mathcal{O}_{x^q}| + |\mathcal{N}_{x^q}|}, \quad \forall x_* \in \mathcal{O}_{x^q}, x \in \mathcal{N}_{x^q}.$$

Recall that non-comparable tokens (see Section 4) under a query x^q never appears in any training sentence data with the same query x^q. Thus, Proposition 2 implies an interesting generalization capability – each x^q-partial order can automatically incorporate more relationships to expand the query-dependent partial orders. Combining Proposition 2 with Theorem 1, we obtain the following theorem on $W_{\rm kq}^{(t)}$.

Theorem 3 *Under the conditions and notations in Proposition 2, let* $T = \Omega(\log(1/\epsilon))$*, and* $t = \Omega(\log(1/\epsilon))$*. Then there exists a constant* C_0 *such that*

$$(x_* - x_0)^\top W_{\rm kq}^{(t)} x^q \geq C_0 t, \quad (x_0 - x)^\top W_{\rm kq}^{(t)} x^q \geq C_0 t, \quad \forall x_* \in \mathcal{O}_{x^q}, x_0 \in \mathcal{M}_{x^q}, x \in \mathcal{N}_{x^q}.$$

Moreover, if the trained transformer takes input X *with query* x^q *that consists of a non-comparable token* x_0 *and non-optimal tokens, then the prediction made by* $\mathrm{T}_{\theta^{(t)}}(X)$ *is* $\mathrm{n}(x_0)$ *with high probability.*

Theorem 3 suggests that a new partial order is created by the trained transformer. Specifically, it inserts the non-comparable tokens between the optimal and non-optimal tokens. The trained transformer can generalize the token prediction to such new sentences as given in Theorem 3.

We use Example 1 to illustrate the generalization ability described above.

Example 2 (Generalization to unseen data in Example 1) *Recall that in Example 1, the training dataset consists of four sentences: SVOP, VOP, OPP, and PSV. Consider the partial order $>_P$ under the punctuation mark P. We have that O$>_P$P and O is an optimal token, P is a non-optimal token, and S,V are non-comparable tokens. We then have the following non-trivial prediction by the trained transformer.*

Case 1. *Non-comparable tokens are learned to be "larger" than non-optimal tokens.*

Consider a new (unseen) input sentence SP. Since S is non-comparable before training, but is "larger" than P under the trained key-query matrix $W_{\mathrm{kq}}^{(t)}$, the next predicted token is $\mathrm{n}(S)$ = V.

Case 2. *Optimal tokens remain optimal over all tokens after training.*

Consider a new (unseen) input OSP. O is optimal and S is still "smaller" than O under the trained P-partial order. The trained transformer will consistently predict P.

In both of the above cases, the trained transformer provides desirable prediction for the unseen sentences. We further note that the effectiveness of both cases can vary during the inference time of the trained transformer. For instance, if the input sequence is SP (subject-punctuation), the output is SPV (subject-P-verb), which follows a logical subject-verb order and is desirable. However, in cases where the input is VP (verb-punctuation), it may be preferable to terminate the sequence after the verb, i.e., VPP, as the verb alone can suffice to convey the intended meaning.

7 Experiment

In this section, we verify our theoretical findings via an experiment on a synthetic dataset. Specifically, we randomly generate a realizable dataset as described in Assumption 1 with $|\mathcal{V}| = 20$. Then, we train W_{ov} and W_{kq} by Algorithm 1, each with 900 iterations. The parameters are chosen as $d = |\mathcal{V}|$, $\eta_0 = 0.2/\sqrt{d}$, and $\eta = 0.05/\sqrt{d}$. In Figure 2, the first three plots show the dynamics of the training stage 1, which indicates the convergence of the loss $\mathcal{L}_0(W_{\mathrm{ov}}^{(t)})$ to its minimum value, the convergence of $W_{\mathrm{ov}}^{(t)}$ in direction to W_{ov}^*, and the linear increase of the norm $\|W_{\mathrm{ov}}^{(t)}\|$, respectively. These results verify Proposition 1. The last three plots show the dynamics of the training stage 2, which indicates the convergence of the loss $\mathcal{L}(\theta^{(t)})$, the convergence of $W_{\mathrm{kq}}^{(t)}$ in direction to W_{kq}^*, and the linear increase of the norm $\|W_{\mathrm{kq}}^{(t)}\|$. These results verify Theorem 1 and Theorem 2. All experiments are conducted on a PC equipped with an i5-12400F processor and 16GB of memory.

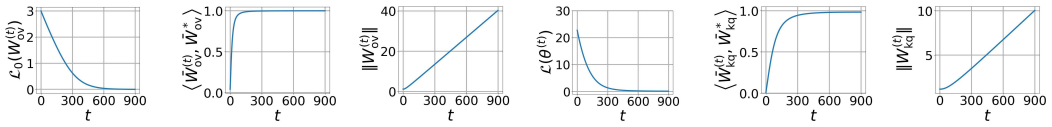

Figure 2: Training dynamics of single-layer transformer for NTP.

8 Conclusion

In this work, we investigated the training dynamics of a single-layer transformer for NTP. We first characterized two structural properties of the training dataset under the realizable setting where the training loss can be made arbitrarily close to zero. These properties allow us to define two max-margin solutions for both the feed-forward layer and the self-attention layer. Then, we showed that both layers converge in direction to their corresponding max-margin solutions sub-linearly, which further yields a linear convergence of the training loss for NTP. We further showed that the well trained transformer can have non-trivial prediction ability on unseen data, which sheds light on the generalization capability of transformers. Our experiments verify our theoretical findings.

Acknowledgments and Disclosure of Funding

The work of R. Huang and J. Yang was supported in part by the U.S. National Science Foundation under grants NSF CNS-1956276 and ECCS-2133170. The work of Y. Liang was supported in part by the U.S. National Science Foundation under grants ECCS-2413528 and DMS-2134145.

References

Ahn, K., Cheng, X., Daneshmand, H., and Sra, S. (2024). Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36.

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

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

Belkin, M. (2024). The necessity of machine learning theory in mitigating ai risk. *ACM/JMS Journal of Data Science*.

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., et al. (2021). On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258*.

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

Chen, S. and Li, Y. (2024). Provably learning a multi-head attention layer. *arXiv preprint arXiv:2402.04084*.

Chen, S., Sheen, H., Wang, T., and Yang, Z. (2024). Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*.

Cheng, X., Chen, Y., and Sra, S. (2023). Transformers implement functional gradient descent to learn non-linear functions in context. *arXiv preprint arXiv:2312.06528*.

Cui, Y., Ren, J., He, P., Tang, J., and Xing, Y. (2024). Superiority of multi-head attention in in-context linear regression. *arXiv preprint arXiv:2401.17426*.

Deora, P., Ghaderi, R., Taheri, H., and Thrampoulidis, C. (2023). On the optimization and generalization of multi-head attention. *arXiv preprint arXiv:2310.12680*.

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

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

Dryer, M. S. (1991). Svo languages and the ov: Vo typology1. *Journal of linguistics*, 27(2):443–482.

Edelman, B. L., Goel, S., Kakade, S., and Zhang, C. (2022). Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR.

Frei, S., Vardi, G., Bartlett, P. L., Srebro, N., and Hu, W. (2022). Implicit bias in leaky relu networks trained on high-dimensional data. *arXiv preprint arXiv:2210.07082*.

Fu, D., Chen, T.-Q., Jia, R., and Sharan, V. (2023). Transformers learn higher-order optimization methods for in-context learning: A study with linear models. *arXiv preprint arXiv:2310.17086*.

Giannou, A., Rajput, S., Sohn, J.-y., Lee, K., Lee, J. D., and Papailiopoulos, D. (2023). Looped transformers as programmable computers. In *International Conference on Machine Learning*, pages 11398–11442. PMLR.

Huang, Y., Cheng, Y., and Liang, Y. (2023). In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*.

Ji, Z., Srebro, N., and Telgarsky, M. (2021). Fast margin maximization via dual acceleration. In *International Conference on Machine Learning*, pages 4860–4869. PMLR.

Ji, Z. and Telgarsky, M. (2021). Characterizing the implicit bias via a primal-dual analysis. In *Algorithmic Learning Theory*, pages 772–804. PMLR.

Kim, J. and Suzuki, T. (2024). Transformers learn nonlinear features in context: Nonconvex mean-field dynamics on the attention landscape. *arXiv preprint arXiv:2402.01258*.

Kou, Y., Chen, Z., and Gu, Q. (2024). Implicit bias of gradient descent for two-layer relu and leaky relu networks on nearly-orthogonal data. *Advances in Neural Information Processing Systems*, 36.

Lehecka, T. (2015). Collocation and colligation. In *Handbook of pragmatics online*. Benjamins.

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

Li, Y., Huang, Y., Ildiz, M. E., Rawat, A. S., and Oymak, S. (2024). Mechanics of next token prediction with self-attention. In *International Conference on Artificial Intelligence and Statistics*, pages 685–693. PMLR.

Lin, L., Bai, Y., and Mei, S. (2023). Transformers as decision makers: Provable in-context reinforcement learning via supervised pretraining. *arXiv preprint arXiv:2310.08566*.

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

Nacson, M. S., Lee, J., Gunasekar, S., Savarese, P. H. P., Srebro, N., and Soudry, D. (2019). Convergence of gradient descent on separable data. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 3420–3428. PMLR.

Nichani, E., Damian, A., and Lee, J. D. (2024). How transformers learn causal structure with gradient descent. *arXiv preprint arXiv:2402.14735*.

Phuong, M. and Lampert, C. H. (2020). The inductive bias of relu networks on orthogonally separable data. In *International Conference on Learning Representations*.

Soudry, D., Hoffer, E., Nacson, M. S., Gunasekar, S., and Srebro, N. (2018). The implicit bias of gradient descent on separable data. *Journal of Machine Learning Research*, 19(70):1–57.

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

Tarzanagh, D. A., Li, Y., Zhang, X., and Oymak, S. (2023b). Max-margin token selection in attention mechanism. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Thrampoulidis, C. (2024). Implicit bias of next-token prediction.

Tian, Y., Wang, Y., Chen, B., and Du, S. S. (2023a). Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in Neural Information Processing Systems*, 36:71911–71947.

Tian, Y., Wang, Y., Zhang, Z., Chen, B., and Du, S. (2023b). Joma: Demystifying multilayer transformers via joint dynamics of mlp and attention. *arXiv preprint arXiv:2310.00535*.

Tsai, Y.-H. H., Bai, S., Liang, P. P., Kolter, J. Z., Morency, L.-P., and Salakhutdinov, R. (2019). Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558. NIH Public Access.

Vardi, G. (2023). On the implicit bias in deep-learning algorithms. *Communications of the ACM*, 66(6):86–93.

Vasudeva, B., Deora, P., and Thrampoulidis, C. (2024). Implicit bias and fast convergence rates for self-attention. *arXiv preprint arXiv:2402.05738*.

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

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

Yannakakis, M. (1982). The complexity of the partial order dimension problem. *SIAM Journal on Algebraic Discrete Methods*, 3(3):351–358.

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

Contents

A Expression of Gradients

We first provide the general formula for the gradients of both layers.

$$\nabla_{W_{\mathrm{ov}}} \mathcal{L}_0(W_{\mathrm{ov}}) = \sum_{x \in \mathcal{V}} \left(\mathrm{T}_0(x) - e_{\mathrm{In}(x)} \right) x^\top, \tag{5}$$

$$\begin{aligned} &\nabla_{W_{\mathrm{kq}}} \mathcal{L}(\theta) \\ &\quad = \sum_n \pi^{(n)} X^{(n)} \left(\mathrm{diag}(\phi_\theta(X^{(n)}) - \phi_\theta(X^{(n)}) \phi_\theta(X^{(n)})^\top \right) (X^{(n)})^\top W_{\mathrm{ov}}^\top \left(\mathrm{T}_\theta(X^{(n)}) - p^{(n)} \right) (X_{-1}^{(n)})^\top. \end{aligned} \tag{6}$$

B Proof of Proposition 1

Recall that we use the loss,

$$\mathcal{L}_0(W_{\mathrm{ov}}) = -\sum_{x \in \mathcal{V}} \log \frac{\exp\left(e_{\mathrm{In}(x)}^\top W_{\mathrm{ov}} x \right)}{\sum_{i \in [|\mathcal{V}|]} \exp\left(e_i^\top W_{\mathrm{ov}} x \right)}.$$

The updating rule of W_{ov} is that

$$W_{\mathrm{ov}}^{(t+1)} = W_{\mathrm{ov}}^{(t)} - \eta_0 \frac{\nabla_{W_{\mathrm{ov}}} \mathcal{L}_0(W_{\mathrm{ov}}^{(t)})}{\| \nabla_{W_{\mathrm{ov}}} \mathcal{L}_0(W_{\mathrm{ov}}^{(t)}) \|}. \tag{7}$$

We know that $\mathcal{L}_0$ is convex respect to W_{ov}. Therefore, we have

$$\left\langle W_{\mathrm{ov}}^{(t)} - W_{\mathrm{ov}}', \nabla_{W_{\mathrm{ov}}} \mathcal{L}_0(\theta^{(t)}) \right\rangle \geq \mathcal{L}_0(\theta^{(t)}) - \mathcal{L}_0(\theta').$$

It is clear that the loss function $\mathcal{L}$ reaches the minimum 0 when $W_{\mathrm{ov}} = \Delta W_{\mathrm{ov}}^*$, as $\Delta \to \infty$.

