On the Convergence of Gradient Descent on Learning Transformers with Residual Connections

Zhen Qin, **Jinxin Zhou**[*] **and** **Zhihui Zhu**[†]

Department of Computer Science and Engineering,
The Ohio State University,
{qin.660, zhou.3820, zhu.3440}@osu.edu

Abstract

Transformer models have emerged as fundamental tools across various scientific and engineering disciplines, owing to their outstanding performance in diverse applications. Despite this empirical success, the theoretical foundations of Transformers remain relatively underdeveloped, particularly in understanding their training dynamics. Existing research predominantly examines isolated components–such as self-attention mechanisms and feedforward networks–without thoroughly investigating the interdependencies between these components, especially when residual connections are present. In this paper, we aim to bridge this gap by analyzing the convergence behavior of a structurally complete yet single-layer Transformer, comprising self-attention, a feedforward network, and residual connections. We demonstrate that, under appropriate initialization, gradient descent exhibits a linear convergence rate, where the convergence speed is determined by the minimum and maximum singular values of the output matrix from the attention layer. Moreover, our analysis reveals that residual connections serve to ameliorate the ill-conditioning of this output matrix, an issue stemming from the low-rank structure imposed by the softmax operation, thereby promoting enhanced optimization stability. We also extend our theoretical findings to a multi-layer Transformer architecture, confirming the linear convergence rate of gradient descent under suitable initialization. Empirical results corroborate our theoretical insights, illustrating the beneficial role of residual connections in promoting convergence stability.

1 Introduction

Transformer model architectures [VSP+17] have gained widespread recognition and popularity in various scientific and engineering applications, consistently achieving outstanding performance across numerous domains. Notably, Transformers have achieved substantial success in natural language processing tasks [RWC+19, BMR+20], recommendation systems [ZZS+18, CZL+19], reinforcement learning [CLR+21, JLL21], computer vision [DBK+20], multi-modal signal processing [TBL+19], quantum information [MSD+25], and communication systems [KLJ+23]. One prominent example is their impressive performance in large language models such as GPT-4 [AAA+23], where Transformers play a central role in achieving unprecedented language generation capabilities. However, despite their empirical success and widespread adoption, the theoretical understanding of Transformers remains limited, posing significant challenges in analyzing their fundamental mechanisms and performance guarantees.

[*] The first two authors contributed to this work equally.

[†] Corresponding author.

Preprint. Under review.

To address this gap, an expanding body of theoretical literature has explored various facets of Transformer models, including the impact of initialization [MBG+24], sample complexity guarantees [IHL+24], scaling limits [BCP24], and implicit regularization effects [ATLZO23, TLTO23]. A crucial research direction emerging from these efforts is the investigation of training dynamics in Transformers, with particular emphasis on the two fundamental components within each layer: the feedforward network and the self-attention mechanism. Notably, most theoretical studies tend to examine these components independently, often focusing on their properties in isolation. Among these efforts, one important research direction explores the convergence behavior of deep (linear) neural networks [ACGH18, Sha19, NM20, Cha22, QTZ24], showing that with proper initialization, such networks can achieve a linear convergence rate. In addition to these studies, a considerable number of works [ZFB24, HCL23, CSWY24, YHLC24, ZSLS25, LWL+24, MHM23, ACDS23, RWL24, FCJS24, TWCD23, HLY24] have focused on self-attention mechanism-based in-context learning, particularly examining the training dynamics and landscape. However, despite the growing interest in in-context learning, this line of research does not adequately capture the complexities of the training process underlying self-attention mechanisms. The primary reason for this is that in-context learning dynamics are primarily concerned with how models adapt to new tasks during inference, while the training dynamics of self-attention mechanisms within Transformer models focus on the process through which self-attention weights are learned during training. Moreover, in-context learning inherently relies on a specific input matrix structure, which is not directly applicable to the training dynamics of self-attention mechanisms. In response to this gap, recent work [SHZ+24] investigates the training dynamics of gradient descent for self-attention mechanisms within Transformer models, providing insights into how gradient-based optimization influences the evolution of weight matrices during the training process.

Beyond investigating individual components, only recent work [WLCC23] studies the convergence rate of a softmax attention layer combined with a feedforward network using a single ReLU activation function, without imposing any specific restrictions on the input matrix. This work shows that, with proper initialization, gradient descent can achieve a linear convergence rate. Moreover, it demonstrates that under suitable conditions, commonly used initialization schemes such as LeCun and He, and NTK initialization satisfy the requirements for convergence. However, a key limitation of this analysis is the assumption that the weight matrices in the feedforward network are identity matrices, which restricts its applicability to more general settings. Consequently, the results only partially capture the training dynamics of a single-layer Transformer. Furthermore, while previous works have advanced the theoretical understanding of residual connections–such as [HM16], which proved that deep linear residual networks, and [LCZ+19], which showed that a two-layer non-overlapping convolutional residual network does not suffer from spurious local optima due to the use of residual connections, and [SWJS24], which demonstrated that residual connections alleviate the oversmoothing problem in graph neural networks—the theoretical understanding of their role in Transformers remains underdeveloped. As illustrated in Figure 1, a single-layer Transformer with residual connections exhibits a faster convergence rate compared to its counterpart without residual connections. This phenomenon can be intuitively explained by the occurrence of rank collapse [NAB+22] in certain scenarios, leading to an ill-conditioned output matrix in the softmax attention layer, which, in turn, may hinder the convergence speed during Transformer training. In this regard, residual connections could help mitigate such issues by stabilizing the training process. To the best of our knowledge, there is no theoretical understanding of the convergence behavior of learning Transformers with a feedforward network and residual connections.

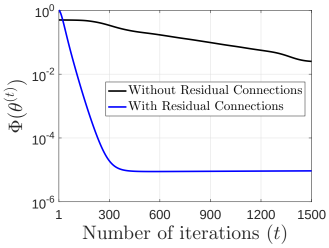


Figure 1: Comparison of training dynamics of single-layer Transformers with and without residual connections (for specific settings, refer to “The importance of residual connections” in Section 4).

Contribution In this paper, we investigate the convergence behavior for learning a single-layer Transformer equipped with single-head self-attention, a feedforward network, and residual connections. We demonstrate that, with proper initialization, gradient descent achieves a linear convergence rate. Our analysis further shows that the convergence speed is influenced by the minimum and maximum singular values of the output matrix from the attention layer, with residual connections

Table 1: Comparisons with existing theoretical works that study the convergence behavior of learning Transformers.

Reference	nonlinear attention	feedforward network	residual connection
[SHZ+24]	✓	×	×
[WLCC23]	✓	×	×
Ours	✓	✓	✓

playing a critical role in mitigating ill-conditioning. Moreover, our convergence analysis can be extended to multi-layer Transformer architecture, indicating that gradient descent retains a linear convergence rate when appropriately initialized. The convergence rate depends on the output matrix of the last attention layer, emphasizing the pivotal role of residual connections in ensuring stable and efficient training. Experimental results validate our theoretical findings, confirming the linear convergence rate of gradient descent and demonstrating the beneficial effect of introducing residual connections.

Notation We use bold capital letters (e.g., $\boldsymbol{Y}$) to denote matrices, bold lowercase letters (e.g., $\boldsymbol{y}$) to denote vectors, and italic letters (e.g., y) to denote scalar quantities. Elements of matrices are denoted in parentheses, as in Matlab notation. For example, $\boldsymbol{Y}(s_1, s_2)$ denotes the element in position (s_1, s_2) of the matrix $\boldsymbol{Y}$. The inner product of $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$ and $\boldsymbol{B} \in \mathbb{R}^{d_1 \times d_2}$ can be denoted as $\langle \boldsymbol{A}, \boldsymbol{B} \rangle = \sum_{s_1=1}^{d_1} \sum_{s_2=1}^{d_2} \boldsymbol{A}(s_1, s_2)\boldsymbol{B}(s_1, s_2)$. $\|\boldsymbol{X}\|$ and $\|\boldsymbol{X}\|_F$ respectively represent the spectral norm and Frobenius norm of $\boldsymbol{X}$. $\sigma_i(\boldsymbol{X})$ is the i-th singular value of $\boldsymbol{X}$ and the condition number $\boldsymbol{X}$ is defined as $\kappa(\boldsymbol{X}) = \frac{\|\boldsymbol{X}\|}{\sigma_{\min}(\boldsymbol{X})}$. For two positive quantities $a, b \in \mathbb{R}$, the inequality $b \lesssim a$ or $b = O(a)$ means $b \leq ca$ for some universal constant c; likewise, $b \gtrsim a$ or $b = \Omega(a)$ represents $b \geq ca$ for some universal constant c.

2 Related Works

Convergence analysis for neural networks The convergence properties of neural networks, particularly in the context of gradient-based optimization, have been extensively studied in the literature. A notable line of research has focused on the convergence behavior of linear residual networks, a subclass of linear neural networks. In particular, it has been established that these networks converge to the global minimum at a linear rate [BHL18, ZLG20]. Similarly, the convergence speed of gradient descent when training deep linear neural networks has been rigorously analyzed, revealing the conditions under which convergence to the global minimum occurs [ACGH18]. Additionally, the convergence analysis of orthonormal deep linear neural networks, which introduces structural regularization through orthogonality constraints, has also been investigated [QTZ24]. On the other hand, [Sha19] showed that gradient descent with random initialization may require exponentially many iterations with respect to the number of layers to converge in one-dimensional deep linear neural networks, indicating the potential difficulties caused by depth and initialization choices. Furthermore, the convergence rate of training more complex architectures, such as recurrent neural network and deep neural networks, has also been explored. Specifically, [AZLS19] examined the convergence behavior in the context of recurrent neural network training, while [NM20, Cha22] provided convergence guarantees for deep neural networks under specific conditions.

Convergence analysis for in-context learning Many works have been proposed to investigate the training dynamics of in-context learning, focusing particularly on understanding how models adapt to new tasks through contextual information. In [ZFB24], it was demonstrated that for a single-layer linear self-attention model, gradient flow with a carefully chosen random initialization converges to a global minimum, achieving a small prediction error when trained on anisotropic Gaussian data. Similarly, [HCL23] made a pioneering contribution by analyzing the training dynamics of softmax attention, providing convergence results for a one-layer transformer with single-head attention on linear regression tasks. Building on these foundations, [CSWY24, YHLC24, ZSLS25] extended the analysis to transformers with multi-head softmax attention trained via gradient descent for in-context learning tasks, offering conditions under which convergence is guaranteed. Additionally, [LWL+24] examined the training process of a one-layer, single-head transformer with softmax attention for in-context learning on binary classification tasks, yielding insights into the convergence behavior specific to this scenario. An alternative perspective on convergence was presented in [MHM23], where it was shown that a transformer performing a single gradient descent step on a least squares linear regression objective can act as the global minimizer of the pre-training loss. Moreover, [ACDS23] investigated the loss landscape of linear transformers trained for in-context learning, offering a comprehensive analysis of how the optimization landscape influences the training dynamics. While gradient-based methods have been extensively studied, recent works have also explored alternative optimization techniques. For instance, [FCJS24] investigated the convergence rate of in-context learning using second-order optimization methods, emphasizing the potential for faster convergence compared to first-order approaches.

3 Problem Setting

This section formalizes the architecture of the Transformer model and the corresponding training objective. We consider a simplified setting involving a single-layer Transformer equipped with single-head self-attention, a feedforward network and residual connections. Given an input matrix $\boldsymbol{X} \in \mathbb{R}^{M\times d}$, where M denotes the number of discrete tokens and d is the embedding dimension, the model is specified as follows:

$$F_{\boldsymbol{\Theta}}(\boldsymbol{X}) = (\text{FFN}(\text{Attn}(\boldsymbol{X}) + \boldsymbol{X}) + \text{Attn}(\boldsymbol{X}) + \boldsymbol{X})\boldsymbol{W}_U, \tag{1}$$

where $\boldsymbol{\Theta} = \{\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V, \boldsymbol{W}_U\}$. For analytical simplicity, layer normalization is not incorporated in our formulation, aligning with prior convergence analyses that likewise omit it in [SHZ+24, WLCC23]. We now provide a detailed introduction to each component of the model as follows.

- Self-Attention Mechanism: A self-attention mechanism is a central component of the Transformer architecture, enabling contextualization of token representations across layers. We denote the attention head function by Attn$(\cdot)$. In this work, we consider the standard softmax attention [VSP+17], for which Attn$(\cdot)$ is defined as follows:

$$\text{Attn}(\boldsymbol{X}) := \phi_s\left(\frac{\boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top}{\sqrt{d_{QK}}}\right)\boldsymbol{X}\boldsymbol{W}_V \in \mathbb{R}^{M\times d}, \tag{2}$$

  where $\boldsymbol{W}_Q \in \mathbb{R}^{d\times d_{QK}}$, $\boldsymbol{W}_K \in \mathbb{R}^{d\times d_{QK}}$ and $\boldsymbol{W}_V \in \mathbb{R}^{d\times d}$ denote the query, key, and value weight matrices, respectively. The function $\phi_s(\cdot)$ denotes the row-wise softmax operation.

- Feedforward Network: The feedforward network (FFN) in each Transformer block comprises two learnable weight matrices, $\boldsymbol{W}_1 \in \mathbb{R}^{d\times d_1}$ and $\boldsymbol{W}_2 \in \mathbb{R}^{d_1\times d}$. In line with recent architectures such as LLaMA [TMS+23], PaLM [CND+23], and OLMo [GBW+24], we omit bias terms. The feedforward function FFN$(\cdot)$ is defined as

$$\text{FFN}(\boldsymbol{Z}(\boldsymbol{X})) := \phi_r(\boldsymbol{Z}(\boldsymbol{X})\boldsymbol{W}_1)\boldsymbol{W}_2 \in \mathbb{R}^{M\times d}, \tag{3}$$

  where $\boldsymbol{Z}(\boldsymbol{X}) = \text{Attn}(\boldsymbol{X}) + \boldsymbol{X}$ and ϕ_r is a general element-wise activation function.

- Residual Connections: Residual connections are incorporated to promote stable training and improve gradient flow in deep Transformer architectures. For each sub-layer (e.g., self-attention or feedforward), the residual connection adds the input (denoted by $\boldsymbol{X}$ or $\boldsymbol{Z}(\boldsymbol{X})$) of the sub-layer to its output.

- Unembedding Layer: In the final layer of the Transformer, the residual stream is projected into the vocabulary space via the unembedding matrix $\boldsymbol{W}_U \in \mathbb{R}^{d\times N}$. Without loss of generality, we include the unembedding matrix; for the convergence analysis of intermediate layers, this matrix can be treated as a constant 1.

For clarity, the full sequence of the operations described above is illustrated in Figure 2.

Based on the Transformer model described above, we consider the supervised learning setting with a dataset $\{\boldsymbol{X}_p, \boldsymbol{Y}_p\}_{p=1}^P \in \{\mathbb{R}^{M\times d}, \mathbb{R}^{M\times N}\}$. The training objective is to minimize the following squared Frobenius norm loss:

$$L(\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V, \boldsymbol{W}_U) = \frac{1}{2}\sum_{p=1}^{P}\|F_{\boldsymbol{\Theta}}(\boldsymbol{X}_p) - \boldsymbol{Y}_p\|_F^2 = \frac{1}{2}\|\overline{F}_{\boldsymbol{\Theta}}(\boldsymbol{X}) - \overline{\boldsymbol{Y}}\|_F^2, \tag{4}$$

where $\overline{F}_{\boldsymbol{\Theta}}(\boldsymbol{X}) = \left[F_{\boldsymbol{\Theta}}(\boldsymbol{X}_1)^\top \cdots F_{\boldsymbol{\Theta}}(\boldsymbol{X}_p)^\top\right]^\top \in \mathbb{R}^{MP\times N}$ and $\overline{\boldsymbol{Y}} = \left[\boldsymbol{Y}_1^\top \cdots \boldsymbol{Y}_P^\top\right]^\top \in \mathbb{R}^{MP\times N}$.

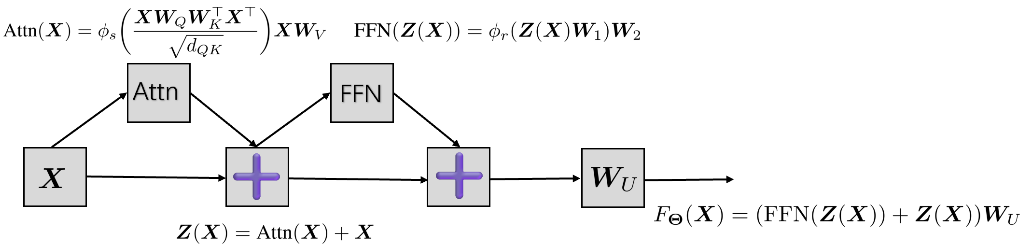

$\mathrm{Attn}(\boldsymbol{X}) = \phi_s\left(\frac{\boldsymbol{X}\boldsymbol{W}_Q\boldsymbol{W}_K^\top\boldsymbol{X}^\top}{\sqrt{d_{QK}}}\right)\boldsymbol{X}\boldsymbol{W}_V$
$\mathrm{FFN}(\boldsymbol{Z}(\boldsymbol{X})) = \phi_r(\boldsymbol{Z}(\boldsymbol{X})\boldsymbol{W}_1)\boldsymbol{W}_2$
Attn
FFN
$\boldsymbol{X}$
$\boldsymbol{W}_U$
$\boldsymbol{Z}(\boldsymbol{X}) = \mathrm{Attn}(\boldsymbol{X}) + \boldsymbol{X}$
$F_{\boldsymbol{\Theta}}(\boldsymbol{X}) = (\mathrm{FFN}(\boldsymbol{Z}(\boldsymbol{X})) + \boldsymbol{Z}(\boldsymbol{X}))\boldsymbol{W}_U$

Figure 2: Illustration of a single-layer Transformer with single-head attention, a feedforward network and residual connections.