Lemma 1 *Under the initialization $W_{\mathrm{ov}}^{(0)}$ and the updating rule Equation* (7) *with step size η, the following inequality holds.*

$$t\eta_0 + \|W_{\mathrm{ov}}^{(0)}\| \geq \|W_{\mathrm{ov}}^{(t)}\| \geq \frac{t\eta_0}{2\|W_{\mathrm{ov}}^*\|} - \|W_{\mathrm{ov}}^{(0)}\|.$$

Proof. Using Equation (5), we have

$$\begin{aligned} \left\langle W_{\mathrm{ov}}^*, \nabla_{W_{\mathrm{ov}}} \mathcal{L}_0(W_{\mathrm{ov}}^{(t)}) \right\rangle &= \sum_{x \in \mathcal{V}} \left(\mathrm{T}_0^{(t)}(x) - e_{\mathrm{In}(x)} \right)^\top W_{\mathrm{ov}}^* x \\ &= \sum_{x \in \mathcal{V}} \sum_{i \in [|\mathcal{V}|]} [\mathrm{T}_0^{(t)}(x)]_i (e_i - e_{\mathrm{In}(x)})^\top W_{\mathrm{ov}}^* x \\ &\overset{(a)}{\leq} -\sum_{x \in \mathcal{V}} \sum_{i \neq \mathrm{In}(x)} [\mathrm{T}_0^{(t)}(x)]_i, \end{aligned} \tag{8}$$

where (a) is due the constraints that W_{ov}^* satisfies. On the other hand,

$$\|\nabla_{W_{\mathrm{ov}}} \mathcal{L}(W_{\mathrm{ov}}^{(t)})\| = \left\langle \frac{\nabla_{W_{\mathrm{ov}}} \mathcal{L}(W_{\mathrm{ov}}^{(t)})}{\|\nabla_{W_{\mathrm{ov}}} \mathcal{L}(W_{\mathrm{ov}}^{(t)})\|}, \nabla_{W_{\mathrm{ov}}} \mathcal{L}_0(W_{\mathrm{ov}}^{(t)}) \right\rangle$$

$$= \sum_{x\in\mathcal{V}}\sum_{i}[\mathrm{T}_0^{(t)}(x)]_i(e_i - e_{\mathrm{In}(x)})^\top \frac{\nabla_{W_{\mathrm{ov}}}\mathcal{L}(\theta^{(t)})}{\|\nabla_{W_{\mathrm{ov}}}\mathcal{L}(\theta^{(t)})\|}x$$
$$\overset{(a)}{\leq} 2\sum_{x\in\mathcal{V}}\sum_{i\neq\mathrm{In}(x)}[\mathrm{T}_0^{(t)}(x)]_i,$$

where (a) follows from $\|AB\| \leq \|A\|\|B\|$ for any matrices A and B. Thus, we obtain

$$\left\langle W_{\mathrm{ov}}^*, \frac{\nabla_{W_{\mathrm{ov}}}\mathcal{L}(\theta^{(t)})}{\|\nabla_{W_{\mathrm{ov}}}\mathcal{L}(\theta^{(t)})\|}\right\rangle \leq -1/2$$

To lower bound the norm of W_{ov}, we recall the updating rule (Equation (7)).

$$\begin{aligned}\|W_{\mathrm{ov}}^{(t)}\| &= \left\| W_{\mathrm{ov}}^{(0)} - \sum_{t'<t}\eta_0\frac{\nabla_{W_{\mathrm{ov}}}\mathcal{L}_0(W_{\mathrm{ov}}^{(t')})}{\|\nabla_{W_{\mathrm{ov}}}\mathcal{L}_0(W_{\mathrm{ov}}^{(t')})\|}\right\| \\ &\geq \left\langle W_{\mathrm{ov}}^{(0)} - \sum_{t'<t}\eta_0\frac{\nabla_{W_{\mathrm{ov}}}\mathcal{L}_0(W_{\mathrm{ov}}^{(t')})}{\|\nabla_{W_{\mathrm{ov}}}\mathcal{L}_0(W_{\mathrm{ov}}^{(t')})\|}, \frac{W_{\mathrm{ov}}}{\|W_{\mathrm{ov}}^*\|}\right\rangle \\ &\geq \frac{t\eta_0}{2\|W_{\mathrm{ov}}^*\|} - \|W_{\mathrm{ov}}^{(0)}\|.\end{aligned}$$

For the LHS of the inequality, it suffices to note that at each iteration, the norm $\|W_{\mathrm{ov}}^{(t)}\|$ increases most η_0 due to normalized gradient descent. ■

Lemma 2 *At each iteration t, the following inequality holds.*

$$\left\langle \frac{W_{\mathrm{ov}}^{(t)}}{\|W_{\mathrm{ov}}^{(t)}\|}, \nabla_{W_{\mathrm{ov}}}\mathcal{L}_0(W_{\mathrm{ov}}^{(t)})\right\rangle \geq \left(1 + \frac{2\|W_{\mathrm{ov}}^*\|}{\|W_{\mathrm{ov}}^{(t)}\|}\log(2|\mathcal{V}|)\right)\left\langle \frac{W_{\mathrm{ov}}^*}{\|W_{\mathrm{ov}}^*\|}, \nabla_{W_{\mathrm{ov}}}\mathcal{L}_0(W_{\mathrm{ov}}^{(t)})\right\rangle$$

Proof. First, we consider the case when $W_{\mathrm{ov}}^{(t)} = W_{\mathrm{ov}}^* \frac{\|W_{\mathrm{ov}}^{(t)}\|}{\|W_{\mathrm{ov}}^*\|}$. Due to Equation (8) in Lemma 1, we have

$$\left\langle \frac{W_{\mathrm{ov}}^*}{\|W_{\mathrm{ov}}^*\|}, \nabla_{W_{\mathrm{ov}}}\mathcal{L}_0(W_{\mathrm{ov}}^{(t)})\right\rangle < 0$$

In this case, the result is trivial.

Then, we consider the case when $W_{\mathrm{ov}}^{(t)} \neq W_{\mathrm{ov}}^* \frac{\|W_{\mathrm{ov}}^{(t)}\|}{\|W_{\mathrm{ov}}^*\|}$. Due the optimality of W_{ov}^*, which achieves the minimum norm satisfying the constraints in Equation (2), we must have that for some $x_0 \in \mathcal{V}$, there exists $i_0 \neq \mathrm{In}(x_0)$ such that the following inequality holds

$$(e_{\mathrm{In}(x_0)} - e_i)^\top W_{\mathrm{ov}}^{(t)} x_0 < \frac{\|W_{\mathrm{ov}}^{(t)}\|}{\|W_{\mathrm{ov}}^*\|}.$$

Therefore, the loss on $W_{\mathrm{ov}}^{(t)}$ can be lower bounded as follows.

$$\begin{aligned}\mathcal{L}(W_{\mathrm{ov}}^{(t)}) &= \sum_{x\in\mathcal{V}}\log\left(1 + \sum_i \exp\left((e_i - e_{\mathrm{In}(x)})^\top W_{\mathrm{ov}}^{(t)}x\right)\right) \\ &> \log\left(1 + \exp\left(-\|W_{\mathrm{ov}}^{(t)}\|/\|W_{\mathrm{ov}}^*\|\right)\right) \\ &\overset{(a)}{>} \frac{1}{2}\exp\left(-\|W_{\mathrm{ov}}^{(t)}\|/\|W_{\mathrm{ov}}^*\|\right),\end{aligned}$$

where (a) is due to the fact that $\log(1+x) \geq x/2$ when $0 < x < 1$. On the other hand, let $W'_{\mathrm{ov}} = \left(\frac{\|W_{\mathrm{ov}}^{(t)}\|}{\|W_{\mathrm{ov}}^*\|} + 2\log(2|\mathcal{V}|)\right)W_{\mathrm{ov}}^*$. Then, the loss on W'_{ov} has the following upper bound.

$$
\begin{aligned}
\mathcal{L}(W'_{\text{ov}}) &= \sum_{x\in\mathcal{V}} \log\left(1+\sum_i \exp\left((e_i - e_{\text{In}(x)})^\top W_{\text{ov}}^{(t)} x\right)\right) \\
&\le \sum_{x\in\mathcal{V}} \log\left(1+(|\mathcal{V}|-1)\exp\left(-\|W_{\text{ov}}^{(t)}\|/\|W_{\text{ov}}^*\| - \log(2|\mathcal{V}|)\right)\right) \\
&\overset{(a)}{\le} \sum_{x\in\mathcal{V}} |\mathcal{V}| \exp\left(-\|W_{\text{ov}}^{(t)}\|/\|W_{\text{ov}}^*\| - 2\log(2|\mathcal{V}|)\right) \\
&\le \frac{1}{2}\exp(-\|W_{\text{ov}}^{(t)}\|/\|W_{\text{ov}}^*\|),
\end{aligned}
$$

where (a) is due to the fact that $\log(1+x) < x$ when $x > 0$. Thus, $\mathcal{L}(W_{\text{ov}}^{(t)}) > \mathcal{L}(W'_{\text{ov}})$. Due to the convextiy of $\mathcal{L}_0$, we have

$$
\begin{aligned}
0 &< \left\langle W_{\text{ov}}^{(t)} - W'_{\text{ov}}, \nabla_{W_{\text{ov}}}\mathcal{L}_0(W_{\text{ov}}^{(t)})\right\rangle \\
&= \left\langle W_{\text{ov}}^{(t)}, \nabla_{W_{\text{ov}}}\mathcal{L}_0(W_{\text{ov}}^{(t)})\right\rangle - \left(\frac{\|W_{\text{ov}}^{(t)}\|}{\|W_{\text{ov}}^*\|} + 2\log(2|\mathcal{V}|)\right)\left\langle W_{\text{ov}}^*, \nabla_{W_{\text{ov}}}\mathcal{L}_0(W_{\text{ov}}^{(t)})\right\rangle,
\end{aligned}
$$

which finishes the proof.

■

Proposition 3 (Restatement of Proposition 1) *Under the zero initialization $W_{\text{ov}}^{(0)} = 0$ and updating rule Equation* (7)*, for any $t \ge 2$, the following inequality holds.*

$$
\left\langle \frac{W_{\text{ov}}^{(t)}}{\|W_{\text{ov}}^{(t)}\|}, \frac{W_{\text{ov}}^*}{\|W_{\text{ov}}^*\|}\right\rangle \ge 1 - \frac{12\|W_{\text{ov}}^*\|^3\log(2|\mathcal{V}|)\log t}{t\eta_0}.
$$

Moreover, $\frac{t\eta_0}{2\|W_{\text{ov}}^*\|} \le \|W_{\text{ov}}^{(t)}\| \le t\eta_0$.

Proof. The second argument about the norm of $W_{\text{ov}}^{(t)}$ follows directly from Lemma 1. We aim to prove the first part as follows.

Let $\alpha_t = \frac{2\|W_{\text{ov}}^*\|}{\|W_{\text{ov}}^{(t)}\|}\log(2|\mathcal{V}|)$. By Lemma 2 and the updating rule Equation (7), we have

$$
\begin{aligned}
\left\langle W_{\text{ov}}^{(t+1)} - W_{\text{ov}}^{(t)}, \frac{W_{\text{ov}}^*}{\|W_{\text{ov}}^*\|}\right\rangle &= -\eta_0\left\langle \nabla_{W_{\text{ov}}}\mathcal{L}(W_{\text{ov}}^{(t)}), \frac{W_{\text{ov}}^*}{\|W_{\text{ov}}^*\|}\right\rangle \\
&\ge -\frac{\eta_0}{1+\alpha_t}\left\langle \nabla_{W_{\text{ov}}}\mathcal{L}(W_{\text{ov}}^{(t)}), \frac{W_{\text{ov}}^{(t)}}{\|W_{\text{ov}}^{(t)}\|}\right\rangle \\
&= \frac{1}{1+\alpha_t}\left\langle W_{\text{ov}}^{(t+1)} - W_{\text{ov}}^{(t)}, \frac{W_{\text{ov}}^{(t)}}{\|W_{\text{ov}}^{(t)}\|}\right\rangle \\
&= \left(1-\frac{\alpha_t}{1+\alpha_t}\right)\left\langle W_{\text{ov}}^{(t+1)} - W_{\text{ov}}^{(t)}, \frac{W_{\text{ov}}^{(t)}}{\|W_{\text{ov}}^{(t)}\|}\right\rangle \\
&= \frac{1}{2\|W_{\text{ov}}^{(t)}\|}\left(\|W_{\text{ov}}^{(t+1)}\|^2 - \|W_{\text{ov}}^{(t+1)} - W_{\text{ov}}^{(t)}\|^2 - \|W_{\text{ov}}^{(t)}\|^2\right) \\
&\quad - \frac{\alpha_t}{1+\alpha_t}\left\langle W_{\text{ov}}^{(t+1)} - W_{\text{ov}}^{(t)}, \frac{W_{\text{ov}}^{(t)}}{\|W_{\text{ov}}^{(t)}\|}\right\rangle
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(a)}{=} \frac{\|W_{\text{ov}}^{(t+1)}\|^2 - \|W_{\text{ov}}^{(t)}\|^2}{2\|W_{\text{ov}}^{(t)}\|} - \frac{\eta^2}{2\|W_{\text{ov}}^{(t)}\|} \\
&\quad + \frac{\eta_0\alpha_t}{1+\alpha_t}\left\langle \frac{\nabla_{W_{\text{ov}}}\mathcal{L}(\theta^{(t)})}{\|\nabla_{W_{\text{ov}}}\mathcal{L}(\theta^{(t)})\|}, \frac{W_{\text{ov}}^{(t)}}{\|W_{\text{ov}}^{(t)}\|}\right\rangle \\
&\overset{(b)}{\geq} \|W_{\text{ov}}^{(t+1)}\| - \|W_{\text{ov}}^{(t)}\| - \frac{\eta_0^2}{2\|W_{\text{ov}}^{(t)}\|} - \frac{\eta_0\alpha_t}{1+\alpha_t},
\end{aligned}
$$

where (a) follows from that $\|W_{\text{ov}}^{(t+1)} - W_{\text{ov}}^{(t)}\| = \eta_0$, and (b) is due to the fact that $x^2 - y^2 \geq 2y(x-y)$ for any $x, y \in \mathbb{R}$.

Summing over t starting from 2, we have

$$
\left\langle W_{\text{ov}}^{(t)} - W_{\text{ov}}^{(2)}, \frac{W_{\text{ov}}^*}{\|W_{\text{ov}}^*\|}\right\rangle \geq \|W_{\text{ov}}^{(t)}\| - \|W_{\text{ov}}^{(2)}\| - \sum_{t'=2}^{t-1}\frac{\eta_0^2}{2\|W_{\text{ov}}^{(t')}\|} - \sum_{t'=2}^{t-1}\frac{\eta_0\alpha_{t'}}{1+\alpha_{t'}}.
$$

Furthermore, due to Lemma 1,

$$
\begin{aligned}
\sum_{t'=2}^{t-1}\frac{1}{\|W_{\text{ov}}^{(t')}\|} &\leq \sum_{t'=2}^{t-1}\frac{2\|W_{\text{ov}}^*\|/\eta_0}{t} \\
&\leq \frac{2\|W_{\text{ov}}^*\|}{\eta_0}\log t.
\end{aligned}
$$

Similarly,

$$
\begin{aligned}
\sum_{t'=2}^{t-1}\frac{\alpha_{t'}}{1+\alpha_{t'}} &\leq \sum_{t'=2}^{t-1}\frac{2\|W_{\text{ov}}^*\|\log(2|\mathcal{V}|)}{\|W_{\text{ov}}^{(t')}\|} \\
&\leq \frac{4\|W_{\text{ov}}^*\|^2\log(2|\mathcal{V}|)}{\eta_0}\log t
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
\left\langle \frac{W_{\text{ov}}^{(t)}}{\|W_{\text{ov}}^{(t)}\|}, \frac{W_{\text{ov}}^*}{\|W_{\text{ov}}^*\|}\right\rangle &\geq 1 - \frac{\|W_{\text{ov}}^{(2)}\| + 2\eta_0\|W_{\text{ov}}^*\|\log t + 4\|W_{\text{ov}}^*\|^2\log(2|\mathcal{V}|)\log t}{\|W_{\text{ov}}^{(t)}\|} \\
&\overset{(a)}{\geq} 1 - \frac{12\|W_{\text{ov}}^*\|^3\log(2|\mathcal{V}|)\log t}{t\eta_0},
\end{aligned}
$$

where (a) follows from Lemma 1 and $\|W_{\text{ov}}^{(2)}\| \leq 2\eta_0 \leq \|W_{\text{ov}}^*\|$, and $\|W_{\text{ov}}^{(t)}\| \geq t\eta_0/(2\|W_{\text{ov}}^*\|)$

■

C Proof of Theorem 1 and Theorem 2

C.1 Supporting Lemmas

Lemma 3 *With zero initialization, under the updating rule Equation* (7)*, for any iteration* t*,* $W_{\text{ov}}^{(t)}$ *satisfies that*

$$
(e_i - e_{i'})^\top W_{\text{ov}}^{(t)} x = 0, \quad \forall i, i' \neq \text{In}(x).
$$

Proof. The proof follows directly from induction and the fact that

$$
(e_i - e_{i'})^\top W_{\text{ov}}^{(t+1)} x = (e_i - e_{i'})^\top W_{\text{ov}}^{(t)} x - \frac{\eta_0([\mathrm{T}_0^{(t)}]_i - [\mathrm{T}_0^{(t)}]_{i'})}{\|\nabla_{W_{\text{ov}}}\mathcal{L}_0(W_{\text{ov}}^{(t)})\|}, \quad \forall i, i' \neq \text{In}(x).
$$

■

Corollary 1 *Under the settings in Proposition 1, let* $T \geq 384\|W_{\text{ov}}^*\|^5 \log(2|\mathcal{V}|)\log T/\eta_0$, *and* $\Delta = T\eta_0/(4\|W_{\text{ov}}^*\|^2)$

$$\begin{cases}(e_{\text{In}(x)} - e_i)^\top W_{\text{ov}} x \in (\Delta, 3\Delta), \forall i \neq \text{In}(x) \\ (e_i - e_{i'})^\top W_{\text{ov}} x = 0, \forall i, i' \neq \text{In}(x)\end{cases}$$

Proof. The second equality follows directly from Lemma 3.

To show the first equation, we analyze

$$\begin{aligned}
&(e_{\text{In}(x)} - e_i)^\top W_{\text{ov}}^{(T)} x \\
&\quad = (e_{\text{In}(x)} - e_i)^\top \frac{W_{\text{ov}}^* \|W_{\text{ov}}^{(T)}\|}{\|W_{\text{ov}}^*\|} x + \|W_{\text{ov}}^{(T)}\|(e_{\text{In}(x)} - e_i)^\top \left(\frac{W_{\text{ov}}^{(T)}}{\|W_{\text{ov}}^{(T)}\|} - \frac{W_{\text{ov}}^*}{\|W_{\text{ov}}^*\|}\right) W_{\text{ov}}^{(T)} x \\
&\quad \overset{(a)}{=} \frac{\|W_{\text{ov}}^{(T)}\|}{\|W_{\text{ov}}^*\|} - 2\sqrt{2}\|W_{\text{ov}}^{(T)}\| \sqrt{\frac{12\|W_{\text{ov}}^*\|^3 \log(2|\mathcal{V}|)\log T}{T\eta_0}} \\
&\quad \overset{(b)}{\geq} \frac{T\eta_0}{2\|W_{\text{ov}}^*\|^2} - \sqrt{24T\eta_0\|W_{\text{ov}}^*\|\log(2|\mathcal{V}|)\log T} \\
&\quad \geq \frac{T\eta_0}{4\|W_{\text{ov}}^*\|^2},
\end{aligned}$$

where (a) follows from Proposition 1, and (b) is due to Lemma 1. On the other hand, we also have

$$\begin{aligned}
(e_{\text{In}(x)} - e_i)^\top W_{\text{ov}}^{(T)} x &\leq \frac{\|W_{\text{ov}}^{(T)}\|}{\|W_{\text{ov}}^*\|} + 2\sqrt{2}\|W_{\text{ov}}^{(T)}\| \sqrt{\frac{12\|W_{\text{ov}}^*\|^3 \log(2|\mathcal{V}|)\log T}{T\eta_0}} \\
&\leq \frac{3T\eta_0}{4\|W_{\text{ov}}^*\|^2}
\end{aligned}$$

The proof is finished. ■

Thus, for simplicity, we further assume that $(e_{\text{In}(x)} - e_i)W_{\text{ov}}^{(T)} x = \Delta$ for all x, because $(e_{\text{In}(x)} - e_i)\hat{W}_{\text{ov}} x = \Theta(\Delta)$ for large enough iteration. Next, we provide the general form of the projection of the gradient of Key-Query matrix W_{kq} follows from a notation for the token weight.

The token weight $\varphi_\ell^{(n,t)}$ of the token $x_\ell^{(n)}$ in the sentence $X^{(n)} = [x_1^{(n)}, \ldots, x_L^{(n)}]$ under $\theta = (W_{\text{ov}}^{(T)}, W_{\text{kq}})$ is calculated as

$$\varphi_\ell^{(n,t)} = \frac{\exp\left((x_\ell^{(n)})^\top W_{\text{kq}} X_{-1}^{(n)}\right)}{\sum_{\ell'=1}^L \exp\left((x_{\ell'}^{(n)})^\top W_{\text{kq}} X_{-1}^{(n)}\right)} \tag{9}$$

Lemma 4 (Projection of gradient of W_{kq}) *If* W_{ov} *satisfies that*

$$\begin{cases}(e_{\text{In}(x)} - e_i)^\top W_{\text{ov}} x = \Delta, \forall i \neq \text{In}(x) \\ (e_i - e_{i'})^\top W_{\text{ov}} x = 0, \forall i, i' \neq \text{In}(x)\end{cases}$$

we have

$$\begin{aligned}
&\left\langle \nabla_{W_{\text{kq}}} \mathcal{L}(\theta), W'_{\text{kq}} \right\rangle \\
&\quad = \Delta \sum_n \pi^{(n)} ([\mathrm{T}_\theta^{(n)}]_{\text{In}(X^{(n)})} - 1) \sum_{\ell_* \in l(n)} \varphi_{\ell_*}^{(n,\theta)} \left(x_{\ell_*}^{(n)} - \sum_{\ell'} \varphi_{\ell'}^{(n,\theta)} x_{\ell'}^{(n)}\right)^\top W'_{\text{kq}} X_{-1}^{(n)} \\
&\qquad + \Delta \sum_n \pi^{(n)} \sum_{\ell \notin l(n)} [\mathrm{T}_\theta^{(n)}]_{\text{In}(x_\ell^{(n)})} \varphi_\ell^{(n,\theta)} \left(x_\ell^{(n)} - \sum_{\ell'} \varphi_{\ell'}^{(n,\theta)} x_{\ell'}^{(n)}\right)^\top W'_{\text{kq}} X_{-1}^{(n)}.
\end{aligned}$$

Proof. Recall that $l(n)$ is the set of indices of the optimal tokens in the sample $X^{(n)}$. Thus, for any $\ell_* \in l(n)$, $\mathrm{In}(x_{\ell_*}^{(n)}) = \mathrm{In}(X^{(n)})$. In addition, we denote $\mathrm{T}_\theta(X^{(n)})$ by $\mathrm{T}_\theta^{(n)}$ for simplicity. From Equation (6), we have

$$
\begin{aligned}
&\left\langle \nabla_{W_{\mathrm{kq}}} \mathcal{L}(\theta), W'_{\mathrm{kq}} \right\rangle \\
&= \sum_n \pi^{(n)} \sum_\ell (\mathrm{T}_\theta^{(n)} - p^{(n)})^\top W_{\mathrm{ov}} x_\ell^{(n)} \varphi_\ell^{(n,\theta)} \left(x_\ell^{(n)} - \sum_{\ell'} \varphi_{\ell'}^{(n,\theta)} x_{\ell'}^{(n)} \right)^\top W'_{\mathrm{kq}} X_{-1}^{(n)} \\
&\overset{(a)}{=} \sum_n \pi^{(n)} \sum_\ell \sum_i [\mathrm{T}_\theta^{(n)}]_i (e_i - e_{\mathrm{In}(X^{(n)})})^\top W_{\mathrm{ov}} x_\ell^{(n)} \varphi_\ell^{(n,\theta)} \left(x_\ell^{(n)} - \sum_{\ell'} \varphi_{\ell'}^{(n,\theta)} x_{\ell'}^{(n)} \right)^\top W'_{\mathrm{kq}} X_{-1}^{(n)} \\
&= \sum_n \pi^{(n)} \sum_{\ell \in l(n)} \sum_i [\mathrm{T}_\theta^{(n)}]_i (e_i - e_{\mathrm{In}(X^{(n)})})^\top W_{\mathrm{ov}} x_\ell^{(n)} \varphi_\ell^{(n,\theta)} \left(x_\ell^{(n)} - \sum_{\ell'} \varphi_{\ell'}^{(n,\theta)} x_{\ell'}^{(n)} \right)^\top W'_{\mathrm{kq}} X_{-1}^{(n)} \\
&\quad + \sum_n \pi^{(n)} \sum_{\ell \notin l(n)} \sum_i [\mathrm{T}_\theta^{(n)}]_i (e_i - e_{\mathrm{In}(X^{(n)})})^\top W_{\mathrm{ov}} x_\ell^{(n)} \varphi_\ell^{(n,\theta)} \left(x_\ell^{(n)} - \sum_{\ell'} \varphi_{\ell'}^{(n,\theta)} x_{\ell'}^{(n)} \right)^\top W'_{\mathrm{kq}} X_{-1}^{(n)} \\
&= \sum_n \pi^{(n)} \sum_{\ell \in l(n)} \sum_{i \neq \mathrm{In}(X^{(n)})} [\mathrm{T}_\theta^{(n)}]_i (-\Delta) \varphi_\ell^{(n,\theta)} \left(x_\ell^{(n)} - \sum_{\ell'} \varphi_{\ell'}^{(n,\theta)} x_{\ell'}^{(n)} \right)^\top W'_{\mathrm{kq}} X_{-1}^{(n)} \\
&\quad + \sum_n \pi^{(n)} \sum_{\ell \notin l(n)} [\mathrm{T}_\theta^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,\theta)} \left(x_\ell^{(n)} - \sum_{\ell'} \varphi_{\ell'}^{(n,\theta)} x_{\ell'}^{(n)} \right)^\top W'_{\mathrm{kq}} X_{-1}^{(n)} \\
&= \Delta \sum_n \pi^{(n)} ([\mathrm{T}_\theta^{(n)}]_{\mathrm{In}(X^{(n)})} - 1) \sum_{\ell_* \in l(n)} \varphi_{\ell_*}^{(n,\theta)} \left(x_{\ell_*}^{(n)} - \sum_{\ell'} \varphi_{\ell'}^{(n,\theta)} x_{\ell'}^{(n)} \right)^\top W'_{\mathrm{kq}} X_{-1}^{(n)} \\
&\quad + \Delta \sum_n \pi^{(n)} \sum_{\ell \notin l(n)} [\mathrm{T}_\theta^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,\theta)} \left(x_\ell^{(n)} - \sum_{\ell'} \varphi_{\ell'}^{(n,\theta)} x_{\ell'}^{(n)} \right)^\top W'_{\mathrm{kq}} X_{-1}^{(n)},
\end{aligned}
$$

where (a) is due to the fact that $\sum_{i \in [|\mathcal{V}|]} [\mathrm{T}_\theta^{(n)}]_i = 1$ for any θ, n.

■

Main Steps

The proof consists of three main steps. First, we show that the optimal token weight has a lower bound. Then, we show that the gradient aligns with the optimal direction. Third, we show that the norm of Key-Query matrix grows linearly. Combining these three steps, we can prove the Theorem 1 and Theorem 2.

Recall that the updating rule for $W_{\mathrm{kq}}^{(t)}$ is

$$
W_{\mathrm{kq}}^{(t+1)} = W_{\mathrm{kq}}^{(t)} - \eta \frac{\nabla_{W_{\mathrm{kq}}} \mathcal{L}(\theta^{(t)})}{\|\nabla_{W_{\mathrm{kq}}} \mathcal{L}(\theta^{(t)})\|}. \tag{10}
$$

C.2 Step 1

We first show that the optiaml token weight has a lower bound during the training.

Lemma 5 (Lower bound of optimal token weight) *Under the zero initialization and updating rule Equation* (10)*, for any iteration t, and any sample $X^{(n)}$, if $l(n)$ is the set of indices of the optimal token in $X^{(n)}$, the following inequality holds.*

$$
\varphi_\ell^{(n,t)} \geq \varphi_\ell^{(n,0)} \geq 1/L_{\max}, \quad \forall \ell \in l(n)
$$

Proof.

First, we introduce the notation that $\varphi_+^{(n,t)} = \sum_{\ell_* \in l(n)} \varphi_{\ell_*}^{(n,t)}$ as the summation of optimal token weights, and $\varphi_-^{(n,t)} = 1 - \varphi_+^{(n,t)}$ as the summation of non-optimal token weights.

At $t = 0$, due to zero initialization, we have $\varphi_\ell^{(n,0)} = 1/L^{(n)}$. Moreover, by Assumption 3, for any $\ell \notin l(n)$, we have $\varphi_+^{(n,0)} \geq q_n(x_\ell)\varphi_\ell^{(n,0)}$.

We perform induction the hypothesis: $\varphi_+^{(n,t)} \geq \varphi_+^{(n,t-1)}$ and $\varphi_{\ell_*}^{(n,t)} \geq \varphi_\ell^{(n,t)}$ for all $\ell_* \in l(n)$ and $\ell \notin l(n)$.

Suppose the hypothesis holds for iteration t. Let $x_{\ell_*}^{(n)}$ be the optimal token in the sequence $X^{(n)}$.