4 Convergence Analysis

In this section, we provide a theoretical analysis of the optimization method associated with problem (4). Specifically, we investigate the convergence behavior of the vanilla gradient descent (GD) algorithm, which updates each parameter according to

$$\boldsymbol{W}_b^{(t+1)} = \boldsymbol{W}_b^{(t)} - \mu\nabla_{\boldsymbol{W}_b} L(\boldsymbol{W}_1^{(t)}, \boldsymbol{W}_2^{(t)}, \boldsymbol{W}_Q^{(t)}, \boldsymbol{W}_K^{(t)}, \boldsymbol{W}_V^{(t)}, \boldsymbol{W}_U^{(t)}), \tag{5}$$

where $b \in \{1, 2, Q, K, V, U\}$ and μ is the learning rate. All detailed gradient expressions are presented in Appendix A. To facilitate our analysis, we introduce the following assumption on the Lipschitz condition of the activation functions.

Assumption 1 (Lipschitz condition of the activation function) *Let σ_r be a non-decreasing function satisfying $|\sigma_r(x) - \sigma_r(y)| \leq |x - y|$ for every $x, y \in \mathbb{R}$.*

This condition holds for several commonly used activation functions. For instance, the ReLU function $\phi_r(x) = \max\{0, x\}$ satisfies this assumption. Moreover, smooth approximations of ReLU, such as the Gaussian-smoothed ReLU discussed in [NM20], also comply with this property.

To simplify notation and analysis, we vectorize the model output $\overline{F}_{\boldsymbol{\Theta}}(\boldsymbol{X})$ and the ground truth matrix $\overline{\boldsymbol{Y}}$. Specifically, we let $f_{\boldsymbol{\theta}}(\boldsymbol{X}) = \mathrm{vec}(\overline{F}_{\boldsymbol{\Theta}}(\boldsymbol{X}))$ and $\boldsymbol{y} = \mathrm{vec}(\overline{\boldsymbol{Y}})$, where the parameter vector $\boldsymbol{\theta}$ is constructed by stacking the vectorized weights as $\boldsymbol{\theta} = \left[\mathrm{vec}(\boldsymbol{W}_1)^\top\ \mathrm{vec}(\boldsymbol{W}_2)^\top\ \mathrm{vec}(\boldsymbol{W}_U)^\top\ \mathrm{vec}(\boldsymbol{W}_V)^\top\ \mathrm{vec}(\boldsymbol{W}_Q)^\top\ \mathrm{vec}(\boldsymbol{W}_K)^\top\right]^\top$. Using this compact formulation, the original objective (4) can be equivalently rewritten as a standard least-squares loss:

$$\Phi(\boldsymbol{\theta}) = \frac{1}{2}\|f_{\boldsymbol{\theta}}(\boldsymbol{X}) - \boldsymbol{y}\|_2^2. \tag{6}$$

This reformulation allows us to analyze the convergence behavior of GD in a more tractable form, leveraging standard tools from vector-valued function analysis. Now, we present our convergence analysis as follows.

Theorem 1 *Given a dataset $\{\boldsymbol{X}_p, \boldsymbol{Y}_p\}_{p=1}^P$, we consider a single-layer Transformer architecture in (1). We define $\overline{\lambda} = \max_{b\in\{1,2,U,V,Q,K\}} \|\boldsymbol{W}_b^{(0)}\|$, $\underline{\lambda} = \min_{b\in\{1,2,U,V,Q,K\}} \sigma_{\min}(\boldsymbol{W}_b^{(0)})$ and $\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p) = \phi_s(\boldsymbol{X}_p\boldsymbol{W}_Q^{(0)}\boldsymbol{W}_K^{(0)\top}\boldsymbol{X}_p^\top/\sqrt{d_{QK}})\boldsymbol{X}_p\boldsymbol{W}_V^{(0)} + \boldsymbol{X}_p$. In addition, we assume that the matrices $\{\boldsymbol{W}_b^{(0)}\}_{b\in\{1,2,U,V,Q,K\}}$ are either full row rank or full column rank, and the initialization satisfies*

$$\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}) \lesssim \frac{\underline{\lambda}^2(\min_p \sigma_{\min}(\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)})))^2}{\max\{1, \overline{\lambda}^7\} \cdot \max_p \|\boldsymbol{X}_p\|^5 \max_{i,p} \|\boldsymbol{X}_p(i,:)\|_2 \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2}. \tag{7}$$

Define $\alpha = \frac{\sigma_{\min}^2(\boldsymbol{W}_U^{(0)}) \min_p \sigma_{\min}^2(\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)}))}{16}$. Using the gradient descent in (5), we have

$$\Phi(\boldsymbol{\theta}^{(t+1)}) \leq (1 - \mu\alpha)\Phi(\boldsymbol{\theta}^{(t)}), \tag{8}$$

where the learning rate satisfies $\mu \le \min\{\frac{1}{C}, \frac{1}{\alpha}\}$*, and we define*

$$\begin{aligned}C = \widetilde{C} \cdot \max\{&\overline{\lambda}^2 \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|, \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|, \max_p \|\boldsymbol{X}_p\|(\overline{\lambda}^3 + \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|),\\ &(\overline{\lambda}^3 + \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|) \max_p \|\boldsymbol{X}_p\|^3 \overline{\lambda}^2\}(\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|(1 + \overline{\lambda}^2)\\ &+ \max_p \|\boldsymbol{X}_p\|^3(\overline{\lambda}^3 + \overline{\lambda}^5) + \max_p \|\boldsymbol{X}_p\|^2 \max_{i,p} \|\boldsymbol{X}_p(i,:)\|_2(\overline{\lambda}^3 + \overline{\lambda}^5)),\end{aligned} \tag{9}$$

where $\widetilde{C}$ *is a positive constant.*

The proof has been provided in Appendix C. According to Theorem 1, when the initialization condition is satisfied, GD in (5) achieves a linear convergence rate. We further remark that the specific convergence rate is influenced by the parameters α and C, both of which are dependent on the initialization. To facilitate the subsequent analysis, we assume that the elements of the weight matrices $\{\boldsymbol{W}_b^{(0)}\}_{b=1,2,U,V,Q,K}$ are drawn from a standard Gaussian distribution with varying variances. Based on these assumptions, the weight matrices are full row or full column rank with high probability. Moreover, employing the ReLU activation function, we ensure that the conditions $d_1/4 \ge d$ and $N/4 \ge d$ hold. The initializations are specified as $\boldsymbol{W}_1^{(0)}(i,j) \sim \mathcal{N}(0, \gamma_1^2)$ and $\boldsymbol{W}_U^{(0)}(i,j) \sim \mathcal{N}(0, \gamma_U^2)$. Under these settings, with high probability, α constitutes a strictly positive lower bound rather than being zero.

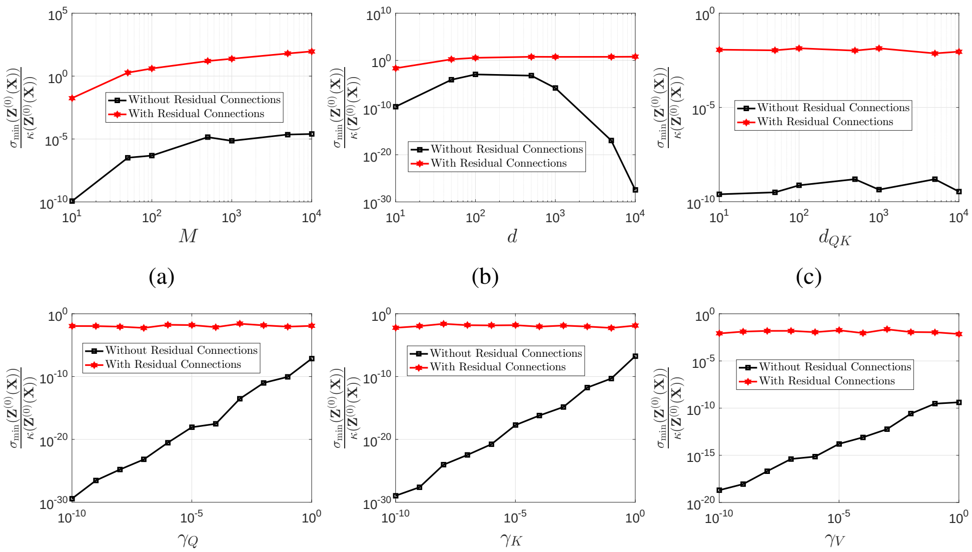
$$\begin{aligned}\alpha &= \frac{\sigma_{\min}^2(\boldsymbol{W}_U^{(0)}) \min_p \sigma_{\min}^2(\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)}))}{32}\\ &\ge \frac{d_1\gamma_1^2\gamma_U^2(\mu_1(\sigma_r))^2}{128}(\frac{\sqrt{N}}{2} - \sqrt{d})^2 \min_p \sigma_{\min}^2(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)),\end{aligned} \tag{10}$$

where the first inequality follows from the singular value inequality for Gaussian matrices, as stated in Lemma 4 of Appendix E, along with results from [NM20, Lemma C.2] and [NMM21, Lemma 5.3]. Here, $\mu_1(\sigma_r)$ denotes the first Hermite coefficient of the ReLU function, satisfying $\mu_1(\sigma_r) > 0$.

On the other hand, when the weight matrices are initialized with Gaussian distributions, their spectral norms and minimum singular values can be bounded, as established in Lemma 4 in Appendix E. In this case, the positive values $\overline{\lambda}$, $\underline{\lambda}$, $\|\boldsymbol{X}_p\|$ and $\|\boldsymbol{X}_p(i,:)\|_2$ exert the same influence on the convergence analysis for Transformer model regardless of the presence of residual connections. Consequently, leveraging the lower bound in (10), when $\frac{1}{C} < \frac{1}{\alpha}$, the convergence rate $1 - \mu\alpha$ is predominantly influenced by the term $\frac{\min_p \sigma_{\min}^2(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p))}{\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|}$. Here $\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p) = \frac{\phi_s(\boldsymbol{X}_p\boldsymbol{W}_Q^{(0)}\boldsymbol{W}_K^{(0)\top}\boldsymbol{X}_p^\top}{\sqrt{d_{QK})}}\boldsymbol{X}_p\boldsymbol{W}_V^{(0)} + \boldsymbol{X}_p$ is determined by the residual associated with the input data $\boldsymbol{X}_p$. Now, we select $1 - C_1\frac{\min_p \sigma_{\min}^2(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p))}{\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|}$, where C_1 is a positive constant chosen to ensure that this term remains less than 1, as the convergence rate index to elucidate the role of the residual connection.

The importance of residual connections From a theoretical perspective, it is essential to recognize that when the constant C_1 is fixed, there may exist extreme cases in which $\text{Attn}(\boldsymbol{X})$ becomes ill-conditioned, potentially compromising the model's stability and convergence. For instance, as established by [NAB+22, Theorem A.2], consider the scenario where $d < \infty$ is fixed and $d_{QK} \to \infty$. In this regime, for any input $\boldsymbol{X}$, it can be shown that $\text{Attn}(\boldsymbol{X}) \to \frac{1}{M}\boldsymbol{1}_{M\times M}\boldsymbol{X}\boldsymbol{W}_V$, where $\boldsymbol{1}_{M\times M}$ denotes an $M \times M$ matrix with all elements equal to one. Consequently, in the absence of a residual connection (i.e., $\boldsymbol{Z}^{(0)}(\boldsymbol{X}) = \text{Attn}(\boldsymbol{X})$), the resulting matrix $\boldsymbol{Z}^{(0)}(\boldsymbol{X})$ is rank-one, yielding $\sigma_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X})) \to 0$, implying that the convergence curve will remain flat. However, when the residual connection is introduced, the resulting matrix $\boldsymbol{Z}^{(0)}(\boldsymbol{X}) \to \frac{1}{M}\boldsymbol{1}_{M\times M}\boldsymbol{X}\boldsymbol{W}_V + \boldsymbol{X}$ maintains full rank as long as the input matrix $\boldsymbol{X}$ is of full rank, thereby ensuring that the minimum singular value $\sigma_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}))$ is strictly positive. *This enhancement in the rank structure fundamentally guarantees a reduction in the risk of convergence stagnation, thereby highlighting the crucial role of residual connections in maintaining the stability and effectiveness of the model under extreme conditions.*

For further illustration, we consider data $\boldsymbol{X}$ following $\mathcal{N}(0,1)$ and conduct a series of simulations to investigate the variation in the magnitude of $\frac{\sigma_{\min}^2(\boldsymbol{Z}^{(0)}(\boldsymbol{X}))}{\|\boldsymbol{Z}^{(0)}(\boldsymbol{X})\|} = \frac{\sigma_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}))}{\kappa(\boldsymbol{Z}^{(0)}(\boldsymbol{X}))}$ under different parameter



(a) (b) (c)

(d) (e) (f)

Figure 3: Magnitude of $\frac{\sigma_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}))}{\kappa(\boldsymbol{Z}^{(0)}(\boldsymbol{X}))}$ with and without residual connection for (a) different M, (b) different d, (c) different d_{QK}, (d) different γ_Q, (e) different γ_K, and (f) different γ_V. Here, in (a), $d = d_{QK} = 10$, $\gamma_Q = \gamma_K = \gamma_V = 0.1$; in (b), $M = d_{QK} = 10$, $\gamma_Q = \gamma_K = \gamma_V = 0.1$; in (c), $M = d = 10$, $\gamma_Q = \gamma_K = \gamma_V = 0.1$; in (d), $M = d = d_{QK} = 10$, $\gamma_K = \gamma_V = 0.1$; in (e), $M = d = d_{QK} = 10$, $\gamma_Q = \gamma_V = 0.1$; in (f), $M = d = d_{QK} = 10$, $\gamma_Q = \gamma_K = 0.1$.

settings. The elements of $\boldsymbol{W}_Q \in \mathbb{R}^{d\times d_{QK}}$, $\boldsymbol{W}_K \in \mathbb{R}^{d\times d_{QK}}$ and $\boldsymbol{W}_V \in \mathbb{R}^{d\times d}$ are independently drawn from the Gaussian distributions $\mathcal{N}(0,\gamma_Q^2)$, $\mathcal{N}(0,\gamma_K^2)$ and $\mathcal{N}(0,\gamma_V^2)$, respectively. From Figure 3a, we observe that as the number of discrete tokens M increases, the magnitude of $\frac{\sigma_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}))}{\kappa(\boldsymbol{Z}^{(0)}(\boldsymbol{X}))}$ consistently rises, regardless of the presence of residual connections. In contrast, Figure 3b reveals that when varying the dimensionality d, there exists an optimal value where the magnitude of $\frac{\sigma_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}))}{\kappa(\boldsymbol{Z}^{(0)}(\boldsymbol{X}))}$ reaches its maximum for the model without the residual connection. Moreover, as shown in Figure 3c, varying the intermediate dimension d_{QK} exhibits a relatively stable magnitude of $\frac{\sigma_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}))}{\kappa(\boldsymbol{Z}^{(0)}(\boldsymbol{X}))}$ across both cases, indicating that this parameter has a limited effect on the model's stability. Furthermore, as depicted in Figures 3d to 3f, increasing the variance of the projection matrices leads to a stable magnitude of $\frac{\sigma_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}))}{\kappa(\boldsymbol{Z}^{(0)}(\boldsymbol{X}))}$ when residual connections are present. In contrast, the magnitude increases significantly when residual connections are absent. This disparity arises because when the variances γ_Q^2, γ_K^2, or γ_V^2 are small, the residual term $\boldsymbol{X}$ in the expression $\boldsymbol{Z}(\boldsymbol{X}) = \mathrm{Attn}(\boldsymbol{X}) + \boldsymbol{X}$ predominantly contributes to the output, thereby stabilizing the magnitude.

Lastly, we validate our theoretical findings through an experiment on synthetic data. Specifically, we generate matrices $\{\boldsymbol{X}_p, \boldsymbol{Y}_p\}_{p=1}^{10} \in \{\mathbb{R}^{10\times 100}, \mathbb{R}^{10\times 1}\}$, where each element follows the standard Gaussian distribution. The parameters are set as follows: $d = d_{QK} = d_1/4 = 1000$, learning rate $\mu = 0.1$, and each element in the initial weight matrices $\{\boldsymbol{W}_1^{(0)}, \boldsymbol{W}_2^{(0)}, \boldsymbol{W}_Q^{(0)}, \boldsymbol{W}_K^{(0)}, \boldsymbol{W}_V^{(0)}, \boldsymbol{W}_U^{(0)}\}$ follows $\mathcal{N}(0, 10^{-3})$ and $\mathcal{N}(0, 10^{-2})$ for the cases with and without residual connections, respectively. We compute $\frac{\min_p \sigma_{\min}^2(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p))}{\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|}$ for Transformers trained with and without residual connections, obtaining values of 25.38 and 1.51, respectively. As shown in Figure 1, the convergence rate of Transformer with residual connections is notably faster compared to those without, which is consistent with our theoretical findings.