Fix a sample $X^{(n')}$. we have

$$
\begin{aligned}
&(x_{\ell_*}^{(n')})^\top \nabla_{W_{\mathrm{kq}}} \mathcal{L}^{(t)}(\theta^{(t)}) X_{-1}^{(n')} \\
&\quad = \sum_n \pi^{(n)} ([\mathrm{T}_\theta^{(n)}]_{\mathrm{In}(X^{(n)})} - 1) \varphi_+^{(n,t)} \left(\sum_{\ell' \notin l(n)} \varphi_{\ell'}^{(n,t)} (x_{\ell_*}^{(n)} - x_{\ell'}^{(n)})^\top x_{\ell_*}^{(n')} \right) \left\langle X_{-1}^{(n)}, X_{-1}^{(n')} \right\rangle \\
&\qquad + \sum_n \pi^{(n)} \sum_{\ell \notin l(n)} [\mathrm{T}_\theta^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,t)} \left(\sum_{\ell'} \varphi_{\ell'}^{(n,t)} (x_\ell^{(n)} - x_{\ell'}^{(n)})^\top x_{\ell_*}^{(n')} \right) \left\langle X_{-1}^{(n)}, X_{-1}^{(n')} \right\rangle
\end{aligned}
$$

Because $\sum_{\ell' \notin l(n)} (x_{\ell'}^{(n)}) x_{\ell_*}^{(n')} = 0$ due to Assumption 1, and $(x_{\ell_*}^{(n)})^\top x_{\ell_*}^{(n')} \geq 0$, we immediately have

$$
(x_{\ell_*}^{(n')})^\top \nabla_{W_{\mathrm{kq}}} \mathcal{L}^{(t)}(\theta^{(t)}) X_{-1}^{(n')} \leq 0.
$$

Let $x_{\ell_0}^{(n')}$ be any non-optiaml token in the sequence $X^{(n')}$. Then, we have

$$
\begin{aligned}
&(x_{\ell_0}^{(n')})^\top \nabla_{W_{\mathrm{kq}}} \mathcal{L}^{(t)}(\theta^{(t)}) X_{-1}^{(n')} \\
&= \sum_n \pi^{(n)} ([\mathrm{T}_\theta^{(n)}]_{\mathrm{In}(X^{(n)})} - 1) \varphi_+^{(n,t)} \left(\sum_{\ell' \notin l(n)} \varphi_{\ell'}^{(n,t)} (x_{\ell_*}^{(n)} - x_{\ell'}^{(n)})^\top x_{\ell_0}^{(n')} \right) \left\langle X_{-1}^{(n)}, X_{-1}^{(n')} \right\rangle \\
&\quad + \sum_n \pi^{(n)} \sum_{\ell \notin l(n)} [\mathrm{T}_\theta^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,t)} \left(\sum_{\ell'} \varphi_{\ell'}^{(n,t)} (x_\ell^{(n)} - x_{\ell'}^{(n)})^\top x_{\ell_0}^{(n')} \right) \left\langle X_{-1}^{(n)}, X_{-1}^{(n')} \right\rangle \\
&= \sum_n \pi^{(n)} ([\mathrm{T}_\theta^{(n)}]_{\mathrm{In}(X^{(n)})} - 1) \varphi_+^{(n,t)} \left(\sum_{\ell' \notin l(n)} \varphi_{\ell'}^{(n,t)} (-x_{\ell'}^{(n)})^\top x_{\ell_0}^{(n')} \right) \left\langle X_{-1}^{(n)}, X_{-1}^{(n')} \right\rangle \\
&\quad + \sum_n \pi^{(n)} \sum_{\ell \notin l(n)} [\mathrm{T}_\theta^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,t)} \left(\sum_{\ell' \notin l(n)} \varphi_{\ell'}^{(n,t)} (x_\ell^{(n)} - x_{\ell'}^{(n)})^\top x_{\ell_0}^{(n')} \right) \left\langle X_{-1}^{(n)}, X_{-1}^{(n')} \right\rangle \\
&\geq \sum_n \pi^{(n)} \left((1 - [\mathrm{T}_\theta^{(n)}]_{\mathrm{In}(X^{(n)})}) \varphi_{\ell_+}^{(n,t)} - \sum_{\ell \notin l(n)} [\mathrm{T}_\theta^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,t)} \right) \left(\sum_{\ell' \notin l(n)} \varphi_{\ell'}^{(n,t)} (x_{\ell'}^{(n)})^\top x_{\ell_0}^{(n')} \right) \left\langle X_{-1}^{(n)}, X_{-1}^{(n')} \right\rangle \\
&\geq 0,
\end{aligned}
$$

where the last inequality is due to Assumption 3, the induction hypothesis and $\sum_{i \in [|\mathcal{V}|]} [\mathrm{T}_\theta^{(n)}]_i = 1$

Therefore, for any n, we have

$$
\varphi_{\ell_*}^{(n,t+1)} = \frac{\exp\left((x_{\ell_*}^{(n)})^\top W_{\mathrm{kq}}^{(t+1)} X_{-1}^{(n)} \right)}{\sum_\ell \exp\left((x_\ell^{(n)})^\top W_{\mathrm{kq}}^{(t+1)} X_{-1}^{(n)} \right)}
$$

$$
\begin{aligned}
&= \frac{\exp\left((x_{\ell_*}^{(n)})^\top W_{\mathrm{kq}}^{(t)} X_{-1}^{(n)} - \eta (x_{\ell_*}^{(n)})^\top \nabla_{W_{\mathrm{kq}}} \mathcal{L}^{(t)}(\theta^{(t)}) X_{-1}^{(n)} / \|\nabla_{W_{\mathrm{kq}}} \mathcal{L}^{(t)}(\theta^{(t)})\|\right)}{\sum_\ell \exp\left((x_\ell^{(n)})^\top W_{\mathrm{kq}}^{(t)} X_{-1}^{(n)} - \eta (x_\ell^{(n)})^\top \nabla_{W_{\mathrm{kq}}} \mathcal{L}^{(t)}(\theta^{(t)}) X_{-1}^{(n)} / \|\nabla_{W_{\mathrm{kq}}} \mathcal{L}^{(t)}(\theta^{(t)})\|\right)} \\
&= \frac{\exp\left((x_{\ell_*}^{(n)})^\top W_{\mathrm{kq}}^{(t)} X_{-1}^{(n)}\right)}{\sum_\ell \exp\left((x_\ell^{(n)})^\top W_{\mathrm{kq}}^{(t)} X_{-1}^{(n)} + \eta (x_{\ell_*}^{(n)} - x_\ell^{(n)})^\top \nabla_{W_{\mathrm{kq}}} \mathcal{L}^{(t)}(\theta^{(t)}) X_{-1}^{(n)} / \|\nabla_{W_{\mathrm{kq}}} \mathcal{L}^{(t)}(\theta^{(t)})\|\right)} \\
&\geq \frac{\exp\left((x_{\ell_*}^{(n)})^\top W_{\mathrm{kq}}^{(t)} X_{-1}^{(n)}\right)}{\sum_\ell \exp\left((x_\ell^{(n)})^\top W_{\mathrm{kq}}^{(t)} X_{-1}^{(n)}\right)} \\
&= \varphi_{\ell_*}^{(n,t)},
\end{aligned}
$$

which implies that $\varphi_+^{(n,t+1)} \geq \varphi_+^{(n,t)}$.

For the second argument in the hypothesis, we examine $\varphi_{\ell_*}^{(n,t+1)} / \varphi_\ell^{(n,t+1)}$ for any $\ell \notin l(n)$. We have

$$
\begin{aligned}
\frac{\varphi_{\ell_*}^{(n,t+1)}}{\varphi_\ell^{(n,t+1)}} &= \exp\left((x_{\ell_*}^{(n)} - x_\ell^{(n)})^\top W_{\mathrm{kq}}^{(t+1)} X_{-1}^{(n)}\right) \\
&= \exp\left((x_{\ell_*}^{(n)} - x_\ell^{(n)})^\top W_{\mathrm{kq}}^{(t)} X_{-1}^{(n)}\right) \exp\left(-\frac{\eta}{\|\nabla_{W_{\mathrm{kq}}} \mathcal{L}^{(t)}(\theta^{(t)})\|} (x_{\ell_*}^{(n)} - x_\ell^{(n)})^\top \nabla_{W_{\mathrm{kq}}} \mathcal{L}^{(t)}(\theta^{(t)}) X_{-1}^{(n)}\right) \\
&\geq \frac{\varphi_{\ell_*}^{(n,t)}}{\varphi_\ell^{(n,t)}} \\
&\geq 1.
\end{aligned}
$$

The proof is finished.

■

C.3 Step 2

The following lemma shows that the norm of the Key-Query Matrix increases linearly with the number of iterations.

Lemma 6 *Under the initialization $W_{\mathrm{kq}}^{(0)}$ and the updating rule Equation* (10)*, for each iteration t, the following inequality holds.*

$$
t\eta + \|W_{\mathrm{kq}}^{(0)}\| \geq \|W_{\mathrm{kq}}^{(t)}\| \geq \frac{t\eta}{2L_{\max}\|W_{\mathrm{kq}}^*\|} - \|W_{\mathrm{kq}}^{(0)}\|.
$$

Proof.

We examine the gradient $\nabla_{W_{\mathrm{kq}}} \mathcal{L}(\theta)$ projected onto the optimal direction $W_{\mathrm{kq}}^* / \|W_{\mathrm{kq}}^*\|$.

$$
\begin{aligned}
&\left\langle \nabla_{W_{\mathrm{kq}}} \mathcal{L}(\theta^{(t)}), W_{\mathrm{kq}}^* \right\rangle \\
&= \sum_n \pi^{(n)} \Delta \sum_{\ell_* \in l(n)} ([\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_{\ell_*}^{(n)})} - 1) \varphi_{\ell_*}^{(n,t)} \left(a_{\ell_*}^{(n,*)} - \sum_{\ell'} \varphi_{\ell'}^{(n,t)} a_{\ell'}^{(n,*)} \right) \\
&\quad + \sum_n \pi^{(n)} \Delta \sum_{\ell \notin l(n)} [\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,t)} \left(a_\ell^{(n,*)} - \sum_{\ell'} \varphi_{\ell'}^{(n,t)} a_{\ell'}^{(n,*)} \right) \\
&= \sum_n \pi^{(n)} \Delta ([\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(X^{(n)})} - 1) \varphi_+^{(n,t)} \left(\sum_{\ell' \notin l(n)} \varphi_{\ell'}^{(n,t)} (a_{\ell_*}^{(n,*)} - a_{\ell'}^{(n,*)}) \right)
\end{aligned}
$$

$$
\begin{aligned}
&+\sum_n \pi^{(n)}\Delta \sum_{\ell\notin l(n)} [\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_\ell)}\varphi^{(n,t)}_\ell \left(\sum_{\ell'\in l(n)} \varphi^{(n,t)}_{\ell'}(a^{(n,*)}_\ell - a^{(n,*)}_{\ell'})\right)\\
&\le \sum_n \pi^{(n)}\Delta([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(X^{(n)})}-1)\varphi^{(n,t)}_+\varphi^{(n,t)}_-\\
&\quad+\sum_n \pi^{(n)}\Delta \sum_{\ell\notin l(n)} [\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_\ell)}\varphi^{(n,t)}_\ell(-\varphi^{(n,t)}_+),
\end{aligned}
$$

where $\varphi^{(n,t)}_+ = \sum_{\ell_*\in l(n)}\varphi^{(n,t)}_{\ell_*}$ is the summation of optimal token weights, and $\varphi^{(n,t)}_- = 1-\varphi^{(n,t)}_+$ is the summation of non-optimal token weights.

On the other hand,

$$
\begin{aligned}
&\|\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta^{(t)})\|\\
&\quad=\left\langle \nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta^{(t)}), \frac{\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta^{(t)})}{\|\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta^{(t)})\|}\right\rangle\\
&\quad=\sum_n \pi^{(n)}\Delta\sum_{\ell_*\in l(n)}([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell_*})}-1)\varphi^{(n,t)}_{\ell_*}\left(x^{(n)}_{\ell_*}-\sum_{\ell'}\varphi^{(n,t)}_{\ell'}x^{(n)}_{\ell'}\right)^\top \frac{\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta^{(t)})}{\|\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta^{(t)})\|}X^{(n)}_{-1}\\
&\qquad+\sum_n \pi^{(n)}\Delta\sum_{\ell\notin l(n)}[\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_\ell)}\varphi^{(n,t)}_\ell\left(x^{(n)}_\ell-\sum_{\ell'}\varphi^{(n,t)}_{\ell'}x^{(n)}_{\ell'}\right)^\top \frac{\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta^{(t)})}{\|\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta^{(t)})\|}X^{(n)}_{-1}\\
&\quad\le 2\sum_n \pi^{(n)}\Delta(1-[\mathrm{T}^{(n)}_\theta]_{\mathrm{In}(X^{(n)})})\varphi^{(n,t)}_+\varphi^{(n,t)}_- + 2\sum_n \pi^{(n)}\Delta\sum_{\ell\notin l(n)}[\mathrm{T}^{(n)}_\theta]_{\mathrm{In}(x^{(n)}_\ell)}\varphi^{(n,t)}_\ell
\end{aligned}
$$

Thus, we have

$$
\left\langle \frac{\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta)}{\left\|\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta)\right\|}, \frac{W^*_{\mathrm{kq}}}{\|W^*_{\mathrm{kq}}\|}\right\rangle \le -\frac{\min_n \varphi^{(n,t)}_+}{2\|W^*_{\mathrm{kq}}\|} \le -\frac{1}{2L_{\max}\|W^*_{\mathrm{kq}}\|}
$$

By the updating rule Equation (10), we have

$$
\begin{aligned}
\|W^{(t)}_{\mathrm{kq}}\| &= \left\|W^{(0)}_{\mathrm{kq}} - \sum_{t'=0}^{t-1}\eta\frac{\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta^{(t')})}{\left\|\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta^{(t')})\right\|}\right\|\\
&\ge \left\langle W^{(0)}_{\mathrm{kq}}, \frac{W^*_{\mathrm{kq}}}{\|W^*_{\mathrm{kq}}\|}\right\rangle - \sum_{t'\le t-1}\eta\left\langle \frac{\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta^{(t')})}{\left\|\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta^{(t')})\right\|}, \frac{W^*_{\mathrm{kq}}}{\|W^*_{\mathrm{kq}}\|}\right\rangle\\
&\ge \sum_{t'<t}\frac{\eta}{2L_{\max}\|W^*_{\mathrm{kq}}\|} - \|W^{(0)}_{\mathrm{kq}}\|\\
&= \frac{t\eta}{2L_{\max}\|W^*_{\mathrm{kq}}\|} - \|W^{(0)}_{\mathrm{kq}}\|.
\end{aligned}
$$

In addition, by the triangle inequality,

$$
\begin{aligned}
\|W^{(t)}_{\mathrm{kq}}\| &= \left\|W^{(0)}_{\mathrm{kq}} - \sum_{t'=0}^{t-1}\eta\frac{\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta^{(t')})}{\left\|\nabla_{W_{\mathrm{kq}}}\mathcal{L}(\theta^{(t')})\right\|}\right\|\\
&\le t\eta + \|W^{(0)}_{\mathrm{kq}}\|.
\end{aligned}
$$

The proof is completed. ■

C.4 Step 3

We next show that the gradient $\nabla_{W_{\text{kq}}}\mathcal{L}(\theta^{(t)})$ is close to the optimal direction W^*_{kq}.

Lemma 7 (Gradient aligns with the optimal direction) *Let* $t_0 = \lceil \frac{8L_{\max}\|W^*_{\text{kq}}\|^2}{\eta} \rceil$. *Then, for any* $t \geq t_0$, *we have*

$$\left\langle \nabla_{W_{\text{kq}}}\mathcal{L}(\theta^{(t)}), W^{(t)}_{\text{kq}} \right\rangle \geq (1+\alpha_t)\left\langle \nabla_{W_{\text{kq}}}\mathcal{L}(\theta^{(t)}), W^*_{\text{kq}} \right\rangle \frac{\|W^{(t)}_{\text{kq}}\|}{\|W^*_{\text{kq}}\|}$$

where

$$\alpha_t = \frac{4NL^2_{\max}\|W^*_{\text{kq}}\|^2}{\|W^{(t)}_{\text{kq}}\|}\left(1+\log\left(2L_{\max}\|W^{(t)}_{\text{kq}}\|\right)\right)$$

Proof. During the proof, we denote

$$\begin{cases} a^{(n,t)}_\ell = (x^{(n)}_\ell)^\top W^{(t)}_{\text{kq}} X^{(n)}_{-1} \\ a^{(n,*)}_\ell = (x^{(n)}_\ell)^\top W^*_{\text{kq}} X^{(n)}_{-1} \frac{\|W^{(t)}_{\text{kq}}\|}{\|W^*_{\text{kq}}\|} \end{cases}$$

$$\beta_0 = \frac{2L^2_{\max}\|W^*_{\text{kq}}\|^2}{\|W^{(t)}_{\text{kq}}\|}(1+\log(2L_{\max}\|W^{(t)}_{\text{kq}}\|)).$$

We point out a few facts that will be frequently used in the proof.

If $a^{(n,t)}_\ell \leq a^{(n,t)}_{\ell'} - C_0$, then we have

$$\varphi^{(n,t)}_\ell = \varphi^{(n,t)}_{\ell'} \exp\left(a^{(n,t)}_\ell - a^{(n,t)}_{\ell'}\right) \leq \exp(-C_0) \tag{11}$$

The same result holds if $a^{(n,t)}_{\ell'}$ is replaced any convex combination of a set of $a^{(n,t)}_{\ell'}$'s.

We start the proof by noting that W^*_{kq} is the minimum unique solution to the problem

$$W^*_{\text{kq}} = \arg\min \|W\|, \quad \text{s.t.} \quad (x^{(n)}_{\ell_*} - x^{(n)}_\ell) W X^{(n)}_{-1} \geq 1, \quad \forall \ell_* \in l^{(n)}, \ell \notin l^{(n)}, \forall n. \tag{12}$$

Therefore, if $W^{(t)}_{\text{kq}} \frac{\|W^*_{\text{kq}}\|}{\|W^{(t)}_{\text{kq}}\|} = W^*_{\text{kq}}$, the results is trivial since $\left\langle \nabla_{W_{\text{kq}}}\mathcal{L}(\theta^{(t)}), W^*_{\text{kq}} \right\rangle \leq 0$.

In the following, we focus on the case when $W^{(t)}_{\text{kq}} \frac{\|W^*_{\text{kq}}\|}{\|W^{(t)}_{\text{kq}}\|} \neq W^*_{\text{kq}}$. Then, there must be at least a sentence $X^{(n)}$, such that $W^{(t)}_{\text{kq}} \frac{\|W^*_{\text{kq}}\|}{\|W^{(t)}_{\text{kq}}\|}$ violates the contraint on $X^{(n)}$. In other words, we must have

$$a^{(n,t)}_{\ell_*} - a^{(n,t)}_\ell = (x^{(n)}_{\ell_*} - x^{(n)}_\ell) W^{(t)}_{\text{kq}} X^{(n)}_{-1} \leq \frac{\|W^{(t)}_{\text{kq}}\|}{\|W^*_{\text{kq}}\|}.$$

This implies that for those n, we must have $\varphi^{(n,t)}_\ell \geq \exp(-\frac{\|W^{(t)}_{\text{kq}}\|}{\|W^*_{\text{kq}}\|})$

Thus, we consider two types of samples in the folloiwng.

Type 1. Let us consider $X^{(n)}$ such that $\varphi^{(n,t)}_- \geq \exp(-(1+\beta_0/2)\|W^{(t)}_{\text{kq}}\|/\|W^*_{\text{kq}}\|)$.

Recall that the inner product between the gradient $\nabla_{W_{\text{kq}}}\mathcal{L}(\theta^{(t)})$ and any other Key-Query matrix $\theta' = W'_{\text{kq}}$ has the following form (Lemma 4).

$$\left\langle \nabla_{W_{\text{kq}}}\mathcal{L}(\theta^{(t)}), W'_{\text{kq}} \right\rangle$$

$$= \Delta \sum_n \pi^{(n)} ([\mathrm{T}_\theta^{(n)}]_{\mathrm{In}(X^{(n)})} - 1) \sum_{\ell_* \in l(n)} \varphi_{\ell_*}^{(n,\theta)} \left(x_{\ell_*}^{(n)} - \sum_{\ell'} \varphi_{\ell'}^{(n,\theta)} x_{\ell'}^{(n)} \right)^\top W'_{\mathrm{kq}} X_{-1}^{(n)}$$

$$+ \Delta \sum_n \pi^{(n)} \sum_{\ell \notin l(n)} [\mathrm{T}_\theta^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,\theta)} \left(x_\ell^{(n)} - \sum_{\ell'} \varphi_{\ell'}^{(n,\theta)} x_{\ell'}^{(n)} \right)^\top W'_{\mathrm{kq}} X_{-1}^{(n)}$$

Let $L_n(\theta) = -\log e_{\mathrm{In}(X^{(n)})}^\top \mathrm{T}_\theta(X^{(n)})$ be the loss on sample $X^{(n)}$.

To proceed, we examine the gradient on each sample $X^{(n)}$ with $\varphi_-^{(n,t)} \geq \exp(-(1 + \beta_0/2)\|W_{\mathrm{kq}}^{(t)}\|/\|W_{\mathrm{kq}}^*\|)$, which can be divided into two parts. $\left\langle \nabla_{W_{\mathrm{kq}}} L_n(\theta^{(t)}), W_{\mathrm{kq}} \right\rangle = \Delta(A^{(n,t)} + B^{(n,t)})$, where

$$\begin{cases} A^{(n,t)} = \sum_{\ell_* \in l(n)} ([\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_{\ell_*}^{(n)})} - 1) \varphi_{\ell_*}^{(n,t)} \left(a_{\ell_*}^{(n,t)} - \sum_{\ell'} \varphi_{\ell'}^{(n,t)} a_{\ell'}^{(n,t)} \right), \\ B^{(n,t)} = \sum_{\ell \notin l(n)} [\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,t)} \left(a_\ell^{(n,t)} - \sum_{\ell'} \varphi_{\ell'}^{(n,t)} a_{\ell'}^{(n,t)} \right). \end{cases}$$

We further let

$$A^{(n,*)} = \sum_{\ell_* \in l(n)} ([\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{I_n}(x_\ell^{(n)})} - 1) \varphi_{\ell_*}^{(n,t)} \left(a_{\ell_*}^{(n,*)} - \sum_{\ell'} \varphi_{\ell'}^{(n,t)} a_{\ell'}^{(n,*)} \right),$$

and

$$B^{(n,*)} = \sum_{\ell \notin l(n)} [\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{I_n}(x_\ell^{(n)})} \varphi_\ell^{(n,t)} \left(a_\ell^{(n,*)} - \sum_{\ell'} \varphi_{\ell'}^{(n,t)} a_{\ell'}^{(n,*)} \right).$$

Thus, we aim to find the relationship $A^{(n,t)} + B^{(n,t)}$ between $A^{(n,*)} + B^{(n,*)}$.

We first provide the upper bounds for $A^{(n,*)}$ and $B^{(n,*)}$.

$$\begin{aligned} A^{(n,*)} &= \sum_{\ell_* \in l(n)} ([\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_{\ell_*}^{(n)})} - 1) \varphi_{\ell_*}^{(n,t)} \left(a_{\ell_*}^{(n,*)} - \sum_{\ell'} \varphi_{\ell'}^{(n,t)} a_{\ell'}^{(n,*)} \right) \\ &= \sum_{\ell_* \in l(n)} ([\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_{\ell_*}^{(n)})} - 1) \varphi_{\ell_*}^{(n,t)} \left(\sum_{\ell' \notin l(n)} \varphi_{\ell'}^{(n,t)} (a_{\ell_*}^{(n,*)} - a_{\ell'}^{(n,*)}) \right) \\ &\overset{(a)}{\leq} ([\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(X^{(n)})} - 1) \varphi_+^{(n,t)} \varphi_-^{(n,t)} \frac{\|W_{\mathrm{kq}}^{(t)}\|}{\|W_{\mathrm{kq}}^*\|}, \end{aligned}$$

where (a) is due to the fact that $(x_{\ell_*}^{(n)} - x_\ell^{(n)}) W_{\mathrm{kq}}^* X_{-1}^{(n)} \geq 1$, and $a_\ell^{(n,*)} = (x_\ell^{(n)})^\top W_{\mathrm{kq}}^* X_{-1}^{(n)} \|W_{\mathrm{kq}}^{(t)}\|/\|W_{\mathrm{kq}}^*\|$.

On the other hand

$$\begin{aligned} A^{(n,t)} &= \sum_{\ell \in l(n)} ([\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} - 1) \varphi_\ell^{(n,t)} \left(a_\ell^{(n,t)} - \sum_{\ell'} \varphi_{\ell'}^{(n,t)} a_{\ell'}^{(n,t)} \right) \\ &= \sum_{\ell \in l(n)} ([\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} - 1) \varphi_\ell^{(n,t)} \left(\sum_{\ell' \notin l(n)} \varphi_{\ell'}^{(n,t)} (a_\ell^{(n,t)} - a_{\ell'}^{(n,t)}) \right) \end{aligned}$$

$$
\begin{aligned}
&= \max_{\mathtt{T}} \left\{ \sum_{\ell \in l(n)} ([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell})} - 1)\varphi^{(n,t)}_{\ell} \left(\sum_{\substack{\ell' \notin l(n) \\ \mathrm{diff} < \mathtt{T}}} \varphi^{(n,t)}_{\ell'} \underbrace{(a^{(n,t)}_{\ell} - a^{(n,t)}_{\ell'})}_{\mathrm{diff}} \right) \right. \\
&\quad \left. + \sum_{\ell \in l(n)} ([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell})} - 1)\varphi^{(n,t)}_{\ell} \left(\sum_{\substack{\ell' \notin l(n) \\ \mathrm{diff} > \mathtt{T}}} \varphi^{(n,t)}_{\ell'} \underbrace{(a^{(n,t)}_{\ell} - a^{(n,t)}_{\ell'})}_{\mathrm{diff}} \right) \right\}. \\
&\overset{(a)}{\geq} \max_{\mathtt{T}} \left\{ \sum_{\ell \in l(n)} ([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell})} - 1)\varphi^{(n,t)}_{\ell} \left(\sum_{\ell' \notin l(n)} \varphi^{(n,t)}_{\ell'} \mathtt{T} \right) \right. \\
&\quad \left. + \sum_{\ell \in l(n)} ([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell})} - 1)\varphi^{(n,t)}_{\ell} \left(2 \sum_{\ell' \notin l(n)} \exp(-\mathtt{T}) \|W^{(t)}_{\mathrm{kq}}\| \right) \right\} \\
&\geq \max_{\mathtt{T}} \left\{ \sum_{\ell \in l(n)} ([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell})} - 1)\varphi^{(n,t)}_{\ell} \left(\varphi^{(n,t)}_{-} \mathtt{T} + 2L_{\max} \exp(-\mathtt{T}) \|W^{(t)}_{\mathrm{kq}}\| \right) \right\} \\
&\overset{(b)}{\geq} \sum_{\ell \in l(n)} ([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell})} - 1)\varphi^{(n,t)}_{\ell} \varphi^{(n,t)}_{-} \left(1 + \log \frac{2L_{\max} \|W^{(t)}_{\mathrm{kq}}\|}{\varphi^{(n,t)}_{-}} \right),
\end{aligned}
$$

where (a) is due to Equation (11), and (b) is obtained by choosing $\mathtt{T} = \log \frac{2L_{\max}\|W^{(t)}_{\mathrm{kq}}\|}{\varphi^{(n,t)}_{-}}$.