Global minimum Next, we illustrate that the solution can converge to a global minimum, denoted as $\boldsymbol{\theta}^\star$, satisfying $\Phi(\boldsymbol{\theta}^\star) = 0$. This result highlights the theoretical guarantee of achieving optimal parameter estimation in Transformer models through gradient descent.

Corollary 1 *Under the same setting as in Theorem 1, utilizing gradient descent, the following convergence rate holds:*

$$\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^\star\|_2 \leq (1-\mu\alpha)^{\frac{k}{2}} \frac{C_W}{\alpha} \Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}). \tag{11}$$

where the learning rate satisfies $\mu \leq \min\{\frac{1}{C}, \frac{1}{\alpha}\}$ in which C and α have been defined in Theorem 1. In addition, we define $C_W = O(\overline{\lambda} \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\| + (1+\overline{\lambda}^2)(\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\| + \overline{\lambda} \max_p \|\boldsymbol{X}_p\|) + \max_{i,p} \|\boldsymbol{X}_p(i,:)\|_2 \|\boldsymbol{X}_p\|^2 (1+\overline{\lambda}^2)\overline{\lambda}^3)$.

The proof has been provided in Appendix D. This result guarantees convergence to the global minimum, given that the step size μ is properly chosen and the initialization conditions are satisfied. Specifically, (11) highlights that as the number of iterations k increases, the parameter estimate $\boldsymbol{\theta}^{(k)}$ converges to $\boldsymbol{\theta}^\star$ at a linear convergence rate.

Extension to an L-Layer Transformer Architecture The previously discussed model can be naturally extended to an L-layer Transformer architecture. Specifically, we consider the following model formulation:

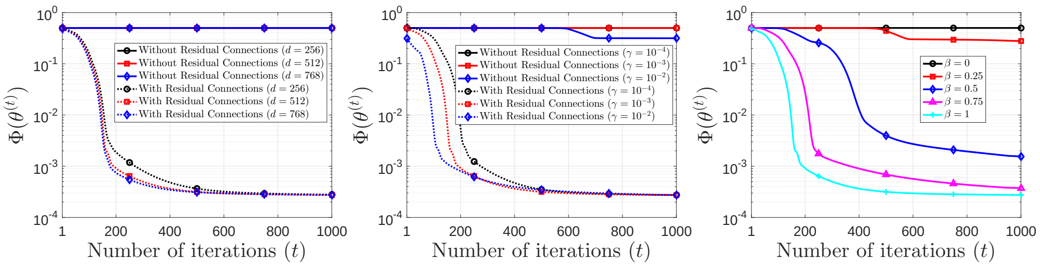
$$\begin{cases} \boldsymbol{Z}_1(\boldsymbol{X}) = \text{Attn}_1(\boldsymbol{X}) + \boldsymbol{X}, \\ \boldsymbol{F}_1(\boldsymbol{X}) = \text{FFN}_1(\boldsymbol{Z}_1) + \boldsymbol{Z}_1 \\ \boldsymbol{Z}_j(\boldsymbol{X}) = \text{Attn}_j(\boldsymbol{F}_{j-1}) + \boldsymbol{F}_{j-1}, j = 2, \ldots, L-1, \\ \boldsymbol{F}_j(\boldsymbol{X}) = \text{FFN}_j(\boldsymbol{Z}_j) + \boldsymbol{Z}_j, j = 2, \ldots, L-1, \\ \boldsymbol{Z}_L(\boldsymbol{X}) = \text{Attn}_L(\boldsymbol{F}_{L-1}) + \boldsymbol{F}_{L-1}, \\ F_{\boldsymbol{\Theta}_L}(\boldsymbol{X}) = (\text{FFN}_L(\boldsymbol{Z}_L) + \boldsymbol{Z}_L)\boldsymbol{W}_U. \end{cases} \tag{12}$$

Here Attn_j and FFN_j for $j = 1, \ldots, L$ denote the softmax attention and feedforward network layers, respectively, each characterized by distinct weight matrices, and $\boldsymbol{\Theta}_L$ is a set of unknown weight matrices. This extension allows us to conduct a convergence analysis of gradient descent within the context of an L-layer Transformer architecture.

To formalize the analysis, we define the objective function as follows: let $f_{\boldsymbol{\theta}_L}(\boldsymbol{X}) = \text{vec}([F_{\boldsymbol{\Theta}_L}(\boldsymbol{X}_1)^\top \cdots F_{\boldsymbol{\Theta}_L}(\boldsymbol{X}_p)^\top]^\top)$ where $\boldsymbol{\theta}_L$ is formed by concatenating all vectorized unknown weight matrices of the Transformer model. We then consider the standard least-squares problem: $\Phi_L(\boldsymbol{\theta}_L) = \frac{1}{2}\|f_{\boldsymbol{\theta}_L}(\boldsymbol{X}) - \boldsymbol{y}\|_2^2$. Following the convergence analysis outlined in Theorem 1, we present the following result as an informal extension. Given appropriate initialization, the estimator obtained via gradient descent approximately satisfies the following bound: $\Phi_L(\boldsymbol{\theta}_L^{(t+1)}) \leq (1-\mu\alpha(L))\Phi_L(\boldsymbol{\theta}_L^{(t)})$. The learning rate μ is chosen to satisfy $\mu \leq \min\{\frac{1}{C(L)}, \frac{1}{\alpha(L)}\}$, where $C(L)$ and $\alpha(L)$ are functions dependent on the number of layers L. This result guarantees a linear convergence rate under the proper initialization condition. Moreover, when considering the last layer, $\boldsymbol{F}_{L-1}(\boldsymbol{X})$ in (12) corresponds to the input matrix in the single-layer Transformer model. Analogous to the previously discussed scenario, given a dataset $\{\boldsymbol{X}_p, \boldsymbol{Y}_p\}_{p=1}^P$, the convergence rate $1-\mu\alpha(L)$ is influenced by the term $1 - C_{1,L} \min_p \sigma_{\min}^2(\boldsymbol{Z}_L^{(0)}(\boldsymbol{X}_p)) / \max_p \|\boldsymbol{Z}_L^{(0)}(\boldsymbol{X}_p)\|$, where $C_{1,L}$ is a positive constant ensuring that this term remains below 1. Here, $\boldsymbol{Z}_L^{(0)}(\boldsymbol{X}_p)$ corresponds to the initial output of the last softmax attention layer. Analogous to the single-layer case, residual connections in the multi-layer setting significantly mitigate the potential ill-conditioning of $\boldsymbol{Z}_L^{(0)}(\boldsymbol{X}_p)$, thereby enhancing numerical stability.

5 Experimental Results

In this section, we investigate the performance of single-layer Transformers with and without residual connections using real-world data. To this end, we employ the Jena Climate Dataset [ASTS20], a multivariate time series dataset widely used in recent studies [DCX+23, VA24]. This dataset comprises weather measurements recorded at 10-minute intervals over several years, capturing various environmental variables, including temperature, pressure, humidity, and wind conditions. Following standard practice [DCX+23, VA24], we adopt a sliding window technique to generate fixed-length input-output pairs for sequence-to-sequence regression, using a window size of $M = 24$. From the complete dataset, we randomly select 5000 instances to train the Transformers. In our experiments, we utilize the ReLU activation function and set the dimension parameters as $d = d_{QK} = \frac{d_1}{4}$. The elements of all weight matrices at initialization ($t = 0$) are drawn from the normal distribution $\mathcal{N}(0, \gamma^2)$. We set the learning rate of gradient descent to $\mu = 0.01$.



(a) (b) (c)

Figure 4: Comparison of training dynamics of 1-layer Transformers with and without residual connections for (a) different d, (b) different γ, (c) different β. Here, in (a) $\gamma = 10^{-3}$, (b) $d = 512$, (c) $\gamma = 10^{-3}$, $d = 512$.

In the first experiment, we examine the performance of Transformers with varying values of d. As shown in Figure 4a, we observe that Transformers equipped with residual connections exhibit stable convergence behavior across different values of d, with the convergence rate improving as the matrix size increases. This phenomenon can be attributed to the enhanced capacity of larger-dimensional models to capture complex patterns. In contrast, the learning dynamics of Transformers without residual connections present a stark difference. Regardless of the matrix size, the convergence stagnates around the initialization values, failing to achieve significant loss reduction even after extensive iterations. This behavior indicates that the optimization process becomes trapped near the initial point, emphasizing the pivotal role of residual connections in breaking such optimization bottlenecks and achieving stable convergence. For $d \in \{256, 512, 768\}$, the values of $\frac{\min_p \sigma^2_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p))}{\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|}$ are respectively $\{0.4, 0.56, 0.63\}$ and $\{1.17, 7.74, 56.3\} \times 10^{-14}$ for models with and without residual connections, respectively, further supporting our previous analysis.

In the second experiment, we investigate the performance of Transformers with varying values of γ. As illustrated in Figure 4b, we systematically vary γ to assess its impact on convergence speed. When employing residual connections, we observe that increasing the standard deviation γ from 10^{-4} to 10^{-2} consistently enhances the convergence rate, highlighting that an appropriate initialization scale is essential for efficient learning. For Transformers without residual connections, the effect of γ is notably different. While setting $\gamma = 10^{-2}$ marginally reduces the training loss compared to smaller initialization scales, the convergence remains slow and susceptible to stagnation. This outcome reveals that, despite slight improvements, the inherent difficulty in training Transformers without residual connections persists, regardless of the choice of initialization scale. Furthermore, for $\gamma \in \{10^{-4}, 10^{-3}, 10^{-2}\}$, the values of $\frac{\min_p \sigma^2_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p))}{\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|}$ are respectively $\{0.27, 0.56, 0.64\}$ and $\{2.26, 7.74, 58.7\} \times 10^{-14}$ for models with and without residual connections, respectively, further illustrating that the convergence rate is significantly influenced by the presence of residual connections.

In the last experiment, we investigate the impact of residual connections on the training dynamics. Drawing inspiration from [ZJG+24], where a residual coefficient β is introduced to construct the model as $F_{\boldsymbol{\Theta}}(\boldsymbol{X}) = (\text{FFN}(\text{Attn}(\boldsymbol{X}) + \beta\boldsymbol{X}) + \beta(\text{Attn}(\boldsymbol{X}) + \beta\boldsymbol{X}))\boldsymbol{W}_U$, it was demonstrated that setting β smaller than 1 can enhance test performance. Motivated by this observation, we systematically examine the effect of varying the residual coefficient β on the convergence rate. As depicted in Figure 4c, we observe that increasing the value of β accelerates convergence. Although the convergence rate for $\beta = 0.5$ or 0.75 is slower than that of $\beta = 1$, it remains significantly faster than the model without residual connections, reinforcing the critical role of residual coefficients in efficient learning. Additionally, for $\beta \in \{0, 0.25, 0.5, 0.75, 1\}$, the values of $\frac{\min_p \sigma^2_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p))}{\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|}$ are given by $\{7.74 \times 10^{-14}, 0.16, 0.32, 0.48, 0.56\}$, highlighting the substantial influence of residual connections on convergence speed.

6 Conclusion

This paper presented a comprehensive analysis of the convergence behavior of a single-layer Transformer architecture, incorporating a single-head self-attention mechanism, a feedforward neural network, and residual connections. Our theoretical investigation established that, under proper initialization, the gradient descent method exhibits a linear convergence rate. The convergence speed is primarily determined by the minimum and maximum singular values of the output matrix produced by the attention layer. Extending this analysis to a multi-layer Transformer architecture, we demonstrated that the convergence rate is influenced by the output matrix of the final attention layer, emphasizing the critical role of residual connections in maintaining numerical stability. Experimental results further substantiated our theoretical insights, illustrating that incorporating residual connections significantly enhances convergence stability. Our findings underscore the importance of residual connections in optimizing Transformer training dynamics, providing a theoretical foundation for their effectiveness in improving convergence rates and numerical stability.

7 Acknowledgement

We acknowledge support from NSF grants CCF-2241298, IIS-2402952 and ECCS-2409701, as well as the ORAU Ralph E. Powe Junior Faculty Enhancement Award.

References

[AAA+23] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[ACDS23] Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, and Suvrit Sra. Transformers learn to implement preconditioned gradient descent for in-context learning. *Advances in Neural Information Processing Systems*, 36:45614–45650, 2023.

[ACGH18] Sanjeev Arora, Nadav Cohen, Noah Golowich, and Wei Hu. A convergence analysis of gradient descent for deep linear neural networks. *arXiv preprint arXiv:1810.02281*, 2018.

[ASTS20] Prabhanshu Attri, Yashika Sharma, Kristi Takach, and Falak Shah. Timeseries forecasting for weather prediction. *Keras Tutorial*, 2020.

[ATLZO23] Davoud Ataee Tarzanagh, Yingcong Li, Xuechen Zhang, and Samet Oymak. Max-margin token selection in attention mechanism. *Advances in neural information processing systems*, 36:48314–48362, 2023.

[AZLS19] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International conference on machine learning*, pages 242–252. PMLR, 2019.

[BCP24] Blake Bordelon, Hamza Chaudhry, and Cengiz Pehlevan. Infinite limits of multi-head transformer dynamics. *Advances in Neural Information Processing Systems*, 37:35824–35878, 2024.

[BHL18] Peter Bartlett, Dave Helmbold, and Philip Long. Gradient descent with identity initialization efficiently learns positive definite linear transformations by deep residual networks. In *International conference on machine learning*, pages 521–530. PMLR, 2018.

[BMR+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Cha22] Sourav Chatterjee. Convergence of gradient descent for deep neural networks. *arXiv preprint arXiv:2203.16462*, 2022.

[CLR+21] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

[CND+23] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023.

[CSWY24] Siyu Chen, Heejune Sheen, Tianhao Wang, and Zhuoran Yang. Training dynamics of multi-head softmax attention for in-context learning: Emergence, convergence, and optimality. *arXiv preprint arXiv:2402.19442*, 2024.

[CZL+19] Qiwei Chen, Huan Zhao, Wei Li, Pipei Huang, and Wenwu Ou. Behavior sequence transformer for e-commerce recommendation in alibaba. In *Proceedings of the 1st international workshop on deep learning practice for high-dimensional sparse data*, pages 1–4, 2019.

[DBK+20] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, G Heigold, S Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2020.

[DCX+23] Ngoc Thanh Duong, Yu-Chieh Chien, Heng Xiang, Sifan Li, Haofei Zheng, Yufei Shi, and Kah-Wee Ang. Dynamic ferroelectric transistor-based reservoir computing for spatiotemporal information processing. *Advanced Intelligent Systems*, 5(6):2300009, 2023.

[EGKZ22] Benjamin L Edelman, Surbhi Goel, Sham Kakade, and Cyril Zhang. Inductive biases and variable creation in self-attention mechanisms. In *International Conference on Machine Learning*, pages 5793–5831. PMLR, 2022.

[FCJS24] Deqing Fu, Tian-qi Chen, Robin Jia, and Vatsal Sharan. Transformers learn to achieve second-order convergence rates for in-context linear regression. *Advances in Neural Information Processing Systems*, 37:98675–98716, 2024.

[GBW+24] Dirk Groeneveld, Iz Beltagy, Pete Walsh, Akshita Bhagia, Rodney Kinney, Oyvind Tafjord, Ananya Harsh Jha, Hamish Ivison, Ian Magnusson, Yizhong Wang, et al. Olmo: Accelerating the science of language models. *arXiv preprint arXiv:2402.00838*, 2024.

[HCL23] Yu Huang, Yuan Cheng, and Yingbin Liang. In-context convergence of transformers. *arXiv preprint arXiv:2310.05249*, 2023.

[HLY24] Ruiquan Huang, Yingbin Liang, and Jing Yang. Non-asymptotic convergence of training transformers for next-token prediction. *arXiv preprint arXiv:2409.17335*, 2024.

[HM16] Moritz Hardt and Tengyu Ma. Identity matters in deep learning. *arXiv preprint arXiv:1611.04231*, 2016.

[IHL+24] Muhammed Emrullah Ildiz, Yixiao Huang, Yingcong Li, Ankit Singh Rawat, and Samet Oymak. From self-attention to markov models: Unveiling the dynamics of generative transformers. In *International Conference on Machine Learning*, pages 20955–20982. PMLR, 2024.

[JLL21] Michael Janner, Qiyang Li, and Sergey Levine. Offline reinforcement learning as one big sequence modeling problem. *Advances in neural information processing systems*, 34:1273–1286, 2021.

[KLJ+23] Seungnyun Kim, Anho Lee, Hyungyu Ju, Khoa Anh Ngo, Jihoon Moon, and Byonghyo Shim. Transformer-based channel parameter acquisition for terahertz ultra-massive mimo systems. *IEEE Transactions on Vehicular Technology*, 72(11):15127–15132, 2023.

[LCZ+19] Tianyi Liu, Minshuo Chen, Mo Zhou, Simon S Du, Enlu Zhou, and Tuo Zhao. Towards understanding the importance of shortcut connections in residual networks. *Advances in neural information processing systems*, 32, 2019.