Recall that $\varphi^{(n,t)}_{-} \geq \exp\left(-(1+\beta_0/2)\|W^{(t)}_{\mathrm{kq}}\|/\|W^{*}_{\mathrm{kq}}\|\right)$ and $\beta_0 \geq \frac{2\|W^{*}_{\mathrm{kq}}\|(1+\log(2L_{\max}\|W^{(t)}_{\mathrm{kq}}\|))}{\|W^{(t)}_{\mathrm{kq}}\|}$. Thus, we further have

$$
\begin{aligned}
A^{(n,t)} &\geq ([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(X^{(n)})} - 1)\varphi^{(n,t)}_{+} \varphi^{(n,t)}_{-} \left(1 + \log \frac{2L_{\max}\|W^{(t)}_{\mathrm{kq}}\|}{\varphi^{(n,t)}_{-}} \right) \\
&\geq ([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(X^{(n)})} - 1)\varphi^{(n,t)}_{+} \varphi^{(n,t)}_{-} \left(1 + \log(2L_{\max}\|W^{(t)}_{\mathrm{kq}}\|) + (1+\beta_0/2)\frac{\|W^{(t)}_{\mathrm{kq}}\|}{\|W^{*}_{\mathrm{kq}}\|} \right) \\
&\geq ([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(X^{(n)})} - 1)\varphi^{(n,t)}_{+} \varphi^{(n,t)}_{-} \left(\frac{\beta_0\|W^{(t)}_{\mathrm{kq}}\|}{2\|W^{*}_{\mathrm{kq}}\|} + (1+\beta_0/2)\frac{\|W^{(t)}_{\mathrm{kq}}\|}{\|W^{*}_{\mathrm{kq}}\|} \right) \\
&= (1+\beta_0)([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(X^{(n)})} - 1)\varphi^{(n,t)}_{+} \varphi^{(n,t)}_{-} \frac{\|W^{(t)}_{\mathrm{kq}}\|}{\|W^{*}_{\mathrm{kq}}\|} \\
&\geq (1+\beta_0)A^{(n,*)}
\end{aligned}
$$

Next, we analyze $B^{(n,t)}$, and further divide $B^{(n,\theta)}$ into $B^{(n,\theta)}_{+}$ and $B^{(n,\theta)}_{-}$:

$$
\begin{cases}
B^{(n,\theta)}_{+} = \sum_{\ell \notin l(n)} [\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell})} \varphi^{(n,t)}_{\ell} \left(\varphi^{(n,t)}_{+} a^{(n,\theta)}_{\ell} - \sum_{\ell' \in l(n)} \varphi^{(n,t)}_{\ell'} a^{(n,\theta)}_{\ell'} \right) \\
B^{(n,\theta)}_{-} = \sum_{\ell \notin l(n)} [\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell})} \varphi^{(n,t)}_{\ell} \left(\varphi^{(n,t)}_{-} a^{(n,\theta)}_{\ell} - \sum_{\ell' \notin l(n)} \varphi^{(n,t)}_{\ell'} a^{(n,\theta)}_{\ell'} \right)
\end{cases}
$$

Due to Proposition 4, we have $B^{(n,*)}_{-} = 0$, and thus

$$
B^{(n,*)} = B^{(n,*)}_{+} \leq \sum_{\ell \notin l(n)} [\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell})} \varphi^{(n,t)}_{\ell} (-\varphi^{(n,t)}_{+}) \frac{\|W^{(t)}_{\mathrm{kq}}\|}{\|W^{*}_{\mathrm{kq}}\|} \leq 0.
$$

We then analyze:

$$
\begin{aligned}
&B_+^{(n,t)} - (1+\beta_0)B_+^{(n,*)} \\
&\quad= \sum_{\ell \notin l(n)} [\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,t)} \left(\varphi_+^{(n,t)} a_\ell^{(n,t)} - \varphi_+^{(n,t)} a_{\ell_*}^{(n,t)} - (1+\beta_0) \left(\varphi_+^{(n,t)} a_\ell^{(n,*)} - \varphi_+^{(n,t)} a_{\ell_*}^{(n,*)} \right) \right) \\
&\quad= \sum_{\ell \notin l(n)} [\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,t)} \varphi_+^{(n,t)} \left(\underbrace{a_\ell^{(n,t)} - a_{\ell_*}^{(n,t)}}_{b_{t,\ell}} - \underbrace{(1+\beta_0) \left(a_\ell^{(n,*)} - a_{\ell_*}^{(n,*)} \right)}_{b_{*,\ell}} \right) \\
&\quad\geq \sum_{\substack{\ell \notin l(n) \\ b_{t,\ell} < b_{*,\ell}}} [\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,t)} \varphi_+^{(n,t)} \left(b_{t,\ell} - b_{*,\ell} \right) \\
&\quad\overset{(a)}{\geq} \sum_{\ell \notin l(n)} [\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_+^{(n,t)} \exp\left(-(1+\beta_0) \frac{\|W_{\mathrm{kq}}^{(t)}\|}{\|W_{\mathrm{kq}}^*\|} \right) \left(-2\|W_{\mathrm{kq}}^{(t)}\| \right) \\
&\quad= -2 \sum_{\ell \notin l(n)} [\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_+^{(n,t)} \exp\left(-(1+\beta_0/2) \frac{\|W_{\mathrm{kq}}^{(t)}\|}{\|W_{\mathrm{kq}}^*\|} \right) \|W_{\mathrm{kq}}^{(t)}\| \exp\left(-\frac{\beta_0}{2} \frac{\|W_{\mathrm{kq}}^{(t)}\|}{\|W_{\mathrm{kq}}^*\|} \right) \\
&\quad\overset{(b)}{\geq} -2L_{\max}(1 - [\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(X^{(n)})}) \varphi_+^{(n,t)} \varphi_-^{(n,t)} \frac{\|W_{\mathrm{kq}}^{(t)}\|}{\|W_{\mathrm{kq}}^*\|} \exp\left(-\frac{\beta_0}{2} \frac{\|W_{\mathrm{kq}}^{(t)}\|}{\|W_{\mathrm{kq}}^*\|} \right) \|W_{\mathrm{kq}}^*\| \\
&\quad\geq 2L_{\max} \exp\left(-\frac{\beta_0}{2} \frac{\|W_{\mathrm{kq}}^{(t)}\|}{\|W_{\mathrm{kq}}^*\|} \right) \|W_{\mathrm{kq}}^*\| A^{(n,*)} \\
&\quad\overset{(c)}{\geq} \frac{\|W_{\mathrm{kq}}^*\|}{\|W_{\mathrm{kq}}^{(t)}\|} A^{(n,*)} \\
&\quad\geq \beta_0 A^{(n,*)},
\end{aligned}
$$

where (a) follows from that $b_{*,\ell} \leq -(1+\beta_0)\|W_{\mathrm{kq}}^{(t)}\|/\|W_{\mathrm{kq}}^*\|$ and Equation (11), and (b) is due to the fact that $\sum_{i \in [|\mathcal{V}|]} [\mathrm{T}_\theta^{(n)}]_i = 1$ for any θ, n, and (c) follows from $\beta_0/2 \geq \frac{\|W_{\mathrm{kq}}^*\|}{\|W_{\mathrm{kq}}^{(t)}\|} \log(2L_{\max}\|W_{\mathrm{kq}}^{(t)}\|)$

For the term $B_-^{(n,t)}$, we have

$$
\begin{aligned}
B_-^{(n,t)} &= \sum_{\ell \notin l(n)} [\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,t)} \left(\varphi_-^{(n,t)} a_\ell^{(n,t)} - \sum_{\ell' \notin l(n)} \varphi_{\ell'}^{(n,t)} a_{\ell'}^{(n,t)} \right) \\
&= \sum_{\ell \notin l(n)} [\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,t)} \varphi_-^{(n,t)} \left(a_\ell^{(n,t)} - \sum_{\ell' \notin l(n)} \frac{\varphi_{\ell'}^{(n,t)}}{\varphi_-^{(n,t)}} a_{\ell'}^{(n,t)} \right) \\
&= \max_{\mathrm{T}>0} \left\{ \sum_{\substack{\ell \notin l(n) \\ \mathrm{diff} > -\mathrm{T}}} [\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,t)} \varphi_-^{(n,t)} \left(\underbrace{a_\ell^{(n,t)} - \sum_{\ell' \notin l(n)} \frac{\varphi_{\ell'}^{(n,t)}}{\varphi_-^{(n,t)}} a_{\ell'}^{(n,t)}}_{\mathrm{diff}} \right) \right. \\
&\qquad \left. + \sum_{\substack{\ell \notin l(n) \\ \mathrm{diff} < -\mathrm{T}}} [\mathrm{T}_{\theta^{(t)}}^{(n)}]_{\mathrm{In}(x_\ell^{(n)})} \varphi_\ell^{(n,t)} \varphi_-^{(n,t)} \left(\underbrace{a_\ell^{(n,t)} - \sum_{\ell' \notin l(n)} \frac{\varphi_{\ell'}^{(n,t)}}{\varphi_-^{(n,t)}} a_{\ell'}^{(n,t)}}_{\mathrm{diff}} \right) \right\}
\end{aligned}
$$

$$
\begin{aligned}
&\overset{(a)}{\geq} \max_{\mathtt{T}>0} \left\{ \sum_{\ell \notin l(n)} [\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell})} \varphi^{(n,t)}_{-} \left(-\varphi^{(n,t)}_{\ell} \mathtt{T} - 2\|W^{(t)}_{\mathrm{kq}}\| \exp(-\mathtt{T}) \right) \right\} \\
&\overset{(b)}{\geq} -L_{\max}(1 - [\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell_*})}) \varphi^{(n,t)}_{-} \left(1 + \log(2\|W^{(t)}_{\mathrm{kq}}\|) \right) \\
&\geq \frac{L_{\max}\|W^{*}_{\mathrm{kq}}\|}{\varphi^{(n,t)}_{+}\|W^{(t)}_{\mathrm{kq}}\|} \left(1 + \log(2\|W^{(t)}_{\mathrm{kq}}\|) \right) A^{(n,*)} \\
&\overset{(c)}{\geq} \frac{L^2_{\max}\|W^{*}_{\mathrm{kq}}\|}{\|W^{(t)}_{\mathrm{kq}}\|} \left(1 + \log(2\|W^{(t)}_{\mathrm{kq}}\|) \right) A^{(n,*)} \\
&\overset{(d)}{\geq} \beta_0 A^{(n,*)},
\end{aligned}
$$

where (a) follows from Equation (11), (b) is optained by choosing $\mathtt{T} = \log(2\|W^{(t)}_{\mathrm{kq}}\|)$, (c) follows from Lemma 5, and (d) is due to the fact that $\beta_0 \geq \frac{L^2_{\max}\|W^{*}_{\mathrm{kq}}\|}{\|W^{(t)}_{\mathrm{kq}}\|}(1 + \log(2\|W^{(t)}_{\mathrm{kq}}\|))$.

So far, we have shown that for if $\varphi^{(n,t)}_{-} \geq \exp(-(1+\beta_0/2)\|W^{(t)}_{\mathrm{kq}}\|/\|W^{*}_{\mathrm{kq}}\|)$, then

$$
\begin{aligned}
A^{(n,t)} + B^{(n,t)} &= A^{(n,t)} + B^{(n,t)}_{+} + B^{(n,t)}_{-} \\
&\geq (1+\beta_0)A^{(n,*)} + \left((1+\beta_0)B^{(n,*)} + \beta_0 A^{(n,*)} \right) + \beta_0 A^{(n,*)} \\
&= (1+3\beta_0)A^{(n,*)} + (1+\beta_0)B^{(n,*)} \\
&\geq (1+3\beta_0)(A^{(n,*)} + B^{(n,*)}).
\end{aligned}
$$

Type 2. Now consider sentence $X^{(n)}$ such that $\varphi^{(n,t)}_{-} < \exp(-(1+\beta_0/2)\|W^{(t)}_{\mathrm{kq}}\|/\|W^{*}_{\mathrm{kq}}\|)$.

Let n_0 be the type 1 sample such that $\varphi^{(n_0,t)}_{-} \geq \exp(-\|W^{(t)}_{\mathrm{kq}}\|/\|W^{*}_{\mathrm{kq}}\|)$

Then, we aim to show that

$$
A^{(n,t)} + B^{(n,t)} \geq \beta_0 A^{(n_0,*)}
$$

Note that

$$
\begin{aligned}
A^{(n,t)} &\geq \sum_{\ell_* \in l(n)} ([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell_*})} - 1) \varphi^{(n,t)}_{\ell_*} \varphi^{(n,t)}_{-} \left(1 + \log \frac{L_{\max}\|W^{(t)}_{\mathrm{kq}}\|}{\varphi^{(n,t)}_{-}} \right) \\
&\geq ([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell_*})} - 1) \exp\left(-(1+\beta_0/2)\frac{\|W^{(t)}_{\mathrm{kq}}\|}{\|W^{*}_{\mathrm{kq}}\|} \right) \left(1 + (1+\beta_0/2)\frac{\|W^{(t)}_{\mathrm{kq}}\|}{\|W^{*}_{\mathrm{kq}}\|} + \log(L_{\max}\|W^{(t)}_{\mathrm{kq}}\|) \right) \\
&\geq (1+\beta_0)([\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell_*})} - 1) \exp\left(-(1+\beta_0/2)\frac{\|W^{(t)}_{\mathrm{kq}}\|}{\|W^{*}_{\mathrm{kq}}\|} \right) \frac{\|W^{(t)}_{\mathrm{kq}}\|}{\|W^{*}_{\mathrm{kq}}\|}
\end{aligned}
$$

and

$$
\begin{aligned}
B^{(n,t)} &= \sum_{\ell \notin l(n)} [\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell})} \varphi^{(n,t)}_{\ell} \left(a^{(n,t)}_{\ell} - \sum_{\ell'} \varphi^{(n,t)}_{\ell'} a^{(n,t)}_{\ell'} \right) \\
&\geq -2(1 - [\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell_*})}) \exp\left(-(1+\beta_0/2)\frac{\|W^{(t)}_{\mathrm{kq}}\|}{\|W^{*}_{\mathrm{kq}}\|} \right) \|W^{(t)}_{\mathrm{kq}}\|
\end{aligned}
$$

Since

$$A^{(n_0,*)} \le ([\mathrm{T}^{(n_0)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n_0)}_{\ell_*})} - 1)\varphi_+^{(n_0,t)}\varphi_-^{(n_0,t)}\frac{\|W^{(t)}_{\mathrm{kq}}\|}{\|W^*_{\mathrm{kq}}\|}$$
$$\le ([\mathrm{T}^{(n_0)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n_0)}_{\ell_*})} - 1)\varphi_+^{(n_0,t)}\exp\left(-\frac{\|W^{(t)}_{\mathrm{kq}}\|}{\|W^*_{\mathrm{kq}}\|}\right)\frac{\|W^{(t)}_{\mathrm{kq}}\|}{\|W^*_{\mathrm{kq}}\|}$$