[LWL+24] Hongkang Li, Meng Wang, Songtao Lu, Xiaodong Cui, and Pin-Yu Chen. Training nonlinear transformers for efficient in-context learning: A theoretical learning and generalization analysis. *arXiv preprint arXiv:2402.15607*, 3, 2024.

[MBG+24] Ashok Vardhan Makkuva, Marco Bondaschi, Adway Girish, Alliot Nagle, Hyeji Kim, Michael Gastpar, and Chanakya Ekbote. Local to global: Learning dynamics and effect of initialization for transformers. *Advances in Neural Information Processing Systems*, 37:86243–86308, 2024.

[MHM23] Arvind Mahankali, Tatsunori B Hashimoto, and Tengyu Ma. One step of gradient descent is provably the optimal in-context learner with one layer of linear self-attention. *arXiv preprint arXiv:2307.03576*, 2023.

[MSD+25] Hailan Ma, Zhenhong Sun, Daoyi Dong, Chunlin Chen, and Herschel Rabitz. Tomography of quantum states from structured measurements via quantum-aware transformer. *IEEE Transactions on Cybernetics*, 2025.

[NAB+22] Lorenzo Noci, Sotiris Anagnostidis, Luca Biggio, Antonio Orvieto, Sidak Pal Singh, and Aurelien Lucchi. Signal propagation in transformers: Theoretical perspectives and the role of rank collapse. *Advances in Neural Information Processing Systems*, 35:27198–27211, 2022.

[NM20] Quynh N Nguyen and Marco Mondelli. Global convergence of deep networks with one wide layer followed by pyramidal topology. *Advances in Neural Information Processing Systems*, 33:11961–11972, 2020.

[NMM21] Quynh Nguyen, Marco Mondelli, and Guido F Montufar. Tight bounds on the smallest eigenvalue of the neural tangent kernel for deep relu networks. In *International Conference on Machine Learning*, pages 8119–8129. PMLR, 2021.

[QTZ24] Zhen Qin, Xuwei Tan, and Zhihui Zhu. Convergence analysis for learning orthonormal deep linear neural networks. *IEEE Signal Processing Letters*, 2024.

[RWC+19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[RWL24] Yunwei Ren, Zixuan Wang, and Jason D Lee. Learning and transferring sparse contextual bigrams with linear transformers. *arXiv preprint arXiv:2410.23438*, 2024.

[Sha19] Ohad Shamir. Exponential convergence time of gradient descent for one-dimensional deep linear neural networks. In *Conference on Learning Theory*, pages 2691–2713. PMLR, 2019.

[SHZ+24] Bingqing Song, Boran Han, Shuai Zhang, Jie Ding, and Mingyi Hong. Unraveling the gradient descent dynamics of transformers. *Advances in Neural Information Processing Systems*, 37:92317–92351, 2024.

[SWJS24] Michael Scholkemper, Xinyi Wu, Ali Jadbabaie, and Michael T Schaub. Residual connections and normalization can provably prevent oversmoothing in gnns. *arXiv preprint arXiv:2406.02997*, 2024.

[TBL+19] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for computational linguistics. Meeting*, volume 2019, page 6558, 2019.

[TLTO23] Davoud Ataee Tarzanagh, Yingcong Li, Christos Thrampoulidis, and Samet Oymak. Transformers as support vector machines. *arXiv preprint arXiv:2308.16898*, 2023.

[TMS$^+$23] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023.

[TWCD23] Yuandong Tian, Yiping Wang, Beidi Chen, and Simon S Du. Scan and snap: Understanding training dynamics and token composition in 1-layer transformer. *Advances in neural information processing systems*, 36:71911–71947, 2023.

[VA24] Alexander Okhuese Victor and Muhammad Intizar Ali. Enhancing time series data predictions: A survey of augmentation techniques and model performances. In *Proceedings of the 2024 Australasian Computer Science Week*, pages 1–13. 2024.

[Ver10] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv:1011.3027*, 2010.

[VSP$^+$17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[WLCC23] Yongtao Wu, Fanghui Liu, Grigorios Chrysos, and Volkan Cevher. On the convergence of encoder-only shallow transformers. *Advances in Neural Information Processing Systems*, 36:52197–52237, 2023.

[YHLC24] Tong Yang, Yu Huang, Yingbin Liang, and Yuejie Chi. In-context learning with representations: Contextual generalization of trained transformers. *arXiv preprint arXiv:2408.10147*, 2024.

[ZFB24] Ruiqi Zhang, Spencer Frei, and Peter L Bartlett. Trained transformers learn linear models in-context. *Journal of Machine Learning Research*, 25(49):1–55, 2024.

[ZJG$^+$24] Xiao Zhang, Ruoxi Jiang, William Gao, Rebecca Willett, and Michael Maire. Residual connections harm generative representation learning. *arXiv preprint arXiv:2404.10947*, 2024.

[ZLG20] Difan Zou, Philip M Long, and Quanquan Gu. On the global convergence of training deep linear resnets. *arXiv preprint arXiv:2003.01094*, 2020.

[ZSLS25] Yedi Zhang, Aaditya K Singh, Peter E Latham, and Andrew Saxe. Training dynamics of in-context learning in linear attention. *arXiv preprint arXiv:2501.16265*, 2025.

[ZZS$^+$18] Guorui Zhou, Xiaoqiang Zhu, Chenru Song, Ying Fan, Han Zhu, Xiao Ma, Yanghui Yan, Junqi Jin, Han Li, and Kun Gai. Deep interest network for click-through rate prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 1059–1068, 2018.

Appendices

A Gradient Derivations

To simplify notation, we omit the explicit dependence of the loss function on all weight matrices and write $\nabla_{\boldsymbol{W}_b} L$ instead of $\nabla_{\boldsymbol{W}_b} L(\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V, \boldsymbol{W}_U)$ for $b \in \{1, 2, Q, K, V, U\}$. The gradients are summarized as follows:

$$\nabla_{\boldsymbol{W}_1} L = \sum_{p=1}^{P} \boldsymbol{Z}(\boldsymbol{X}_p)^\top \Big(\phi_r'(\boldsymbol{Z}(\boldsymbol{X}_p)\boldsymbol{W}_1) \odot \Big((F_{\boldsymbol{\Theta}}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\boldsymbol{W}_U^\top \boldsymbol{W}_2^\top \Big) \Big), \tag{13}$$

$$\nabla_{\boldsymbol{W}_2} L = \sum_{p=1}^{P} (\phi_r(\boldsymbol{Z}(\boldsymbol{X}_p)\boldsymbol{W}_1))^\top (F_{\boldsymbol{\Theta}}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\boldsymbol{W}_U^\top, \tag{14}$$

$$\nabla_{\boldsymbol{W}_U} L = \sum_{p=1}^{P} (\phi_r(\boldsymbol{Z}(\boldsymbol{X}_p)\boldsymbol{W}_1)\boldsymbol{W}_2 + \boldsymbol{Z}(\boldsymbol{X}_p))^\top (F_{\boldsymbol{\Theta}}(\boldsymbol{X}_p) - \boldsymbol{Y}_p), \tag{15}$$

$$\begin{aligned}\nabla_{\boldsymbol{W}_V} L = &\sum_{p=1}^{P} \boldsymbol{X}_p^\top \phi_s\Big(\frac{\boldsymbol{X}_p \boldsymbol{W}_Q \boldsymbol{W}_K^\top \boldsymbol{X}_p^\top}{\sqrt{d_{QK}}}\Big)^\top \Big(\phi_r'(\boldsymbol{Z}(\boldsymbol{X}_p)\boldsymbol{W}_1) \odot \Big((F_{\boldsymbol{\Theta}}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\boldsymbol{W}_U^\top \boldsymbol{W}_2^\top \Big) \Big) \boldsymbol{W}_1^\top \\ &+ \sum_{p=1}^{P} \boldsymbol{X}_p^\top \phi_s\Big(\frac{\boldsymbol{X}_p \boldsymbol{W}_Q \boldsymbol{W}_K^\top \boldsymbol{X}_p^\top}{\sqrt{d_{QK}}}\Big)^\top (F_{\boldsymbol{\Theta}}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\boldsymbol{W}_U^\top,\end{aligned} \tag{16}$$

$$\begin{aligned}\nabla_{\boldsymbol{W}_Q} L = &\sum_{p=1}^{P}\sum_{i=1}^{M} \boldsymbol{X}_p^\top(i,:)(F_{\boldsymbol{\Theta}}(\boldsymbol{X}_p)(i,:) - \boldsymbol{Y}_p(i,:))\boldsymbol{W}_U^\top \boldsymbol{W}_V^\top \boldsymbol{X}_p^\top \phi_s'\Big(\frac{\boldsymbol{X}_p(i,:) \boldsymbol{W}_Q \boldsymbol{W}_K^\top \boldsymbol{X}_p^\top}{\sqrt{d_{QK}}}\Big)\boldsymbol{X}_p \boldsymbol{W}_K \\ &+ \sum_{p=1}^{P}\sum_{i=1}^{M} \boldsymbol{X}_p^\top(i,:)\Big(\phi_r'(\boldsymbol{Z}(\boldsymbol{X}_p)(i,:)\boldsymbol{W}_1) \odot \Big((F_{\boldsymbol{\Theta}}(\boldsymbol{X}_p)(i,:) - \boldsymbol{Y}_p(i,:))\boldsymbol{W}_U^\top \boldsymbol{W}_2^\top \Big) \Big) \\ &\cdot \boldsymbol{W}_1^\top \boldsymbol{W}_V^\top \boldsymbol{X}_p^\top \phi_s'\Big(\frac{\boldsymbol{X}_p(i,:) \boldsymbol{W}_Q \boldsymbol{W}_K^\top \boldsymbol{X}_p^\top}{\sqrt{d_{QK}}}\Big)\boldsymbol{X}_p \boldsymbol{W}_K,\end{aligned} \tag{17}$$

$$\begin{aligned}\nabla_{\boldsymbol{W}_K} L = &\sum_{p=1}^{P}\sum_{i=1}^{M} \boldsymbol{X}_p^\top \Big((F_{\boldsymbol{\Theta}}(\boldsymbol{X}_p)(i,:) - \boldsymbol{Y}_p(i,:))\boldsymbol{W}_U^\top \boldsymbol{W}_V^\top \boldsymbol{X}_p^\top \phi_s'\Big(\frac{\boldsymbol{X}_p \boldsymbol{W}_Q \boldsymbol{W}_K^\top \boldsymbol{X}_p^\top}{\sqrt{d_{QK}}}\Big)\Big)^\top \boldsymbol{X}_p(i,:)\boldsymbol{W}_Q \\ &+ \sum_{p=1}^{P}\sum_{i=1}^{M} \boldsymbol{X}_p^\top \Big(\Big(\phi_r'(\boldsymbol{Z}(\boldsymbol{X}_p)(i,:)\boldsymbol{W}_1) \odot \Big((F_{\boldsymbol{\Theta}}(\boldsymbol{X}_p)(i,:) - \boldsymbol{Y}_p(i,:))\boldsymbol{W}_U^\top \boldsymbol{W}_2^\top \Big) \Big) \\ &\cdot \boldsymbol{W}_1^\top \boldsymbol{W}_V^\top \boldsymbol{X}_p^\top \phi_s'\Big(\frac{\boldsymbol{X}_p(i,:) \boldsymbol{W}_Q \boldsymbol{W}_K^\top \boldsymbol{X}_p^\top}{\sqrt{d_{QK}}}\Big)\Big)^\top \boldsymbol{X}_p(i,:)\boldsymbol{W}_Q.\end{aligned} \tag{18}$$

Note that to derive the last two gradients, we need to rewrite the loss function as $L(\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{W}_Q, \boldsymbol{W}_K, \boldsymbol{W}_V, \boldsymbol{W}_U) = \frac{1}{2}\sum_{p=1}^{P}\sum_{i=1}^{M} \|(\phi_r((\phi_s(\frac{\boldsymbol{X}_p(i,:)\boldsymbol{W}_Q \boldsymbol{W}_K^\top \boldsymbol{X}_p^\top}{\sqrt{d_{QK}}})\boldsymbol{X}_p \boldsymbol{W}_V + \boldsymbol{X}_p(i,:))\boldsymbol{W}_1)\boldsymbol{W}_2 + \phi_s(\frac{\boldsymbol{X}_p(i,:)\boldsymbol{W}_Q \boldsymbol{W}_K^\top \boldsymbol{X}_p^\top}{\sqrt{d_{QK}}})\boldsymbol{X}_p \boldsymbol{W}_V + \boldsymbol{X}_p(i,:))\boldsymbol{W}_U - \boldsymbol{Y}_p(i,:)\|_2^2$. Here, the softmax function $\phi_s : \mathbb{R}^{1\times M} \to \mathbb{R}^{1\times M}$ is defined as

$$\phi_s(\boldsymbol{a}) = \begin{bmatrix} \frac{e^{\boldsymbol{a}(1,:)}}{\sum_{i=1}^{M} e^{\boldsymbol{a}(i,:)}} & \cdots & \frac{e^{\boldsymbol{a}(M,:)}}{\sum_{i=1}^{M} e^{\boldsymbol{a}(i,:)}} \end{bmatrix}. \tag{19}$$

Furthermore, we define the Jacobian of the softmax function, evaluated at

$$\phi_s'\Big(\frac{\boldsymbol{X}_p(i,:) \boldsymbol{W}_Q \boldsymbol{W}_K^\top \boldsymbol{X}_p^\top}{\sqrt{d_{QK}}}\Big) = \text{diag}(\boldsymbol{s}) - \boldsymbol{s}^\top \boldsymbol{s}, \tag{20}$$

where $\boldsymbol{s} = \phi_s(\frac{\boldsymbol{X}_p(i,:)\boldsymbol{W}_Q \boldsymbol{W}_K^\top \boldsymbol{X}_p^\top}{\sqrt{d_{QK}}})$ is the softmax output vector.

B Detailed version of Theorem 1

Before proceeding to the proof of Theorem 1, we first present and prove the following auxiliary theorem.

Theorem 2 *Given a dataset $\{\boldsymbol{X}_p, \boldsymbol{Y}_p\}_{p=1}^{P}$, we consider a single-layer Transformer architecture equipped with a single-head self-attention mechanism, a feedforward neural network and residual connections. The goal is to model or approximate the underlying distribution of the data. Define $\alpha = \frac{\sigma_{\min}^2(\boldsymbol{W}_U^{(0)}) \min_p \sigma_{\min}^2(\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)}))}{16}$. When the initialization satisfies*

$$\begin{aligned}\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}) \ \leq \ \min\Bigg\{ & \frac{2\sqrt{2}\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}{27\sqrt{P}\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|} \cdot \min\left\{\frac{\sigma_{\min}(\boldsymbol{W}_1^{(0)})}{\|\boldsymbol{W}_2^{(0)}\|\|\boldsymbol{W}_U^{(0)}\|}, \frac{\sigma_{\min}(\boldsymbol{W}_2^{(0)})}{\|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_U^{(0)}\|}\right\}, \\ & \frac{2\sqrt{2}\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}{27\sqrt{P}(1+\|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_2^{(0)}\|)} \cdot \min\left\{\frac{\sigma_{\min}(\boldsymbol{W}_U^{(0)})}{\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|}, \frac{\sigma_{\min}(\boldsymbol{W}_V^{(0)})}{\sqrt{M}\max_p \|\boldsymbol{X}_p\|\|\boldsymbol{W}_U^{(0)}\|}\right\}, \\ & \frac{4\sqrt{2}\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}{243\sqrt{MP}\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2(1+\|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_2^{(0)}\|)\|\boldsymbol{W}_U^{(0)}\|\|\boldsymbol{W}_V^{(0)}\|} \\ & \cdot \min\left\{\frac{\sigma_{\min}(\boldsymbol{W}_Q^{(0)})}{\|\boldsymbol{W}_K^{(0)}\|}, \frac{\sigma_{\min}(\boldsymbol{W}_K^{(0)})}{|\|\boldsymbol{W}_Q^{(0)}\|}\right\}, \frac{\min_p \sigma_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p))}{2C_1}, 2\sqrt{C_2}, \\ & \frac{\min_p \sigma_{\min}(\phi_r(\boldsymbol{Z}^{(0)}\boldsymbol{W}_1^{(0)}))}{3C_1\|\boldsymbol{W}_1^{(0)}\| + \frac{27\sqrt{2P}}{4\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}\|\boldsymbol{Z}^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|\|\boldsymbol{W}_U^{(0)}\|}\Bigg\},\end{aligned} \tag{21}$$

where $C_1 = \frac{2187M\sqrt{2P}\max_p\|\boldsymbol{X}_p\|^3(\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2)\|\boldsymbol{W}_U^{(0)}\|\|\boldsymbol{W}_V^{(0)}\|^2(\|\boldsymbol{W}_Q^{(0)}\|^2+\|\boldsymbol{W}_K^{(0)}\|^2)}{32\sqrt{d_{QK}}\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}(1 + \|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_2^{(0)}\|) + \frac{27\sqrt{2}M\sqrt{P}(\max_p\|\boldsymbol{X}_p\|)^2\|\boldsymbol{W}_U^{(0)}\|}{8\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}(1+\|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_2^{(0)}\|)$ *and* $C_2 = \frac{6561\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\sigma_{\min}^2(\boldsymbol{W}_1^{(0)})}{64} + \frac{81\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\sigma_{\min}^2(\boldsymbol{W}_U^{(0)})}{4} + \big(9\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2 + \frac{729\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2}{16}\big)\big(\frac{27M\sqrt{P}(\max_p\|\boldsymbol{X}_p\|)^2\sigma_{\min}^2(\boldsymbol{W}_V^{(0)})}{4} + \frac{2187M\sqrt{P}\max_p\|\boldsymbol{X}_p\|^3(\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2)\|\boldsymbol{W}_V^{(0)}\|^2\|\boldsymbol{W}_Q^{(0)}\|^2\sigma_{\min}^2(\boldsymbol{W}_K^{(0)})}{16d_{QK}} + \frac{2187M\sqrt{P}\max_p\|\boldsymbol{X}_p\|^3(\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2)\|\boldsymbol{W}_V^{(0)}\|^2\|\boldsymbol{W}_K^{(0)}\|^2\sigma_{\min}^2(\boldsymbol{W}_Q^{(0)})}{16d_{QK}}\big)$. *Using the gradient descent in* (5), *we have*

$$\Phi(\boldsymbol{\theta}^{(t+1)}) \leq (1-\mu\alpha)\Phi(\boldsymbol{\theta}^{(t)}), \tag{22}$$

where the learning rate satisfies $\mu \leq \min\{\frac{1}{C}, \frac{1}{\alpha}\}$, *and we define the following two constants* $C^2 = \frac{2187PC_F}{32}\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2 \cdot \|\boldsymbol{W}_U^{(0)}\|^2(\|\boldsymbol{W}_1^{(0)}\|^2 + \|\boldsymbol{W}_2^{(0)}\|^2) + PC_F\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2 \cdot \big(\frac{2187}{16}\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2 + 27\big) + \frac{PC_F}{d_{QK}}(\max_p\|\boldsymbol{X}_p\|)^6\|\boldsymbol{W}_U^{(0)}\|^2\|\boldsymbol{W}_Q^{(0)}\|^2\|\boldsymbol{W}_K^{(0)}\|^2\big(\frac{2187}{16} + \frac{177147}{256}\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\big) + \frac{2187PC_FM}{4}\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2^2(\max_p\|\boldsymbol{X}_p\|)^4\|\boldsymbol{W}_U^{(0)}\|^2\|\boldsymbol{W}_V^{(0)}\|^2 \cdot (\|\boldsymbol{W}_K^{(0)}\|^2 + \|\boldsymbol{W}_Q^{(0)}\|^2) + \frac{177147PC_FM}{64}\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2^2(\max_p\|\boldsymbol{X}_p\|)^4\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2 \cdot \|\boldsymbol{W}_U^{(0)}\|^2\|\boldsymbol{W}_V^{(0)}\|^2(\|\boldsymbol{W}_K^{(0)}\|^2 + \|\boldsymbol{W}_Q^{(0)}\|^2)$ *and* $C_F = \max\big\{\frac{729\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2}{16}, \big(\frac{729\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2}{16} + 9\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\big)\frac{243M\sqrt{P}\max_p\|\boldsymbol{X}_p\|^3(\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2)\|\boldsymbol{W}_V^{(0)}\|^2\|\boldsymbol{W}_Q^{(0)}\|^2}{4d_{QK}}, \big(\frac{729\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2}{16} + 9\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\big)\frac{243M\sqrt{P}\max_p\|\boldsymbol{X}_p\|^3(\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2)\|\boldsymbol{W}_V^{(0)}\|^2\|\boldsymbol{W}_K^{(0)}\|^2}{4d_{QK}}, 27M\sqrt{P}(\max_p\|\boldsymbol{X}_p\|)^2\big(\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2 + \frac{81\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2}{16}\big), 9\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\big\}$.

Proof We show by induction that, for every $t \geq 0$, the following holds:

$$\begin{cases} \|\boldsymbol{W}_{b,a}^{(t)}\| \leq \frac{3}{2}\|\boldsymbol{W}_b^{(0)}\|, b = 1, 2, U, V, Q, K, \\ \sigma_{\min}(\boldsymbol{W}_{b,a}^{(t)}) \geq \frac{\sigma_{\min}(\boldsymbol{W}_b^{(0)})}{2}, b = 1, 2, U, V, Q, K, \\ \|\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\| \leq \frac{3}{2}\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|, p = 1, \dots, P, \\ \sigma_{\min}(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)) \geq \frac{\sigma_{\min}(\boldsymbol{Z}^{(0)})(\boldsymbol{X}_p)}{2}, p = 1, \dots, P, \\ \sigma_{\min}(\phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)})) \geq \frac{\sigma_{\min}(\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)}))}{2}, p = 1, \dots, P, \\ \Phi(\boldsymbol{\theta}^{(t+1)}) \leq (1 - \mu\alpha)\Phi(\boldsymbol{\theta}^{(t)}). \end{cases} \tag{23}$$

where $\boldsymbol{W}_{b,a}^{(t)} = \boldsymbol{W}_b^{(t)} + a(\boldsymbol{W}_b^{(t+1)} - \boldsymbol{W}_b^{(t)})$ and $\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p) = \boldsymbol{Z}^{(t)}(\boldsymbol{X}_p) + a(\boldsymbol{Z}^{(t+1)}(\boldsymbol{X}_p) - \boldsymbol{Z}^{(t)}(\boldsymbol{X}_p))$, $a \in [0, 1]$.

Step I: Proof of first five inequalities in (23).

Part I We begin by bounding $\|\boldsymbol{W}_{1,a}^{(t)}\|$ and $\sigma_{\min}(\boldsymbol{W}_{1,a}^{(t)})$ where $\boldsymbol{W}_{1,a}^{(t)} = \boldsymbol{W}_1^{(t)} + a(\boldsymbol{W}_1^{(t+1)} - \boldsymbol{W}_1^{(t)})$ with $a \in [0, 1]$. Specifically, we first expand $\|\boldsymbol{W}_{1,a}^{(t)} - \boldsymbol{W}_1^{(0)}\|_F$ as follows:

$$\begin{aligned} & \|\boldsymbol{W}_{1,a}^{(t)} - \boldsymbol{W}_1^{(0)}\|_F \\ \leq\ & \mu \sum_{s=0}^{t} \|\nabla_{\boldsymbol{W}_1} L(\boldsymbol{W}_1^{(s)}, \boldsymbol{W}_2^{(s)}, \boldsymbol{W}_Q^{(s)}, \boldsymbol{W}_K^{(s)}, \boldsymbol{W}_V^{(s)}, \boldsymbol{W}_U^{(s)})\|_F \\ \leq\ & \mu \sum_{s=0}^{t} \sum_{p=1}^{P} \|\boldsymbol{Z}^{(s)}(\boldsymbol{X}_p)\| \|(\phi_r'(\boldsymbol{Z}^{(s)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(s)}) \odot ((F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\boldsymbol{W}_U^{(s)\top}\boldsymbol{W}_2^{(s)\top}))\|_F \\ \leq\ & \mu \sum_{s=0}^{t} \sum_{p=1}^{P} \|\boldsymbol{Z}^{(s)}(\boldsymbol{X}_p)\| \|\boldsymbol{W}_2^{(s)}\| \|\boldsymbol{W}_U^{(s)}\| \|F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}\|_F \\ \leq\ & \frac{27\sqrt{2}\mu\sqrt{P}}{8} \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\| \|\boldsymbol{W}_2^{(0)}\| \|\boldsymbol{W}_U^{(0)}\| \sum_{s=0}^{t} (1 - \mu\alpha)^{\frac{s}{2}} \Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}) \\ =\ & \frac{27\sqrt{2}\mu\sqrt{P}}{8} \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\| \|\boldsymbol{W}_2^{(0)}\| \|\boldsymbol{W}_U^{(0)}\| \Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}) \frac{1 - (1 - \mu\alpha)^{\frac{t+1}{2}}}{1 - (1 - \mu\alpha)^{\frac{1}{2}}} \\ \leq\ & \frac{27\sqrt{2}\sqrt{P}}{8\alpha(1 - (1 - \mu\alpha)^{\frac{1}{2}})} \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\| \|\boldsymbol{W}_2^{(0)}\| \|\boldsymbol{W}_U^{(0)}\| \Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}). \end{aligned} \tag{24}$$

where $\mu \leq \frac{1}{\alpha}$. To further guarantee $\|\boldsymbol{W}_{1,a}^{(t)} - \boldsymbol{W}_1^{(0)}\|_F \leq \frac{\sigma_{\min}(\boldsymbol{W}_1^{(0)})}{2}$, the initialization requirement should satisfy

$$\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}) \leq \frac{2\sqrt{2}\alpha(1 - (1 - \mu\alpha)^{\frac{1}{2}})}{27\sqrt{P}} \frac{\sigma_{\min}(\boldsymbol{W}_1^{(0)})}{\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\| \|\boldsymbol{W}_2^{(0)}\| \|\boldsymbol{W}_U^{(0)}\|}. \tag{25}$$

Based on $\sigma_{\min}(\boldsymbol{W}_1^{(0)}) - \sigma_{\min}(\boldsymbol{W}_{1,a}^{(t)}) \leq \|\boldsymbol{W}_{1,a}^{(t)} - \boldsymbol{W}_1^{(0)}\| \leq \frac{\sigma_{\min}(\boldsymbol{W}_1^{(0)})}{2}$, we can obtain

$$\frac{\sigma_{\min}(\boldsymbol{W}_1^{(0)})}{2} \leq \sigma_{\min}(\boldsymbol{W}_{1,a}^{(t)}) \leq \|\boldsymbol{W}_{1,a}^{(t)}\| \leq \frac{3\|\boldsymbol{W}_1^{(0)}\|}{2}. \tag{26}$$

Similarly, we respectively expand other terms $\|\boldsymbol{W}_{b,a}^{(t)} - \boldsymbol{W}_b^{(0)}\|_F, b = 2, U, V, Q, K$, as follows:

$$
\begin{aligned}
&\|\boldsymbol{W}_{2,a}^{(t)} - \boldsymbol{W}_2^{(0)}\|_F \\
\leq\ & \mu \sum_{s=0}^{t} \sum_{p=1}^{P} \|(\boldsymbol{Z}^{(s)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(s)})^\top (F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\boldsymbol{W}_U^{(s)^\top}\|_F \\
\leq\ & \mu \sum_{s=0}^{t} \sum_{p=1}^{P} \|\boldsymbol{Z}^{(s)}(\boldsymbol{X}_p)\|\|\boldsymbol{W}_1^{(s)}\|\|\boldsymbol{W}_U^{(s)}\|\|F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p\|_F \\
\leq\ & \frac{27\sqrt{2}\sqrt{P}}{8\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})} \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|\|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_U^{(0)}\|\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}), \qquad (27)\\
&\|\boldsymbol{W}_{U,a}^{(t)} - \boldsymbol{W}_U^{(0)}\|_F \\
\leq\ & \mu \sum_{s=0}^{t} \sum_{p=1}^{P} \|(\phi_r(\boldsymbol{Z}^{(s)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(s)})\boldsymbol{W}_2^{(s)})^\top (F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\|_F + \|(\boldsymbol{Z}^{(s)}(\boldsymbol{X}_p))^\top (F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\|_F \\
\leq\ & \mu \sum_{s=0}^{t} \sum_{p=1}^{P} \|\boldsymbol{W}_2^{(s)}\|\|(\boldsymbol{Z}^{(s)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(s)})^\top (F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\|_F + \|\boldsymbol{Z}^{(s)}(\boldsymbol{X}_p)\|\|F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p\|_F \\
\leq\ & \mu \sum_{s=0}^{t} \sum_{p=1}^{P} \|\boldsymbol{W}_2^{(s)}\|\|\boldsymbol{Z}^{(s)}(\boldsymbol{X}_p)\|\|\boldsymbol{W}_1^{(s)}\|\|F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p\|_F + \|\boldsymbol{Z}^{(s)}(\boldsymbol{X}_p)\|\|F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p\|_F \\
\leq\ & \frac{27\sqrt{2}\sqrt{P}}{8\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}(1 + \|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_2^{(0)}\|) \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}), \qquad (28)\\
&\|\boldsymbol{W}_{V,a}^{(t)} - \boldsymbol{W}_V^{(0)}\|_F \\
\leq\ & \mu \sum_{s=0}^{t} \sum_{p=1}^{P} \|\boldsymbol{X}_p^\top \phi_s\Big(\frac{\boldsymbol{X}_p\boldsymbol{W}_Q^{(s)}\boldsymbol{W}_K^{(s)^\top}\boldsymbol{X}_p^\top}{\sqrt{d_{QK}}}\Big)^\top \Big(\phi_r'(\boldsymbol{Z}^{(s)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(s)}) \odot \Big((F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p) \\
& \cdot \boldsymbol{W}_U^{(s)^\top}\boldsymbol{W}_2^{(s)^\top}\Big)\Big)\boldsymbol{W}_1^{(s)^\top}\|_F + \|\boldsymbol{X}_p^\top \phi_s\Big(\frac{\boldsymbol{X}_p\boldsymbol{W}_Q^{(s)}\boldsymbol{W}_K^{(s)^\top}\boldsymbol{X}_p^\top}{\sqrt{d_{QK}}}\Big)^\top (F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\boldsymbol{W}_U^{(s)^\top}\|_F \\
\leq\ & \mu \sum_{s=0}^{t} \sum_{p=1}^{P} \sqrt{M}\|\boldsymbol{X}_p\|\|\boldsymbol{W}_1^{(s)}\|\|(F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\boldsymbol{W}_U^{(s)^\top}\boldsymbol{W}_2^{(s)^\top}\|_F \\
& + \sqrt{M}\|\boldsymbol{X}_p\|\|(F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\boldsymbol{W}_U^{(s)^\top}\|_F \\
\leq\ & \mu\sqrt{M} \sum_{s=0}^{t} \sum_{p=1}^{P} \|\boldsymbol{X}_p\|\|\boldsymbol{W}_1^{(s)}\|\|\boldsymbol{W}_2^{(s)}\|\|\boldsymbol{W}_U^{(s)}\|\|F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p\|_F \\
& + \|\boldsymbol{X}_p\|\|\boldsymbol{W}_U^{(s)}\|\|F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p\|_F \\
\leq\ & \frac{27\sqrt{2}\sqrt{MP}\max_p\|\boldsymbol{X}_p\|}{8\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}(1 + \|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_2^{(0)}\|)\|\boldsymbol{W}_U^{(0)}\|\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}), \qquad (29)
\end{aligned}
$$

where the second inequality follows Lemma 2.

$$
\begin{aligned}
&\|\boldsymbol{W}_{Q,a}^{(t)} - \boldsymbol{W}_{Q}^{(0)}\|_F \\
\leq\ & \mu \sum_{s=0}^{t} \sum_{p=1}^{P} \sum_{i=1}^{M} \bigg(2\|\boldsymbol{X}_p(i,:)\|_2 \|F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p)(i,:) - \boldsymbol{Y}_p(i,:)\|_2 \|\boldsymbol{W}_U^{(s)}\|\|\boldsymbol{W}_V^{(s)}\|\|\boldsymbol{X}_p\|^2 \|\boldsymbol{W}_K^{(s)}\| \\
& +2\|\boldsymbol{X}_p(i,:)\|_2 \|F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p)(i,:) - \boldsymbol{Y}_p(i,:)\|_2 \|\boldsymbol{W}_U^{(s)}\|\|\boldsymbol{W}_2^{(s)}\|\|\boldsymbol{W}_1^{(s)}\|\|\boldsymbol{W}_V^{(s)}\|\|\boldsymbol{X}_p\|^2 \|\boldsymbol{W}_K^{(s)}\| \bigg) \\
\leq\ & \mu \sum_{s=0}^{t} \sum_{p=1}^{P} \frac{243\sqrt{M}}{16} \max_i \|\boldsymbol{X}_p(i,:)\|_2 \|\boldsymbol{X}_p\|^2 (1 + \|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_2^{(0)}\|)\|\boldsymbol{W}_U^{(0)}\|\|\boldsymbol{W}_V^{(0)}\|\|\boldsymbol{W}_K^{(0)}\|\|F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p\|_F \\
\leq\ & \frac{243\sqrt{2MP}\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2}{16\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})} (1 + \|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_2^{(0)}\|)\|\boldsymbol{W}_U^{(0)}\|\|\boldsymbol{W}_V^{(0)}\|\|\boldsymbol{W}_K^{(0)}\|\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}), \qquad (30) \\
&\|\boldsymbol{W}_{K,a}^{(t)} - \boldsymbol{W}_{K}^{(0)}\|_F \\
\leq\ & \mu \sum_{s=0}^{t} \sum_{p=1}^{P} \sum_{i=1}^{M} \bigg(2\|\boldsymbol{X}_p(i,:)\|_2 \|F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p)(i,:) - \boldsymbol{Y}_p(i,:)\|_2 \|\boldsymbol{W}_U^{(s)}\|\|\boldsymbol{W}_V^{(s)}\|\|\boldsymbol{X}_p\|^2 \|\boldsymbol{W}_Q^{(s)}\| \\
& +2\|\boldsymbol{X}_p(i,:)\|_2 \|F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p)(i,:) - \boldsymbol{Y}_p(i,:)\|_2 \|\boldsymbol{W}_U^{(s)}\|\|\boldsymbol{W}_2^{(s)}\|\|\boldsymbol{W}_1^{(s)}\|\|\boldsymbol{W}_V^{(s)}\|\|\boldsymbol{X}_p\|^2 \|\boldsymbol{W}_Q^{(s)}\| \bigg) \\
\leq\ & \mu \sum_{s=0}^{t} \sum_{p=1}^{P} \frac{243\sqrt{M}}{16} \max_i \|\boldsymbol{X}_p(i,:)\|_2 \|\boldsymbol{X}_p\|^2 (1 + \|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_2^{(0)}\|)\|\boldsymbol{W}_U^{(0)}\|\|\boldsymbol{W}_V^{(0)}\|\|\boldsymbol{W}_Q^{(0)}\|\|F_{\boldsymbol{\Theta}}^{(s)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p\|_F \\
\leq\ & \frac{243\sqrt{2MP}\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2}{16\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})} (1 + \|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_2^{(0)}\|)\|\boldsymbol{W}_U^{(0)}\|\|\boldsymbol{W}_V^{(0)}\|\|\boldsymbol{W}_Q^{(0)}\|\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}). \qquad (31)
\end{aligned}
$$