We further note that $[\mathrm{T}^{(n_0)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n_0)}_{\ell_*})} < [\mathrm{T}^{(n)}_{\theta^{(t)}}]_{\mathrm{In}(x^{(n)}_{\ell_*})}$ due to $\varphi_+^{(n_0,t)} < \varphi_+^{(n,t)}$. Thus,

$$\begin{aligned}
&A^{(n,t)} + B^{(n,t)} \\
&\ge \exp\left(-\beta_0/2\frac{\|W^{(t)}_{\mathrm{kq}}\|}{\|W^*_{\mathrm{kq}}\|}\right)(1+\beta_0+2\|W_{\mathrm{kq}^*}\|)\frac{A^{(n_0,*)}}{\varphi_+^{(n_0,t)}} \\
&\overset{(a)}{\ge} \beta_0 A^{(n_0,*)},
\end{aligned}$$

where (a) is due to that $\beta_0 \ge \frac{2\|W^*_{\mathrm{kq}}\|(1+2\|W^*_{\mathrm{kq}}\|)}{\|W^{(t)}_{\mathrm{kq}}\|}\log(1+\frac{\|W^{(t)}_{\mathrm{kq}}\|}{2\|W^*_{\mathrm{kq}}\|})$, $\|W^{(t)}_{\mathrm{kq}}\| \ge 2(e-1)\|W^*_{\mathrm{kq}}\|$, and

$$\beta_0 \ge \frac{2\|W^*_{\mathrm{kq}}\|}{\|W^{(t)}_{\mathrm{kq}}\|}\log\frac{1+\beta_0+2\|W^*_{\mathrm{kq}}\|}{\beta_0}.$$

In summary, we have that

$$\begin{aligned}
&\sum_n \pi^{(n)}(A^{(n,t)} + B^{(n,t)}) \\
&= \sum_{n \text{ is type 2}} \pi^{(n)}(A^{(n,t)} + B^{(n,t)}) + \sum_{n \text{ is type 1}} \pi^{(n)}(A^{(n,t)} + B^{(n,t)}) \\
&\ge \max_{n_0 \text{ is type 1}} \beta_0 A^{(n_0,*)} + \sum_{n \text{ is type 1}} \pi^{(n)}((1+3\beta_0)A^{(n,*)} + (1+\beta_1)B^{(n,*)}) \\
&\ge \sum_{n \text{ is type 1}} N\pi^{(n)}\beta_0(A^{(n,*)} + B^{(n,*)}) + (1+3\beta_0)\sum_{n \text{ is type 1}} \pi^{(n)}(A^{(n,*)} + B^{(n,*)}) \\
&\ge (1+(N+3)\beta_0)\sum_{n \text{ is type 1}} \pi^{(n)}(A^{(n,*)} + B^{(n,*)}) \\
&\ge (1+\alpha_t)\sum_{n \text{ is type 2}} \pi^{(n)}(A^{(n,*)} + B^{(n,*)}) + (1+\alpha_t)\sum_{n \text{ is type 1}} \pi^{(n)}(A^{(n_0,*)} + B^{(n_0,*)}) \\
&= (1+\alpha_t)\sum_n \pi^{(n)}(A^{(n,*)} + B^{(n,*)}),
\end{aligned}$$

where $\alpha_t \ge (N+3)\beta_0$. The proof is finished.

■

Now, we are ready to prove Theorem 1.

C.5 Proof of Theorem 1

Proof of Theorem 1.

Recall that $\alpha_t = \frac{4NL^2_{\max}\|W^*_{\mathrm{kq}}\|^2}{\|W^{(t)}_{\mathrm{kq}}\|}\left(1+\log\left(2L_{\max}\|W^{(t)}_{\mathrm{kq}}\|\right)\right)$. By Lemma 7, we have

$$\left\langle W^{(t+1)}_{\mathrm{kq}} - W^{(t)}_{\mathrm{kq}}, \frac{W^*_{\mathrm{kq}}}{\|W^*_{\mathrm{kq}}\|}\right\rangle$$

$$
\begin{aligned}
&= -\eta \left\langle \nabla_{W_{\mathrm{kq}}} \mathcal{L}(\theta^{(t)}), \frac{W_{\mathrm{kq}}^*}{\|W_{\mathrm{kq}}^*\|} \right\rangle \\
&\geq -\frac{\eta}{1+\alpha_t} \left\langle \nabla_{W_{\mathrm{kq}}} \mathcal{L}(\theta^{(t)}), \frac{W_{\mathrm{kq}}^{(t)}}{\|W_{\mathrm{kq}}^{(t)}\|} \right\rangle \\
&= \frac{1}{1+\alpha_t} \left\langle W_{\mathrm{kq}}^{(t+1)} - W_{\mathrm{kq}}^{(t)}, \frac{W_{\mathrm{kq}}^{(t)}}{\|W_{\mathrm{kq}}^{(t)}\|} \right\rangle \\
&= \frac{1}{2\|W_{\mathrm{kq}}^{(t)}\|} \left(\|W_{\mathrm{kq}}^{(t+1)}\|^2 - \|W_{\mathrm{kq}}^{(t+1)} - W_{\mathrm{kq}}^{(t)}\|^2 - \|W_{\mathrm{kq}}^{(t)}\|^2 \right) - \frac{\alpha_t}{1+\alpha_t} \left\langle W_{\mathrm{kq}}^{(t+1)} - W_{\mathrm{kq}}^{(t)}, \frac{W_{\mathrm{kq}}^{(t)}}{\|W_{\mathrm{kq}}^{(t)}\|} \right\rangle \\
&= \frac{\|W_{\mathrm{kq}}^{(t)}\|^2 - \|W_{\mathrm{kq}}^{(t)}\|^2}{2\|W_{\mathrm{kq}}^{(t)}\|} - \frac{\eta^2}{2\|W_{\mathrm{kq}}^{(t)}\|} + \frac{\eta\alpha_t}{1+\alpha_t} \left\langle \frac{\nabla_{W_{\mathrm{kq}}} \mathcal{L}(\theta^{(t)})}{\|\nabla_{W_{\mathrm{kq}}} \mathcal{L}(\theta^{(t)})\|}, \frac{W_{\mathrm{kq}}^{(t)}}{\|W_{\mathrm{kq}}^{(t)}\|} \right\rangle \\
&\geq \|W_{\mathrm{kq}}^{(t+1)}\| - \|W_{\mathrm{kq}}^{(t)}\| - \frac{\eta^2}{2\|W_{\mathrm{kq}}^{(t)}\|} - \frac{\eta\alpha_t}{1+\alpha_t}
\end{aligned}
$$

Let $t_0 = \lceil \frac{8L_{\max}\|W_{\mathrm{kq}}^*\|^2}{\eta} \rceil$ be defined in Lemma 7. Summing over t from t_0, we have

$$
\left\langle W_{\mathrm{kq}}^{(t)} - W_{\mathrm{kq}}^{(t_0)}, \frac{W_{\mathrm{kq}}^*}{\|W_{\mathrm{kq}}^*\|} \right\rangle \geq \|W_{\mathrm{kq}}^{(t)}\| - \|W_{\mathrm{kq}}^{(t_0)}\| - \sum_{t'=t_0}^{t-1} \frac{\eta^2}{2\|W_{\mathrm{kq}}^{(t')}\|} - \sum_{t'=t_0}^{t-1} \frac{\eta\alpha_{t'}}{1+\alpha_{t'}}
$$

By Lemma 6, we have

$$
\begin{aligned}
\sum_{t'=t_0}^{t-1} \frac{1}{\|W_{\mathrm{kq}}^{(t')}\|} &\leq \sum_{t'=t_0}^{t-1} \frac{2L_{\max}\|W_{\mathrm{kq}}^*\|/\eta}{t'} \\
&\leq \frac{2L_{\max}\|W_{\mathrm{kq}}^*\|}{\eta} \log t.
\end{aligned}
$$

Furthermore,

$$
\begin{aligned}
\sum_{t'=t_0}^{t-1} \frac{\alpha_{t'}}{1+\alpha_{t'}} &\leq \sum_{t'=t_0}^{t-1} \alpha_{t'} \\
&= \sum_{t'=t_0}^{t-1} \frac{4NL_{\max}^2\|W_{\mathrm{kq}}^*\|^2}{\|W_{\mathrm{kq}}^{(t')}\|} \left(1 + \log\left(2L_{\max}\|W_{\mathrm{kq}}^{(t')}\|\right)\right) \\
&= \sum_{t'=t_0}^{t-1} \frac{4NL_{\max}^2\|W_{\mathrm{kq}}^*\|^2}{\|W_{\mathrm{kq}}^{(t')}\|} \left(1 + \log\left(2L_{\max}\right)\right) + \sum_{t'=t_0}^{t-1} \frac{4NL_{\max}^2\|W_{\mathrm{kq}}^*\|^2}{\|W_{\mathrm{kq}}^{(t')}\|} \log\|W_{\mathrm{kq}}^{(t')}\| \\
&\leq \sum_{t'=t_0}^{t-1} \frac{8NL_{\max}^3\|W_{\mathrm{kq}}^*\|^3/\eta}{t'} \log(2eL_{\max}) + \sum_{t'=t_0}^{t-1} \frac{8NL_{\max}^3\|W_{\mathrm{kq}}^*\|^3/\eta}{t'} \log\frac{t'}{2L_{\max}\|W_{\mathrm{kq}}^*\|/\eta} \\
&= \sum_{t'=t_0}^{t-1} \frac{8NL_{\max}^3\|W_{\mathrm{kq}}^*\|^3/\eta}{t'} \log(e\eta/\|W_{\mathrm{kq}}^*\|) + \sum_{t'=t_0}^{t-1} \frac{8NL_{\max}^3\|W_{\mathrm{kq}}^*\|^3/\eta}{t'} \log(t') \\
&\overset{(a)}{\leq} \frac{8NL_{\max}^3\|W_{\mathrm{kq}}^*\|^3}{\eta} \log^2 t,
\end{aligned}
$$

where (a) follows from $\eta \leq \|W_{\mathrm{kq}}^*\|/e$.

Therefore, we have

$$
\left\langle W_{\mathrm{kq}}^{(t)} - W_{\mathrm{kq}}^{(t_0)}, \frac{W_{\mathrm{kq}}^*}{\|W_{\mathrm{kq}}^*\|} \right\rangle \geq \|W_{\mathrm{kq}}^{(t)}\| - \|W_{\mathrm{kq}}^{(t_0)}\| - \sum_{t'=t_0}^{t-1} \frac{\eta^2}{2\|W_{\mathrm{kq}}^{(t')}\|} - \sum_{t'=t_0}^{t-1} \frac{\eta\alpha_{t'}}{1+\alpha_{t'}}
$$

$$\geq \|W_{\mathrm{kq}}^{(t)}\| - \|W_{\mathrm{kq}}^{(t_0)}\| - L_{\max}\|W_{\mathrm{kq}}^*\| \log t - 8NL_{\max}^3\|W_{\mathrm{kq}}^*\|^3 \log^2 t.$$

Finally, by Lemma 6, and $t_0 \leq 1 + 8L_{\max}\|W_{\mathrm{kq}}^*\|^2/\eta$, we have $\|W_{\mathrm{kq}}^{(t_0)}\| \leq 9L_{\max}\|W_{\mathrm{kq}}^*\|^2$, and

$$\begin{aligned}\left\langle \frac{W_{\mathrm{kq}}^{(t)}}{\|W_{\mathrm{kq}}^{(t)}\|}, \frac{W_{\mathrm{kq}}^*}{\|W_{\mathrm{kq}}^*\|} \right\rangle &\geq 1 - \frac{2\|W_{\mathrm{kq}}^{(t_0)}\| + L_{\max}\|W_{\mathrm{kq}}^*\| \log t + 8NL_{\max}^3\|W_{\mathrm{kq}}^*\|^3 \log^2 t}{\|W_{\mathrm{kq}}^{(t)}\|} \\ &\geq 1 - \frac{54NL_{\max}^4\|W_{\mathrm{kq}}^*\|^4 \log^2 t}{t\eta}.\end{aligned}$$

The proof is finished.

■

C.6 Proof of Theorem 2

Proof of Theorem 2.

Recall that $T \geq 384\|W_{\mathrm{ov}}^*\|^5 \log(2|\mathcal{V}|) \log T/\eta_0$, and $\Delta = T\eta_0/(4\|W_{\mathrm{ov}}^*\|^2)$ due to Corollary 1 and

$$(e_{\mathrm{In}(x)} - e_v)^\top W_{\mathrm{ov}}^{(T)} x = \frac{T\eta_0}{4\|W_{\mathrm{ov}}^*\|^2}$$

Note that for all $\ell_* \in l(n)$ and $\ell \notin l(n)$

$$\begin{aligned}&(x_\ell^{(n)} - x_{\ell *}^{(n)})^\top W_{\mathrm{kq}}^{(t)} X_{-1}^{(n)} \\ &= \frac{\|W_{\mathrm{kq}}^{(t)}\|}{\|W_{\mathrm{kq}}^*\|}(x_\ell^{(n)} - x_{\ell *}^{(n)})^\top W_{\mathrm{kq}}^* X_{-1}^{(n)} + (x_\ell^{(n)} - x_{\ell *}^{(n)})^\top \left(W_{\mathrm{kq}}^{(t)} - \frac{\|W_{\mathrm{kq}}^{(t)}\|}{\|W_{\mathrm{kq}}^*\|} W_{\mathrm{kq}^*} \right) X_{-1}^{(n)} \\ &\leq -\frac{\|W_{\mathrm{kq}}^{(t)}\|}{\|W_{\mathrm{kq}}^*\|} + 2\|W_{\mathrm{kq}}^{(t)}\| \left\| \frac{W_{\mathrm{kq}}^{(t)}}{\|W_{\mathrm{kq}}^{(t)}\|} - \frac{W_{\mathrm{kq}}^*}{\|W_{\mathrm{kq}}^*\|} \right\| \\ &\leq -\frac{t\eta}{2L_{\max}\|W_{\mathrm{kq}}^*\|^2} + \frac{\sqrt{2}t\eta}{L_{\max}\|W_{\mathrm{kq}}^*\|} \sqrt{\frac{54NL_{\max}^4\|W_{\mathrm{kq}}^*\|^4 \log^2 t}{t\eta}} \\ &\overset{(a)}{\leq} -\frac{t\eta}{4L_{\max}\|W_{\mathrm{kq}}^*\|^2},\end{aligned} \tag{13}$$

where (a) follows from $t \geq 1696NL_{\max}^4\|W_{\mathrm{kq}}^*\|^6 \log^2 t/\eta$.

Therefore,

$$\begin{aligned}\sum_{\ell_* \in l(n)} \varphi_{\ell_*}^{(n,t)} &= \frac{|l(n)|}{|l(n)| + \sum_{\ell' \notin l(n)} \exp\left((x_\ell^{(n)} - x_{\ell *}^{(n)})^\top W_{\mathrm{kq}}^{(t)} X_{-1}^{(n)} \right)} \\ &\geq \frac{|l(n)|}{|l(n)| + (L^{(n)} - |l(n)|) \exp\left(-\frac{t\eta}{4L_{\max}\|W_{\mathrm{kq}}^*\|^2} \right)} \\ &\geq \frac{1}{1 + L_{\max} \exp\left(-\frac{t\eta}{4L_{\max}\|W_{\mathrm{kq}}^*\|^2} \right)} \\ &\geq \frac{1}{1+\epsilon},\end{aligned}$$

where the last inequality follows from $t \geq \frac{4L_{\max}\|W_{\mathrm{kq}}^*\|}{\eta} \log \frac{L_{\max}}{\epsilon}$.