When the initialization requirement satisfies

$$
\begin{aligned}
\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}) \leq \min \Bigg\{ & \frac{2\sqrt{2}\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}{27\sqrt{P}} \frac{\sigma_{\min}(\boldsymbol{W}_2^{(0)})}{\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|\|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_U^{(0)}\|}, \\
& \frac{2\sqrt{2}\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}{27\sqrt{P}(1+\|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_2^{(0)}\|)} \cdot \min\left\{ \frac{\sigma_{\min}(\boldsymbol{W}_U^{(0)})}{\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|}, \frac{\sigma_{\min}(\boldsymbol{W}_V^{(0)})}{\sqrt{M}\max_p \|\boldsymbol{X}_p\|\|\boldsymbol{W}_U^{(0)}\|} \right\}, \\
& \frac{4\sqrt{2}\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}{243\sqrt{MP}\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2(1+\|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_2^{(0)}\|)\|\boldsymbol{W}_U^{(0)}\|\|\boldsymbol{W}_V^{(0)}\|} \\
& \cdot \min\left\{ \frac{\sigma_{\min}(\boldsymbol{W}_Q^{(0)})}{\|\boldsymbol{W}_K^{(0)}\|}, \frac{\sigma_{\min}(\boldsymbol{W}_K^{(0)})}{|\|\boldsymbol{W}_Q^{(0)}\|} \right\} \Bigg\},
\end{aligned}
$$

we can guarantee

$$
\frac{\sigma_{\min}(\boldsymbol{W}_b^{(0)})}{2} \leq \sigma_{\min}(\boldsymbol{W}_{b,a}^{(t)}) \leq \|\boldsymbol{W}_{b,a}^{(t)}\| \leq \frac{3\|\boldsymbol{W}_b^{(0)}\|}{2}, b = 2, U, V, Q, K. \qquad (32)
$$

Part II Next, we will prove $\frac{\sigma_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p))}{2} \leq \sigma_{\min}(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)) \leq \|\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\| \leq \frac{3\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|}{2}$. Specifically, we first expand

$$
\begin{aligned}
& \|\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|_F \\
= \ & \|\phi_s(\frac{\boldsymbol{X}_p\boldsymbol{W}_{Q,a}^{(t)}\boldsymbol{W}_{K,a}^{(t)}{}^\top\boldsymbol{X}_p^\top}{\sqrt{d_{QK}}})\boldsymbol{X}_p\boldsymbol{W}_{V,a}^{(t)} - \phi_s(\frac{\boldsymbol{X}_p\boldsymbol{W}_{Q}^{(0)}\boldsymbol{W}_{K}^{(0)}{}^\top\boldsymbol{X}_p^\top}{\sqrt{d_{QK}}})\boldsymbol{X}_p\boldsymbol{W}_{V}^{(0)}\|_F \\
\leq \ & \|\phi_s(\frac{\boldsymbol{X}_p\boldsymbol{W}_{Q,a}^{(t)}\boldsymbol{W}_{K,a}^{(t)}{}^\top\boldsymbol{X}_p^\top}{\sqrt{d_{QK}}}) - \phi_s(\frac{\boldsymbol{X}_p\boldsymbol{W}_{Q}^{(0)}\boldsymbol{W}_{K}^{(0)}{}^\top\boldsymbol{X}_p^\top}{\sqrt{d_{QK}}})\|_F\|\boldsymbol{X}_p\|\|\boldsymbol{W}_{V,a}^{(t)}\| \\
& +\|\phi_s(\frac{\boldsymbol{X}_p\boldsymbol{W}_{Q}^{(0)}\boldsymbol{W}_{K}^{(0)}{}^\top\boldsymbol{X}_p^\top}{\sqrt{d_{QK}}})\|_F\|\boldsymbol{X}_p\|\|\boldsymbol{W}_{V,a}^{(t)} - \boldsymbol{W}_V^{(0)}\|_F \\
\leq \ & \frac{3\sqrt{M}\|\boldsymbol{X}_p\|^3\|\boldsymbol{W}_V^{(0)}\|}{\sqrt{d_{QK}}}\|\boldsymbol{W}_{Q,a}^{(t)}\boldsymbol{W}_{K,a}^{(t)}{}^\top - \boldsymbol{W}_{Q}^{(0)}\boldsymbol{W}_{K}^{(0)}{}^\top\|_F + \sqrt{M}\|\boldsymbol{X}_p\|\|\boldsymbol{W}_{V,a}^{(t)} - \boldsymbol{W}_V^{(0)}\|_F \\
\leq \ & \sqrt{M}\|\boldsymbol{X}_p\|\|\boldsymbol{W}_{V,a}^{(t)} - \boldsymbol{W}_V^{(0)}\|_F + \frac{9\sqrt{M}\|\boldsymbol{X}_p\|^3\|\boldsymbol{W}_V^{(0)}\|\|\boldsymbol{W}_Q^{(0)}\|}{2\sqrt{d_{QK}}}\|\boldsymbol{W}_{K,a}^{(t)} - \boldsymbol{W}_K^{(0)}\|_F \\
& +\frac{9\sqrt{M}\|\boldsymbol{X}_p\|^3\|\boldsymbol{W}_V^{(0)}\|\|\boldsymbol{W}_K^{(0)}\|}{2\sqrt{d_{QK}}}\|\boldsymbol{W}_{Q,a}^{(t)} - \boldsymbol{W}_Q^{(0)}\|_F \\
\leq \ & \frac{2187\sqrt{2}M\sqrt{P}\max_p\|\boldsymbol{X}_p\|^3(\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2)\|\boldsymbol{W}_U^{(0)}\|\|\boldsymbol{W}_V^{(0)}\|^2(\|\boldsymbol{W}_Q^{(0)}\|^2 + \|\boldsymbol{W}_K^{(0)}\|^2)}{32\sqrt{d_{QK}}\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}(1 \\
& +\|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_2^{(0)}\|)\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}) + \frac{27\sqrt{2}M\sqrt{P}(\max_p\|\boldsymbol{X}_p\|)^2\|\boldsymbol{W}_U^{(0)}\|}{8\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}(1+\|\boldsymbol{W}_1^{(0)}\|\|\boldsymbol{W}_2^{(0)}\|)\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}) \\
:= \ & C_1\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}),
\end{aligned} \tag{33}
$$

where the second inequality follows Lemma 2 and the fourth inequality uses (29), (30) and (31).

Similar with the analysis of (26), when the initialization satisfies $\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}) \leq \frac{\min_p \sigma_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p))}{2C_1}$, we can derive

$$
\frac{\sigma_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p))}{2} \leq \sigma_{\min}(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)) \leq \|\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\| \leq \frac{3\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|}{2}, \tag{34}
$$

where $a \in [0,1]$.

Part III Finally, we derive a lower bound for $\sigma_{\min}(\phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)}))$. Specifically, we have

$$
\begin{aligned}
& \|\phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)}) - \phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)})\|_F \\
\leq \ & \|\phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)}) - \phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)})\|_F + \|\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)}) - \phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)})\|_F \\
\leq \ & \|\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)} - \boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)}\|_F + \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)} - \boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)}\|_F \\
\leq \ & \frac{3\|\boldsymbol{W}_1^{(0)}\|}{2}\|\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|_F + \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|\|\boldsymbol{W}_1^{(t)} - \boldsymbol{W}_1^{(0)}\|_F \\
\leq \ & \frac{3C_1\|\boldsymbol{W}_1^{(0)}\|}{2}\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}) + \frac{27\sqrt{2P}}{8\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\|\boldsymbol{W}_2^{(0)}\|\|\boldsymbol{W}_U^{(0)}\|\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}),
\end{aligned} \tag{35}
$$

where the last line uses (24) and (33). When

$$
\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}) \leq \frac{\min_p \sigma_{\min}(\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)}))}{3C_1\|\boldsymbol{W}_1^{(0)}\| + \frac{27\sqrt{2P}}{4\alpha(1-(1-\mu\alpha)^{\frac{1}{2}})}\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\|\boldsymbol{W}_2^{(0)}\|\|\boldsymbol{W}_U^{(0)}\|}
$$

is satisfied, following the analysis of (26), we have

$$
\sigma_{\min}(\phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)})) \geq \frac{\sigma_{\min}(\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)}))}{2}. \tag{36}
$$

Step II: Establishing the Lipschitz continuity of the gradient. Before proving $\Phi(\boldsymbol{\theta}^{(t+1)}) \leq (1-\mu\alpha)\Phi(\boldsymbol{\theta}^{(t)})$, it is necessary to first establish the Lipschitz continuity of the gradient, i.e., $\|\Phi(\boldsymbol{\theta}_a^{(t)}) - \Phi(\boldsymbol{\theta}^{(t)})\|_2 \leq C\|\boldsymbol{\theta}_a^{(t)} - \boldsymbol{\theta}^{(t)}\|_2$ with $\boldsymbol{\theta}_a^{(t)} = \boldsymbol{\theta}^{(t)} + a(\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^{(t)})$ and $a \in [0,1]$. To proceed, we recall the result in (33), which states that

$$\begin{aligned}
&\|\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\|_F^2 \\
\leq\ & \frac{243M\sqrt{P}\|\boldsymbol{X}_p\|^3(\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2)\|\boldsymbol{W}_V^{(0)}\|^2\|\boldsymbol{W}_Q^{(0)}\|^2}{4d_{QK}}\|\boldsymbol{W}_{K,a}^{(t)} - \boldsymbol{W}_K^{(t)}\|_F^2 \\
&+\frac{243M\sqrt{P}\|\boldsymbol{X}_p\|^3(\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2)\|\boldsymbol{W}_V^{(0)}\|^2\|\boldsymbol{W}_K^{(0)}\|^2}{4d_{QK}}\|\boldsymbol{W}_{Q,a}^{(t)} - \boldsymbol{W}_Q^{(t)}\|_F^2 \\
&+3M\sqrt{P}(\max_p\|\boldsymbol{X}_p\|)^2\|\boldsymbol{W}_{V,a}^{(t)} - \boldsymbol{W}_V^{(t)}\|_F^2,
\end{aligned}\tag{37}$$

and further derive

$$\begin{aligned}
&\|F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p) - F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)\|_F^2 \\
=\ & \|(\phi_r(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)})\boldsymbol{W}_{2,a}^{(t)} + \boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p))\boldsymbol{W}_{U,a}^{(t)} - (\phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)})\boldsymbol{W}_2^{(t)} + \boldsymbol{Z}^{(t)}(\boldsymbol{X}_p))\boldsymbol{W}_U^{(t)}\|_F^2 \\
\leq\ & 2\|\phi_r(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)})\boldsymbol{W}_{2,a}^{(t)}\boldsymbol{W}_{U,a}^{(t)} - \phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)})\boldsymbol{W}_2^{(t)}\boldsymbol{W}_U^{(t)}\|_F^2 \\
&+2\|\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{U,a}^{(t)} - \boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_U^{(t)}\|_F^2 \\
\leq\ & \frac{81\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2}{4}(\|\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)} - \boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)}\|_F^2 + \|\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)} - \boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)}\|_F^2) \\
&+4(\|\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{U,a}^{(t)} - \boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_U^{(t)}\|_F^2 + \|\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_U^{(t)} - \boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_U^{(t)}\|_F^2) \\
\leq\ & \frac{729\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2}{16}(\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\|\boldsymbol{W}_{1,a}^{(t)} - \boldsymbol{W}_1^{(t)}\|_F^2 + \|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\|_F^2) \\
&+9\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\|\boldsymbol{W}_{U,a}^{(t)} - \boldsymbol{W}_U^{(t)}\|_F^2 + 9\|\boldsymbol{W}_U^{(0)}\|^2\|\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\|_F^2 \\
\leq\ & \frac{729\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2}{16}\|\boldsymbol{W}_{1,a}^{(t)} - \boldsymbol{W}_1^{(t)}\|_F^2 + 9\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\|\boldsymbol{W}_{U,a}^{(t)} - \boldsymbol{W}_U^{(t)}\|_F^2 \\
&+\Big(\frac{729\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2}{16} + 9\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\Big) \\
&\cdot\Big(\frac{243M\sqrt{P}\|\boldsymbol{X}_p\|^3(\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2)\|\boldsymbol{W}_V^{(0)}\|^2\|\boldsymbol{W}_Q^{(0)}\|^2}{4d_{QK}}\|\boldsymbol{W}_{K,a}^{(t)} - \boldsymbol{W}_K^{(t)}\|_F^2 \\
&+\frac{243M\sqrt{P}\|\boldsymbol{X}_p\|^3(\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2)\|\boldsymbol{W}_V^{(0)}\|^2\|\boldsymbol{W}_K^{(0)}\|^2}{4d_{QK}}\|\boldsymbol{W}_{Q,a}^{(t)} - \boldsymbol{W}_Q^{(t)}\|_F^2 \\
&+3M\sqrt{P}(\max_p\|\boldsymbol{X}_p\|)^2\|\boldsymbol{W}_{V,a}^{(t)} - \boldsymbol{W}_V^{(t)}\|_F^2\Big) \\
\leq\ & C_F\|\boldsymbol{\theta}_a^{(t)} - \boldsymbol{\theta}^{(t)}\|_2^2,
\end{aligned}\tag{38}$$

where we define $C_F = \max\big\{\frac{729\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2}{16}, 9\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2,$ $\big(\frac{729\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2}{16} + 9\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\big)\frac{243M\sqrt{P}\max_p\|\boldsymbol{X}_p\|^3(\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2)\|\boldsymbol{W}_V^{(0)}\|^2\|\boldsymbol{W}_Q^{(0)}\|^2}{4d_{QK}},$ $\big(\frac{729\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2}{16} + 9\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\big)\frac{243M\sqrt{P}\max_p\|\boldsymbol{X}_p\|^3(\max_{i,p}\|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2)\|\boldsymbol{W}_V^{(0)}\|^2\|\boldsymbol{W}_K^{(0)}\|^2}{4d_{QK}},$ $27M\sqrt{P}(\max_p\|\boldsymbol{X}_p\|)^2\big(\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2 + \frac{81\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2}{16}\big)\big\}$.

In addition, since it has been previously established that $\|\boldsymbol{W}_{b,a}^{(t)} - \boldsymbol{W}_b^{(t)}\|_F^2 \leq \frac{9\sigma_{\min}^2(\boldsymbol{W}_b^{(t)})}{4}, b = 1, 2, U, V, Q, K$, we can further obtain

$$
\begin{aligned}
&\|F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p) - F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)\|_F^2 \\
\leq\ & \frac{6561\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2 \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\sigma_{\min}^2(\boldsymbol{W}_1^{(0)})}{64} + \frac{81 \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\sigma_{\min}^2(\boldsymbol{W}_U^{(0)})}{4} \\
&+\left(9\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2 + \frac{729\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2}{16}\right)\left(\frac{27M\sqrt{P}(\max_p \|\boldsymbol{X}_p\|)^2\sigma_{\min}^2(\boldsymbol{W}_V^{(0)})}{4}\right. \\
&+\frac{2187M\sqrt{P}\max_p \|\boldsymbol{X}_p\|^3(\max_{i,p} \|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2)\|\boldsymbol{W}_V^{(0)}\|^2\|\boldsymbol{W}_Q^{(0)}\|^2\sigma_{\min}^2(\boldsymbol{W}_K^{(0)})}{16d_{QK}} \\
&\left.+\frac{2187M\sqrt{P}\max_p \|\boldsymbol{X}_p\|^3(\max_{i,p} \|\boldsymbol{X}_p(i,:)\|_2\|\boldsymbol{X}_p\|^2)\|\boldsymbol{W}_V^{(0)}\|^2\|\boldsymbol{W}_K^{(0)}\|^2\sigma_{\min}^2(\boldsymbol{W}_Q^{(0)})}{16d_{QK}}\right) \\
:=\ & C_2. \qquad (39)
\end{aligned}
$$

Considering that

$$
\begin{aligned}
\|\nabla_{\boldsymbol{\theta}}\Phi(\boldsymbol{\theta}_a^{(t)}) - \nabla_{\boldsymbol{\theta}}\Phi(\boldsymbol{\theta}^{(t)})\|_2^2 = \sum_{b=1,2,U,V,Q,K} &\|\nabla_{\boldsymbol{W}_b} L(\boldsymbol{W}_{1,a}^{(t)}, \boldsymbol{W}_{2,a}^{(t)}, \boldsymbol{W}_{Q,a}^{(t)}, \boldsymbol{W}_{K,a}^{(t)}, \boldsymbol{W}_{V,a}^{(t)}, \boldsymbol{W}_{U,a}^{(t)}) \\
&-\nabla_{\boldsymbol{W}_b} L(\boldsymbol{W}_1^{(t)}, \boldsymbol{W}_2^{(t)}, \boldsymbol{W}_Q^{(t)}, \boldsymbol{W}_K^{(t)}, \boldsymbol{W}_V^{(t)}, \boldsymbol{W}_U^{(t)})\|_F^2, \qquad (40)
\end{aligned}
$$

we proceed to analyze each $\boldsymbol{W}_b$ term individually for $b = 1, 2, U, V, Q, K$. To simplify notation, we define $\nabla_{\boldsymbol{W}_b} L(\boldsymbol{W}_{1,a}^{(t)}, \boldsymbol{W}_{2,a}^{(t)}, \boldsymbol{W}_{Q,a}^{(t)}, \boldsymbol{W}_{K,a}^{(t)}, \boldsymbol{W}_{V,a}^{(t)}, \boldsymbol{W}_{U,a}^{(t)}) = \nabla_{\boldsymbol{W}_b} L(\{\boldsymbol{W}_{b,a}^{(t)}\}_b)$ and $\nabla_{\boldsymbol{W}_b} L(\boldsymbol{W}_1^{(t)}, \boldsymbol{W}_2^{(t)}, \boldsymbol{W}_Q^{(t)}, \boldsymbol{W}_K^{(t)}, \boldsymbol{W}_V^{(t)}, \boldsymbol{W}_U^{(t)}) = \nabla_{\boldsymbol{W}_b} L(\{\boldsymbol{W}_b^{(t)}\}_b)$ for $b = 1, 2, U, V, Q, K$.

Now, we begin by expanding the expression corresponding to $\boldsymbol{W}_1$.

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{W}_1} L(\{\boldsymbol{W}_{b,a}^{(t)}\}_b) - \nabla_{\boldsymbol{W}_1} L(\{\boldsymbol{W}_b^{(t)}\}_b)\|_F^2 \\
=\ & \|\sum_{p=1}^{P} \boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)^\top(\phi_r'(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)}) \odot ((F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\boldsymbol{W}_{U,a}^{(t)\top}\boldsymbol{W}_{2,a}^{(t)\top})) \\
&-\sum_{p=1}^{P} \boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)^\top(\phi_r'(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)}) \odot ((F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\boldsymbol{W}_U^{(t)\top}\boldsymbol{W}_2^{(t)\top}))\|_F^2 \\
\leq\ & 3\|\sum_{p=1}^{P} \boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)^\top(\phi_r'(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)}) \odot ((F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p) - F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p))\boldsymbol{W}_{U,a}^{(t)\top}\boldsymbol{W}_{2,a}^{(t)\top}))\|_F^2 \\
&+3\|\sum_{p=1}^{P} \boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)^\top(\phi_r'(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)}) \odot ((F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\boldsymbol{W}_{U,a}^{(t)\top}\boldsymbol{W}_{2,a}^{(t)\top}))\|_F^2 \\
&+3\|\sum_{p=1}^{P} \boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)^\top(\phi_r'(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)}) \odot ((F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\boldsymbol{W}_U^{(t)\top}\boldsymbol{W}_2^{(t)\top}))\|_F^2 \\
\leq\ & \frac{2187P}{64}\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\|\boldsymbol{W}_U^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\sum_{p=1}^{P}\|F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p) - F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)\|_F^2 \\
&+\frac{2187P}{16}\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\|\boldsymbol{W}_U^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\Phi(\boldsymbol{\theta}^{(0)}) \\
\leq\ & \frac{2187PC_F}{32}\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\|\boldsymbol{W}_U^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{\theta}_a^{(t)} - \boldsymbol{\theta}^{(t)}\|_2^2, \qquad (41)
\end{aligned}
$$

where the last line follows (38) and $\Phi(\boldsymbol{\theta}^{(0)}) \leq 4\sum_{p=1}^{P}\|F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p) - F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)\|_F^2 \leq 4PC_2$.

Similarly, for $\boldsymbol{W}_2$, we have

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{W}_2} L(\{\boldsymbol{W}_{b,a}^{(t)}\}_b) - \nabla_{\boldsymbol{W}_2} L(\{\boldsymbol{W}_b^{(t)}\}_b)\|_F^2 \\
=& \|\sum_{p=1}^{P} (\phi_r(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)}))^\top (F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p){\boldsymbol{W}_{U,a}^{(t)}}^\top \\
&- \sum_{p=1}^{P} (\phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1}^{(t)}))^\top (F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p){\boldsymbol{W}_{U}^{(t)}}^\top \|_F^2 \\
\leq& 3\|\sum_{p=1}^{P} (\phi_r(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)}))^\top (F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p) - F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)){\boldsymbol{W}_{U,a}^{(t)}}^\top \|_F^2 \\
&+3\|\sum_{p=1}^{P} (\phi_r(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)}))^\top (F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p){\boldsymbol{W}_{U,a}^{(t)}}^\top \|_F^2 \\
&+3\|\sum_{p=1}^{P} (\phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1}^{(t)}))^\top (F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p){\boldsymbol{W}_{U}^{(t)}}^\top \|_F^2 \\
\leq& 3P\sum_{p=1}^{P} \|(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)})^\top (F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p) - F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)){\boldsymbol{W}_{U,a}^{(t)}}^\top \|_F^2 \\
&+3P\sum_{p=1}^{P} \|(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)})^\top (F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p){\boldsymbol{W}_{U,a}^{(t)}}^\top \|_F^2 \\
&+3P\sum_{p=1}^{P} \|(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1}^{(t)})^\top (F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p){\boldsymbol{W}_{U}^{(t)}}^\top \|_F^2 \\
\leq& \frac{2187PC_F}{32} \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2 \|\boldsymbol{W}_U^{(0)}\|^2 \|\boldsymbol{W}_1^{(0)}\|^2 \|\boldsymbol{\theta}_a^{(t)} - \boldsymbol{\theta}^{(t)}\|_2^2, \qquad (42)
\end{aligned}
$$

where the second inequality uses Assumption 1, and the last line follows (38) and $\Phi(\boldsymbol{\theta}^{(0)}) \leq 4\sum_{p=1}^{P} \|F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p) - F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)\|_F^2 \leq 4PC_2$.

In addition, for $\boldsymbol{W}_U$, we have

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{W}_U} L(\{\boldsymbol{W}_{b,a}^{(t)}\}_b) - \nabla_{\boldsymbol{W}_U} L(\{\boldsymbol{W}_b^{(t)}\}_b)\|_F^2 \\
=& \|\sum_{p=1}^{P} (\phi_r(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)})\boldsymbol{W}_{2,a}^{(t)} + \boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p))^\top (F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p) \\
&- \sum_{p=1}^{P} (\phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1}^{(t)})\boldsymbol{W}_{2}^{(t)} + \boldsymbol{Z}^{(t)}(\boldsymbol{X}_p))^\top (F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\|_F^2 \\
\leq& 2\|\sum_{p=1}^{P} (\phi_r(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)})\boldsymbol{W}_{2,a}^{(t)})^\top (F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p) \\
&- \sum_{p=1}^{P} (\phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1}^{(t)})\boldsymbol{W}_{2}^{(t)})^\top (F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\|_F^2 \\
&+2\|\sum_{p=1}^{P} {\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)}^\top (F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p) - \sum_{p=1}^{P} {\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)}^\top (F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p)\|_F^2
\end{aligned}
$$

$$
\begin{aligned}
\leq\ & 6\|\sum_{p=1}^{P}(\phi_r(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)})\boldsymbol{W}_{2,a}^{(t)})^\top(F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p)-F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p))\|_F^2\\
&+6\|\sum_{p=1}^{P}(\phi_r(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)})\boldsymbol{W}_{2,a}^{(t)})^\top(F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)-\boldsymbol{Y}_p)\|_F^2\\
&+6\|\sum_{p=1}^{P}(\phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)})\boldsymbol{W}_2^{(t)})^\top(F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)-\boldsymbol{Y}_p)\|_F^2+6\|\sum_{p=1}^{P}\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)^\top(F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p)-F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p))\|_F^2\\
&+6\|\sum_{p=1}^{P}\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)^\top(F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)-\boldsymbol{Y}_p)\|_F^2+6\|\sum_{p=1}^{P}\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)^\top(F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)-\boldsymbol{Y}_p)\|_F^2\\
\leq\ & 6P\sum_{p=1}^{P}\|(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)}\boldsymbol{W}_{2,a}^{(t)})^\top(F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p)-F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p))\|_F^2\\
&+6P\sum_{p=1}^{P}\|(\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_{1,a}^{(t)}\boldsymbol{W}_{2,a}^{(t)})^\top(F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)-\boldsymbol{Y}_p)\|_F^2\\
&+6P\sum_{p=1}^{P}\|(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)}\boldsymbol{W}_2^{(t)})^\top(F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)-\boldsymbol{Y}_p)\|_F^2+6P\sum_{p=1}^{P}\|\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)^\top(F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p)-F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p))\|_F^2\\
&+6P\sum_{p=1}^{P}\|\boldsymbol{Z}_a^{(t)}(\boldsymbol{X}_p)^\top(F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)-\boldsymbol{Y}_p)\|_F^2+6P\sum_{p=1}^{P}\|\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)^\top(F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)-\boldsymbol{Y}_p)\|_F^2\\
\leq\ & \frac{2187}{32}P\sum_{p=1}^{P}\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\|F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p)-F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)\|_F^2\\
&+\frac{2187}{16}P\sum_{p=1}^{P}\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\|F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)-\boldsymbol{Y}_p\|_F^2\\
&+\frac{27}{2}P\sum_{p=1}^{P}\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\|F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p)-F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)\|_F^2+27P\sum_{p=1}^{P}\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\|F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)-\boldsymbol{Y}_p\|_F^2\\
\leq\ & \frac{2187PC_F}{16}\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\|\boldsymbol{\theta}_a^{(t)}-\boldsymbol{\theta}^{(t)}\|_2^2+27PC_F\max_p\|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2\|\boldsymbol{\theta}_a^{(t)}-\boldsymbol{\theta}^{(t)}\|_2^2,
\end{aligned}
\tag{43}
$$

where the last line follows (38) and $\Phi(\boldsymbol{\theta}^{(0)})\leq 4\sum_{p=1}^{P}\|F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p)-F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)\|_F^2\leq 4PC_2$.

Similarly, for $\boldsymbol{W}_V$, $\boldsymbol{W}_Q$, and $\boldsymbol{W}_K$, by invoking (38) and $\Phi(\boldsymbol{\theta}^{(0)})\leq 4\sum_{p=1}^{P}\|F_{\boldsymbol{\Theta},a}^{(t)}(\boldsymbol{X}_p)-F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p)\|_F^2\leq 4PC_2$, we have

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{W}_V}L(\{\boldsymbol{W}_{b,a}^{(t)}\}_b)-\nabla_{\boldsymbol{W}_V}L(\{\boldsymbol{W}_b^{(t)}\}_b)\|_F^2\\
\leq\ & \frac{177147PC_F}{256d_{QK}}(\max_p\|\boldsymbol{X}_p\|)^6\|\boldsymbol{W}_1^{(0)}\|^2\|\boldsymbol{W}_2^{(0)}\|^2\|\boldsymbol{W}_U^{(0)}\|^2\|\boldsymbol{W}_Q^{(0)}\|^2\|\boldsymbol{W}_K^{(0)}\|^2\|\boldsymbol{\theta}_a^{(t)}-\boldsymbol{\theta}^{(t)}\|_2^2\\
&+\frac{2187PC_F}{16d_{QK}}(\max_p\|\boldsymbol{X}_p\|)^6\|\boldsymbol{W}_U^{(0)}\|^2\|\boldsymbol{W}_Q^{(0)}\|^2\|\boldsymbol{W}_K^{(0)}\|^2\|\boldsymbol{\theta}_a^{(t)}-\boldsymbol{\theta}^{(t)}\|_2^2,
\end{aligned}
\tag{44}
$$

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{W}_Q} L(\{\boldsymbol{W}_{b,a}^{(t)}\}_b) - \nabla_{\boldsymbol{W}_Q} L(\{\boldsymbol{W}_b^{(t)}\}_b)\|_F^2 \\
=&\frac{2187 P C_F M}{4} \max_{i,p} \|\boldsymbol{X}_p(i,:)\|_2^2 (\max_p \|\boldsymbol{X}_p\|)^4 \|\boldsymbol{W}_U^{(0)}\|^2 \|\boldsymbol{W}_V^{(0)}\|^2 \|\boldsymbol{W}_K^{(0)}\|^2 \|\boldsymbol{\theta}_a^{(t)} - \boldsymbol{\theta}^{(t)}\|_2^2 \\
&+\frac{177147 P C_F M}{64} \max_{i,p} \|\boldsymbol{X}_p(i,:)\|_2^2 (\max_p \|\boldsymbol{X}_p\|)^4 \|\boldsymbol{W}_1^{(0)}\|^2 \|\boldsymbol{W}_2^{(0)}\|^2 \|\boldsymbol{W}_U^{(0)}\|^2 \\
&\cdot \|\boldsymbol{W}_V^{(0)}\|^2 \|\boldsymbol{W}_K^{(0)}\|^2 \|\boldsymbol{\theta}_a^{(t)} - \boldsymbol{\theta}^{(t)}\|_2^2, \qquad (45)\\
&\|\nabla_{\boldsymbol{W}_K} L(\{\boldsymbol{W}_{b,a}^{(t)}\}_b) - \nabla_{\boldsymbol{W}_K} L(\{\boldsymbol{W}_b^{(t)}\}_b)\|_F^2 \\
=&\frac{2187 P C_F M}{4} \max_{i,p} \|\boldsymbol{X}_p(i,:)\|_2^2 (\max_p \|\boldsymbol{X}_p\|)^4 \|\boldsymbol{W}_U^{(0)}\|^2 \|\boldsymbol{W}_V^{(0)}\|^2 \|\boldsymbol{W}_Q^{(0)}\|^2 \|\boldsymbol{\theta}_a^{(t)} - \boldsymbol{\theta}^{(t)}\|_2^2 \\
&+\frac{177147 P C_F M}{64} \max_{i,p} \|\boldsymbol{X}_p(i,:)\|_2^2 (\max_p \|\boldsymbol{X}_p\|)^4 \|\boldsymbol{W}_1^{(0)}\|^2 \|\boldsymbol{W}_2^{(0)}\|^2 \|\boldsymbol{W}_U^{(0)}\|^2 \\
&\cdot \|\boldsymbol{W}_V^{(0)}\|^2 \|\boldsymbol{W}_Q^{(0)}\|^2 \|\boldsymbol{\theta}_a^{(t)} - \boldsymbol{\theta}^{(t)}\|_2^2. \qquad (46)
\end{aligned}
$$

Combing (41)-(46) with (40), we have $\|\Phi(\boldsymbol{\theta}_a^{(t)}) - \Phi(\boldsymbol{\theta}^{(t)})\|_2 \leq C\|\boldsymbol{\theta}_a^{(t)} - \boldsymbol{\theta}^{(t)}\|_2$, where $C^2 = \frac{2187 P C_F}{32} \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2 \cdot \|\boldsymbol{W}_U^{(0)}\|^2 (\|\boldsymbol{W}_1^{(0)}\|^2 + \|\boldsymbol{W}_2^{(0)}\|^2) + P C_F \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2 \cdot \big(\frac{2187}{16} \|\boldsymbol{W}_1^{(0)}\|^2 \|\boldsymbol{W}_2^{(0)}\|^2 + 27\big) + \frac{P C_F}{d_{QK}} (\max_p \|\boldsymbol{X}_p\|)^6 \|\boldsymbol{W}_U^{(0)}\|^2 \|\boldsymbol{W}_Q^{(0)}\|^2 \|\boldsymbol{W}_K^{(0)}\|^2 \big(\frac{2187}{16} + \frac{177147}{256} \|\boldsymbol{W}_1^{(0)}\|^2 \|\boldsymbol{W}_2^{(0)}\|^2\big) + \frac{2187 P C_F M}{4} \max_{i,p} \|\boldsymbol{X}_p(i,:)\|_2^2 (\max_p \|\boldsymbol{X}_p\|)^4 \|\boldsymbol{W}_U^{(0)}\|^2 \|\boldsymbol{W}_V^{(0)}\|^2 \cdot (\|\boldsymbol{W}_K^{(0)}\|^2 + \|\boldsymbol{W}_Q^{(0)}\|^2) + \frac{177147 P C_F M}{64} \max_{i,p} \|\boldsymbol{X}_p(i,:)\|_2^2 (\max_p \|\boldsymbol{X}_p\|)^4 \|\boldsymbol{W}_1^{(0)}\|^2 \|\boldsymbol{W}_2^{(0)}\|^2 \cdot \|\boldsymbol{W}_U^{(0)}\|^2 \|\boldsymbol{W}_V^{(0)}\|^2 (\|\boldsymbol{W}_K^{(0)}\|^2 + \|\boldsymbol{W}_Q^{(0)}\|^2)$.