Hence, the loss on the sentence $X^{(n)}$ satisfies that

$$-\log\left(e_{\mathrm{In}(X^{(n)})}^\top \mathrm{T}_{\theta^{(t)}}(X^{(n)}) \right)$$

$$
\begin{aligned}
&= -\log \frac{\exp\left(e_{\mathrm{In}(X^{(n)})}^\top W_{\mathrm{ov}}^{(T)} \sum_\ell x_\ell^{(n)} \varphi_\ell^{(n,t)}\right)}{\sum_{v\le|\mathcal{V}|} \exp\left(e_v^\top W_{\mathrm{ov}}^{(T)} \sum_\ell x_\ell^{(n)} \varphi_\ell^{(n,t)}\right)} \\
&= -\log \frac{1}{1+\sum_{v\neq \mathrm{In}(X^{(n)})} \exp\left((e_v - e_{\mathrm{In}(X^{(n)})})^\top W_{\mathrm{ov}}^{(T)} \sum_\ell x_\ell^{(n)} \varphi_\ell^{(n,t)}\right)} \\
&= \log\left(1+\sum_{v\neq \mathrm{In}(X^{(n)})} \exp\left((e_v - e_{\mathrm{In}(X^{(n)})})^\top W_{\mathrm{ov}}^{(T)} \sum_{\ell\notin l(n)} x_\ell^{(n)} \varphi_\ell^{(n,t)}\right)\right) \\
&= \log\left(1+\sum_{v\neq \mathrm{In}(X^{(n)})} \exp\left(-\Delta \sum_{\ell_*\in l(n)} \varphi_{\ell_*}^{(n,t)} + \Delta \sum_{\ell\notin l(n)} \varphi_\ell^{(n,t)}\right)\right) \\
&\le |\mathcal{V}| \exp\left(-\Delta(2\varphi_+^{(n,t)}-1)\right) \\
&\le |\mathcal{V}| \exp\left(-\Delta + \frac{2\Delta L_{\max}}{L_{\max}+\exp\left(\frac{t\eta}{4L_{\max}\|W_{\mathrm{kq}}^*\|^2}\right)}\right).
\end{aligned}
$$

Thus, the average loss has upper bound, which is

$$
|\mathcal{V}| \exp\left(-\Delta + \frac{2\Delta L_{\max}}{L_{\max}+\exp(C_1 t)}\right),
$$

for $C_1 = \eta/(4L_{\max}\|W_{\mathrm{kq}}^*\|^2)$, and $\Delta = C_0 T$ for $C_0 = \eta_0/(4\|W_{\mathrm{ov}}^*\|^2)$.

■

D Proof of Proposition 2 and Theorem 3

D.1 Proof of Proposition 2

Proposition 4 (Restatement of Proposition 2) *Under Assumption 2, if W_{kq}^* satisfies Equation* (3)*, i.e.,*

$$
W_{\mathrm{kq}}^* = \arg\min \|W\|, \quad \textit{s.t.} \quad (x_{\ell_*}^{(n)} - x_\ell^{(n)})^\top W x \ge 1, \quad \forall \ell \notin l(n), \forall n.
$$

In addition, for each query x^q, if there are k optimal tokens under a x^q-partial order, m non-optimal tokens under x^q-partial order, then, for any optimal token x_*, non-optimal token x, and non-comparable token x_0, we have

$$
x_*^\top W_{\mathrm{kq}}^* x^q = \frac{m}{k+m}, \quad x^\top W_{\mathrm{kq}}^* x^q = -\frac{k}{k+m}, \quad x_0^\top W_{\mathrm{kq}}^* x^q = 0.
$$

A direct result is that

$$
(x_\ell^{(n)} - x_{\ell'}^{(n)})^\top W_{\mathrm{kq}}^* X_{-1}^{(n)} = 0, \quad \forall \ell, \ell' \notin l(n).
$$

Proof. Let $U \in \mathbb{R}^d$ be the rotation matrix such that $Ux = e_{\mathrm{I}(x)}$. Because U preserves Frobenius norm, the optimization problem in Equation (3) can be written as

$$
\tilde{W}^* = \arg\min \|W\|, \quad \text{s.t.}(e_{\mathrm{I}(x_{\ell_*}^{(n)})} - e_{\mathrm{I}(x_\ell^{(n)})})^\top W e_{\mathrm{I}(X_{-1}^{(n)})} \ge 1. \tag{14}
$$

Notably $\tilde{W}^* = U W_{\mathrm{kq}}^* U^\top$.

Note that $\{e_{\mathrm{I}(X_{-1}^{(n)})}\}_n$ forms a standard basis. It suffices to minimize the norm of each column of W subject to the constraint $(e_{\mathrm{I}(x_{\ell_*}^{(n)})} - e_{\mathrm{I}(x_\ell^{(n)})})^\top W e_{\mathrm{I}(X_{-1}^{(n)})} \ge 1$.

Let us consider any column c of W, denoted as $[w_1, \ldots, w_d]^\top$. Without loss of generality, we assume that, for all $X^{(n)}$ with $\mathrm{I}(X_{-1}^{(n)}) = c$, the set of indices of the optimal tokens of those samples are

$\{1, \ldots, k\}$, and the set of indices of the non-optimal tokens are $\{k+1, \ldots, k+m\}$. Then, the optimization problem Equation (14) reduces to the following problem

$$\min w_1^2 + \ldots + w_d^2, \quad \text{s.t.} \quad w_i - w_j \geq 1, \quad \forall i \leq k, j \in A_i \subset \{k+1, \ldots, k+m\}, \tag{15}$$

where A_i is the set of indices of the non-optimal tokens in some samples whose optimal token has index i.

In other words, each column of the solution of Equation (14) is the solution of Equation (15).

Note that Equation (15) is a convex problem with linear constraints. The Lagrangian function is

$$L(\lambda) = \sum_{i=1}^{k+m} w_i^2 + 2\sum_{i=1}^{m}\sum_{j \in A_i} \lambda_{ij}(1 - w_i + w_j),$$

where we directly set $w_j = 0$ for all $j \in \{k+m+1, \ldots, d\}$. That is, non-comparable tokens have value 0.

By KKT-condition, we have

$$\begin{cases} w_i = \sum_{j \in A_i} \lambda_{ij}, & \forall i \leq k \\ w_j = -\sum_{i=1}^{k} \lambda_{ij} \mathbb{1}\{j \in A_i\}, & \forall k+1 \leq j \leq k+m \end{cases}$$

Thus,

$$\begin{aligned} &\min \ w_1^2 + \ldots + w_d^2 \\ &\quad = \max_{\lambda} \left\{ -\sum_{i=1}^{k} \left(\sum_{j \in A_i} \lambda_{ij} \right)^2 - \sum_{j=k+1}^{k+m} \left(\sum_{i=1}^{k} \lambda_{ij} \mathbb{1}\{j \in A_i\} \right)^2 + 2\sum_{i=1}^{m}\sum_{j \in A_i} \lambda_{ij} \right\} \end{aligned}$$

Let

$$L^*(\lambda) = -\sum_{i=1}^{k} \left(\sum_{j \in A_i} \lambda_{ij} \right)^2 - \sum_{j=k+1}^{k+m} \left(\sum_{i=1}^{k} \lambda_{ij} \mathbb{1}\{j \in A_i\} \right)^2 + 2\sum_{i=1}^{m}\sum_{j \in A_i} \lambda_{ij},$$

where $\lambda \geq 0$. The maximum of L^* is achieved when $\nabla_\lambda L^* = 0$. This implies that

$$\sum_{j \in A_{i_0}} \lambda_{i_0 j} + \sum_{i=1}^{k} \lambda_{ij} \mathbb{1}\{j_0 \in A_i\} = 1, \quad \forall 1 \leq i_0 \leq k < j_0 \leq k+m.$$

Hence, we have $w_i - w_j = 1$ for all $1 \leq i \leq k < j \leq k+m$, which means the optimum of the original problem is achieved on the boundary. Therefore, we reduce the original problem to

$$\min_x k(x+1)^2 + mx^2,$$

where $x = w_{k+1} = \ldots = w_{k+m}$. Hence, the optimal solution is $w_1 = \ldots = w_k = m/(m+k)$, and $w_{k+1} = \ldots = w_{k+m} = -k/(k+m)$.

Therefore, the solution of Equation (15) satisfies that the "optimal values" are the same and the "non-optimal values" are the same as well. This fact proves that

$$(x_\ell^{(n)} - x_{\ell'}^{(n)}) W_{\text{kq}}^* X_{-1}^{(n)} = 0, \forall \ell, \ell' \notin l(n).$$

And moreover, if there are k optimal tokens under a x^q-partial order, m non-optimal tokens under x^q-partial order, then, for any optimal token x_* and non-optimal token x, we have

$$x_*^\top W_{\text{kq}}^* x^q = \frac{m}{k+m}, \quad x^\top W_{\text{kq}}^* x^q = -\frac{k}{k+m}.$$

■

D.2 Proof of Theorem 3

Proof. The proof follows similar logic to Theorem 2. By Equation (13), we have for any $x, x' \in \mathcal{V}$

$$\begin{aligned}
&(x - x')^\top W_{\mathrm{kq}}^{(t)} x^q \\
&\geq \frac{\|W_{\mathrm{kq}}^{(t)}\|}{\|W_{\mathrm{kq}}^*\|}(x - x')^\top W_{\mathrm{kq}}^* x^q - \frac{\sqrt{2}t\eta}{L_{\max}\|W_{\mathrm{kq}}^*\|}\sqrt{\frac{54NL_{\max}^4\|W_{\mathrm{kq}}^*\|^4 \log^2 t}{t\eta}} \\
&\overset{(a)}{\geq} \frac{t\eta}{2L_{\max}\|W_{\mathrm{kq}}^*\|^2}(x - x')^\top W_{\mathrm{kq}}^* x^q - \sqrt{108t\eta NL_{\max}^2\|W_{\mathrm{kq}}^*\|^2 \log^2 t},
\end{aligned}$$

where (a) follows from Lemma 6. The first part of Theorem 3 follows from Theorem 3.

For the second part of Theorem 3, Let $X = [x_1, \ldots, x_L]$ such that for $\ell_0 \in l_0 \subset \{1, \ldots, L\}$, $x_{\ell_0} = x$ is a non-comparable token, and other tokens are non-optimal under the x_L-partial order.

Let $\varphi_\ell \propto \exp(x_\ell W_{\mathrm{kq}}^{(t)} x_L)$ for sufficiently large $t = \Omega(\log(1/\epsilon))$ such that $\sum_{\ell_0 \in l_0} \varphi_{\ell_0} \geq 1 - \epsilon$.

Then, we have

$$\begin{aligned}
e_{\mathrm{In}(x)}^\top \mathrm{T}_{\theta^{(t)}}(X) &= \frac{\exp\left(e_{\mathrm{In}(x)}^\top W_{\mathrm{ov}}^{(T)} \sum_\ell x_\ell \varphi_\ell\right)}{\sum_{v \leq |\mathcal{V}|} \exp\left(e_v^\top W_{\mathrm{ov}}^{(T)} \sum_\ell x_\ell \varphi_\ell\right)} \\
&= \frac{1}{1 + \sum_{v \neq \mathrm{In}(x)} \exp\left((e_v - e_{\mathrm{In}(x)})^\top W_{\mathrm{ov}}^{(T)} \sum_\ell x_\ell \varphi_\ell\right)} \\
&= \frac{1}{1 + \sum_{v \neq \mathrm{In}(x)} \exp\left(-\Delta \sum_{\ell_0 \in l_0} \varphi_{\ell_0} + \Delta \sum_{\ell \notin l_0} \varphi_\ell\right)} \\
&\geq \frac{1}{1 + |\mathcal{V}| \exp\left(-\Delta(1 - 2\epsilon)\right)} \\
&\geq 1 - \epsilon_0,
\end{aligned}$$

where the last inequality follows from $T = O(\log(1/\epsilon_0))$. Therefore, the trained transformer will predict $n(x)$, the next token of the non-comparable token.

■

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: We provide details of our three main claims made in the abstract in Section 4, Section 5, and Section 6, respectively.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: Our paper has specified the assumptions (Assumption 1,Assumption 2, and Assumption 3) under which our results hold.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory Assumptions and Proofs**

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

Answer: [Yes]

Justification: We provide the full proofs in the appendix and title them as the "proof of lemmas, propositions, and theorems". Each proof can be easily found from the table of contents.

Guidelines:

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

4. **Experimental Result Reproducibility**

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

Answer: [Yes]

Justification: We provide our experiment details in Section 7, including the model dimension, vocabulary size, the learning rate, and the iteration number. The experiment follows exactly of Algorithm 1.

Guidelines:

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

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

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

   Answer: [Yes]

   Justification: We provide our code in the supplemental. We do not use open source data.

   Guidelines:

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

6. **Experimental Setting/Details**

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

   Answer: [Yes]

   Justification: Our experiment details are provided in Section 7.

   Guidelines:

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

7. **Experiment Statistical Significance**

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

   Answer: [Yes]

   Justification: As our synthetic dataset and training process is fixed, we do not foresee significant errors.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments Compute Resources**

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

   Answer: [Yes]

   Justification: Our experiment is on synthetic data and is computationally lightweight, which does not require any GPU. The running time is also within a hour.

   Guidelines:

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

9. **Code Of Ethics**

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

   Answer: [Yes]

   Justification: We follow exactly the code of ethics.

   Guidelines:

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

10. **Broader Impacts**

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

    Answer: [NA]

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

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.

- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA]

    Justification: Our paper is primarily a theoretical work, and does not have such risk.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [NA]

    Justification: Our paper is primarily a theoretical work, and does not use any assets.

    Guidelines:

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

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New Assets**

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

    Answer: [NA]

    Justification: Our paper is primarily a theoretical work, and does not release any assets.

    Guidelines:

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

14. **Crowdsourcing and Research with Human Subjects**

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

    Answer: [NA]

    Justification: Our paper is primarily a theoretical work, and does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

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

    Answer: [NA]

    Justification: Our paper is primarily a theoretical work, and does not involve crowdsourcing nor research with human subjects.

    Guidelines:

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

- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.