Step III: Proof of last inequality in (23). Using Lemma 1 with the Lipschitz continuity of the gradient, we can analyze the convergence property of $\boldsymbol{\theta}^{(t+1)} = \boldsymbol{\theta}^{(t)} - \mu \nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}^{(t)})$ as follows:

$$
\begin{aligned}
\Phi(\boldsymbol{\theta}^{(t+1)}) &\leq \Phi(\boldsymbol{\theta}^{(t)}) + \langle \nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}^{(t)}), \Phi(\boldsymbol{\theta}^{(t+1)}) - \Phi(\boldsymbol{\theta}^{(t)}) \rangle + \frac{C}{2} \|\Phi(\boldsymbol{\theta}^{(t+1)}) - \Phi(\boldsymbol{\theta}^{(t)})\|_2^2 \\
&= \Phi(\boldsymbol{\theta}^{(t)}) - \mu \|\nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}^{(t)})\|_2^2 + \frac{C\mu^2}{2} \|\nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}^{(t)})\|_2^2 \\
&\leq \Phi(\boldsymbol{\theta}^{(t)}) - \frac{\mu}{2} \|\nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta}^{(t)})\|_2^2 \\
&\leq \Phi(\boldsymbol{\theta}^{(t)}) - \frac{\mu}{2} \|\nabla_{\boldsymbol{W}_2} L(\boldsymbol{W}_1^{(t)}, \boldsymbol{W}_2^{(t)}, \boldsymbol{W}_Q^{(t)}, \boldsymbol{W}_K^{(t)}, \boldsymbol{W}_V^{(t)}, \boldsymbol{W}_U^{(t)})\|_F^2 \\
&\leq \Phi(\boldsymbol{\theta}^{(t)}) - \mu \frac{\sigma_{\min}^2(\boldsymbol{W}_U^{(0)}) \min_p \sigma_{\min}^2(\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)}))}{16} \Phi(\boldsymbol{\theta}^{(t)}) \\
&= (1 - \mu\alpha) \Phi(\boldsymbol{\theta}^{(t)}), \qquad (47)
\end{aligned}
$$

where the second and third inequalities respectively follows $\mu \leq \frac{1}{C}$ and $\|\nabla_{\boldsymbol{\theta}} \Phi(\boldsymbol{\theta})\|_2^2 = \|\nabla_{\boldsymbol{W}_1} L\|_F^2 + \|\nabla_{\boldsymbol{W}_2} L\|_F^2 + \|\nabla_{\boldsymbol{W}_U} L\|_F^2 + \|\nabla_{\boldsymbol{W}_V} L\|_F^2 + \|\nabla_{\boldsymbol{W}_Q} L\|_F^2 + \|\nabla_{\boldsymbol{W}_K} L\|_F^2$. The fourth inequality uses $\|\nabla_{\boldsymbol{W}_2} L(\boldsymbol{W}_1^{(t)}, \boldsymbol{W}_2^{(t)}, \boldsymbol{W}_Q^{(t)}, \boldsymbol{W}_K^{(t)}, \boldsymbol{W}_V^{(t)}, \boldsymbol{W}_U^{(t)})\|_F = \|\sum_{p=1}^P (\phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)}))^\top (F_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}_p) - \boldsymbol{Y}_p) {\boldsymbol{W}_U^{(t)}}^\top\|_F = \|\mathrm{Diag}\{(\phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)}))^\top\}_{p=1}^P \cdot (\overline{F}_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}) - \overline{\boldsymbol{Y}}) {\boldsymbol{W}_U^{(t)}}^\top\|_F \geq \frac{1}{4} \sigma_{\min}^2(\boldsymbol{W}_U^{(0)}) \min_p \sigma_{\min}^2(\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)})) \|\overline{F}_{\boldsymbol{\Theta}}^{(t)}(\boldsymbol{X}) - \overline{\boldsymbol{Y}}\|_F$ where $\mathrm{Diag}\{(\phi_r(\boldsymbol{Z}^{(t)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(t)}))^\top\}_{p=1}^P$ is a diagonal block matrix. Here we define $\alpha = \frac{\sigma_{\min}^2(\boldsymbol{W}_U^{(0)}) \sigma_{\min}^2(\phi_r(\boldsymbol{Z}^{(0)}\boldsymbol{W}_1^{(0)}))}{16}$.

■

C Proof of Theorem 1

Proof To simplify Theorem 2 in Appendix B, we define $\overline{\lambda} = \max_{b=1,2,U,V,Q,K} \|\boldsymbol{W}_b^{(0)}\|$ and $\underline{\lambda} = \min_{b=1,2,U,V,Q,K} \sigma_{\min}(\boldsymbol{W}_b^{(0)})$. Considering that when P is sufficiently large, C_2 in (21) becomes significantly larger than the other terms, we can neglect this term. Consequently, the initialization requirement and the constant C can be respectively simplified as

$$\begin{aligned}
&\Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}) \lesssim \min\Bigg\{ \frac{\underline{\lambda}^2 \min_p \sigma_{\min}(\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)}))}{\max\{1, \max_p \|\boldsymbol{X}_p\|\} \cdot \max\{1, \overline{\lambda}^2\} \cdot \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|}, \\
&\frac{\underline{\lambda}^2 \min_p \sigma_{\min}(\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)}))}{\max_p \|\boldsymbol{X}_p\|^2 \max_{i.p} \|\boldsymbol{X}_p(i,:)\|_2 \max\{\overline{\lambda}^3, \overline{\lambda}^5\}}, \\
&\frac{\underline{\lambda} \min_p \sigma_{\min}(\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)})) \cdot \min\big\{ \min_p \sigma_{\min}(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)), \min_p \sigma_{\min}^2(\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)}))\big\}}{\max\big\{ \max_p \|\boldsymbol{X}_p\|^5 \max_{i,p} \|\boldsymbol{X}_p(i,:)\|_2 \cdot \max\{\overline{\lambda}^5, \overline{\lambda}^7\}, \max_p \|\boldsymbol{X}_p\|^2 \overline{\lambda} \max\{1, \overline{\lambda}^2\}, \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2 \overline{\lambda}^2 \big\}} \Bigg\} \\
&= \overline{C} \frac{\underline{\lambda}^2 (\min_p \sigma_{\min}(\phi_r(\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\boldsymbol{W}_1^{(0)})))^2}{\max\{1, \overline{\lambda}^7\} \cdot \max_p \|\boldsymbol{X}_p\|^5 \max_{i,p} \|\boldsymbol{X}_p(i,:)\|_2 \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|^2}
\end{aligned}$$

and

$$\begin{aligned}
C = \widetilde{C} \cdot \max\{&\overline{\lambda}^2 \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|, \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|, \max_p \|\boldsymbol{X}_p\|(\overline{\lambda}^3 + \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|), \\
&(\overline{\lambda}^3 + \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|) \max_p \|\boldsymbol{X}_p\|^3 \overline{\lambda}^2\}(\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\|(1 + \overline{\lambda}^2) \\
&+ \max_p \|\boldsymbol{X}_p\|^3(\overline{\lambda}^3 + \overline{\lambda}^5) + \max_p \|\boldsymbol{X}_p\|^2 \max_{i,p} \|\boldsymbol{X}_p(i,:)\|_2(\overline{\lambda}^3 + \overline{\lambda}^5)),
\end{aligned} \tag{48}$$

where $\overline{C}$ and $\widetilde{C}$ are positive constants.

This completes the proof. ■

D Proof of Corollary 1

In this section, we will prove $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{\infty}$ is a Cauchy sequence. Let us fix any $\epsilon > 0$. We need to show that there exists $z > 0$ such that for every $i, j \geq z$, $\|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(i)}\|_2 < \epsilon$. Without loss of generality, we assume that $i < j$. Then, we have

$$\begin{aligned}
&\|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(i)}\|_2 \\
&= \sqrt{\sum_{b=1,2,U,V,Q,K} \|\boldsymbol{W}_b^{(j)} - \boldsymbol{W}_b^{(i)}\|_F^2} \\
&\leq \sum_{b=1,2,U,V,Q,K} \|\boldsymbol{W}_b^{(j)} - \boldsymbol{W}_b^{(i)}\|_F \\
&\leq \sum_{b=1,2,U,V,Q,K} \sum_{s=i}^{j-1} \|\boldsymbol{W}_b^{(s+1)} - \boldsymbol{W}_b^{(s)}\|_F \\
&\leq \sum_{b=1,2,U,V,Q,K} \sum_{s=i}^{j-1} \mu \|\nabla_{\boldsymbol{W}_b} L(\boldsymbol{W}_1^{(s)}, \boldsymbol{W}_2^{(s)}, \boldsymbol{W}_Q^{(s)}, \boldsymbol{W}_K^{(s)}, \boldsymbol{W}_V^{(s)}, \boldsymbol{W}_U^{(s)})\|_F \\
&\leq (1 - \mu\alpha)^{\frac{i}{2}} \frac{1 - (1 - \mu\alpha)^{\frac{j-i}{2}}}{1 - (1 - \mu\alpha)^{\frac{1}{2}}} \mu C_W \Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}) \\
&\leq (1 - \mu\alpha)^{\frac{i}{2}} \frac{C_W}{\alpha} \Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}),
\end{aligned} \tag{49}$$

where we define $C_W = \frac{27\sqrt{2P}}{8} \max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\| \|\boldsymbol{W}_U^{(0)}\|(\|\boldsymbol{W}_1^{(0)}\| + \|\boldsymbol{W}_2^{(0)}\|) + \frac{27\sqrt{2P}}{8}(1 + \|\boldsymbol{W}_1^{(0)}\| \|\boldsymbol{W}_2^{(0)}\|)(\max_p \|\boldsymbol{Z}^{(0)}(\boldsymbol{X}_p)\| + \sqrt{M} \max_p \|\boldsymbol{X}_p\| \|\boldsymbol{W}_U^{(0)}\|) + \frac{243\sqrt{2MP}}{16} \max_{i,p} \|\boldsymbol{X}(i,:)\|_2 \|\boldsymbol{X}_p\|^2 (1 + \|\boldsymbol{W}_1^{(0)}\| \|\boldsymbol{W}_2^{(0)}\|) \|\boldsymbol{W}_U^{(0)}\| \|\boldsymbol{W}_V^{(0)}\| (\|\boldsymbol{W}_K^{(0)}\| + \|\boldsymbol{W}_Q^{(0)}\|)$ and fourth inequality follows (24), (27)-(31). In addition, the last line uses $\mu \frac{1-(1-\mu\alpha)^{\frac{j-i}{2}}}{1-(1-\mu\alpha)^{\frac{1}{2}}} = \frac{(1-(1-\mu\alpha))}{\alpha} \frac{1-(1-\mu\alpha)^{\frac{j-i}{2}}}{1-(1-\mu\alpha)^{\frac{1}{2}}} \leq \frac{1}{\alpha}$.

Note that $(1-\mu\alpha)^{\frac{i}{2}} \leq (1-\mu\alpha)^{\frac{z}{2}}$ and thus there exists a sufficiently large z such that $\|\boldsymbol{\theta}^{(j)} - \boldsymbol{\theta}^{(i)}\|_2 < \epsilon$. This shows that $\{\boldsymbol{\theta}^{(t)}\}_{t=1}^{\infty}$ is a Cauchy sequence, and hence convergent to some $\boldsymbol{\theta}^{\star}$. By continuity, $\Phi(\boldsymbol{\theta}^{\star}) = \Phi(\lim_{t\to\infty} \boldsymbol{\theta}^{(t)}) = \lim_{t\to\infty} \Phi(\boldsymbol{\theta}^{(t)}) = 0$, and thus $\boldsymbol{\theta}^{\star}$ is a global minimizer. The rate of convergence is

$$\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{\star}\|_2 = \lim_{t\to\infty} \|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^{(t)}\|_2 \leq (1-\mu\alpha)^{\frac{k}{2}} \frac{C_W}{\alpha} \Phi^{\frac{1}{2}}(\boldsymbol{\theta}^{(0)}). \tag{50}$$

E Technical tools used in the proofs

Lemma 1 *([NM20, Lemma 4.3]) Let $f : \mathbb{R}^n \to \mathbb{R}$ be a twice continuously differentiable function. Let $x, y \in \mathbb{R}^n$ be given, and assume that $\|\nabla f(z) - \nabla f(x)\|_2 \leq C\|z - x\|_2$ for every $z = x + t(y - x)$ with $t \in [0, 1]$. Then,*

$$f(y) \leq f(x) + \langle \nabla f(x), y - x \rangle + \frac{C}{2}\|x - y\|^2. \tag{51}$$

Lemma 2 *For any matrix $\boldsymbol{A} \in \mathbb{R}^{d_1 \times d_2}$, we apply the row-wise softmax function $\phi_s(\cdot)$ to obtain $\phi_s(\boldsymbol{A}) \in \mathbb{R}^{d_1 \times d_2}$, where each row is given by*

$$\phi_s(\boldsymbol{A}(i,:)) = \left[\frac{e^{\boldsymbol{A}(i,1)}}{\sum_{j=1}^{d_2} e^{\boldsymbol{A}(i,j)}} \quad \cdots \quad \frac{e^{\boldsymbol{A}(i,d_2)}}{\sum_{j=1}^{d_2} e^{\boldsymbol{A}(i,j)}} \right]. \tag{52}$$

First, we have

$$1 \leq \|\phi_s(\boldsymbol{A})\|_F^2 \leq d_1 \tag{53}$$

and

$$\|\phi_s(\boldsymbol{A}_1) - \phi_s(\boldsymbol{A}_2)\|_F^2 \leq 4d_1\|\boldsymbol{A}_1 - \boldsymbol{A}_2\|_F^2, \tag{54}$$

where $\boldsymbol{A}_1, \boldsymbol{A}_2 \in \mathbb{R}^{d_1 \times d_2}$.

Second, the Jacobian $\phi_s'(\boldsymbol{A}(i,:)), i = 1, \ldots, d_1$ of the softmax can be written as

$$\phi_s'(\boldsymbol{A}(i,:)) = \mathit{diag}(\phi_s(\boldsymbol{A}(i,:))) - \phi_s^{\top}(\boldsymbol{A}(i,:))\phi_s(\boldsymbol{A}(i,:)) \in \mathbb{R}^{d_2 \times d_2}, \tag{55}$$

this further implies $\|\phi_s'(\boldsymbol{A}(i,:))\|_F \leq 2$.

Proof According to [WLCC23, Lemma 8], we have $\frac{1}{\sqrt{d_1}} \leq \|\phi_s(\boldsymbol{A}(i,:))\|_2 \leq 1$ and further obtain

$$1 \leq \|\phi_s(\boldsymbol{A})\|_F^2 \leq d_1. \tag{56}$$

In addition, we can derive

$$\begin{aligned} \|\phi_s(\boldsymbol{A}_1) - \phi_s(\boldsymbol{A}_2)\|_F^2 &\leq \|\phi_s(\boldsymbol{A}_1) - \phi_s(\boldsymbol{A}_2)\|_1^2 \\ &\leq 4(\sum_{i=1}^{d_1} \|\boldsymbol{A}_1(i,:) - \boldsymbol{A}_2(i,:)\|_\infty)^2 \\ &\leq 2d_1 \sum_{i=1}^{d_1} \|\boldsymbol{A}_1(i,:) - \boldsymbol{A}_2(i,:)\|_\infty^2 \\ &\leq 4d_1\|\boldsymbol{A}_1 - \boldsymbol{A}_2\|_F^2, \end{aligned} \tag{57}$$

where the second inequality follows [EGKZ22, Corollary A.7].

In the end, the Jacobian of the softmax has been derived from [WLCC23, Lemma 11]. ■

Lemma 3 *([Ver10, Corollary 5.35]) For any matrix $\boldsymbol{W} \in \mathbb{R}^{d_1 \times d_2}$ where $d_1 > d_2$ and each element is sampled independently from $\mathcal{N}(0, 1)$, for every $\zeta \geq 0$, with probability at least $1 - 2\exp(-\zeta^2/2)$ one has:*

$$\sqrt{d_1} - \sqrt{d_2} - \zeta \leq \phi_{\mathit{min}}(\boldsymbol{W}) \leq \|\boldsymbol{W}\| \leq \sqrt{d_1} + \sqrt{d_2} + \zeta. \tag{58}$$

As a direct consequence, we have

Lemma 4 *For any matrix $\boldsymbol{W} \in \mathbb{R}^{d_1 \times d_2}$ where $d_1/4 > d_2$ and each element is sampled independently from $\mathcal{N}(0, \gamma^2)$, with probability at least $1 - 2\exp(-d_1/8)$, one has:*

$$\gamma(\frac{\sqrt{d_1}}{2} - \sqrt{d_2}) \leq \phi_{min}(\boldsymbol{W}) \leq \|\boldsymbol{W}\| \leq \gamma(\frac{3\sqrt{d_1}}{2} + \sqrt{d_2}). \tag{59}$$

Theorem 3 *(Hoeffding's inequality) Let $Z_1, \ldots, Z_n$ be independent bounded random variables with $Z_i \in [a, b]$ for all i, where $-\infty < a \leq b < \infty$. Then*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}(Z_i - \mathbb{E}[Z_i])\right| \geq t\right) \leq 2\exp\left(-\frac{2nt^2}{(b-a)^2}\right) \tag{60}$$

for all $t \geq 0$.