On the Convergence of Encoder-only Shallow Transformers

Yongtao Wu
LIONS, EPFL
yongtao.wu@epfl.ch

Fanghui Liu
University of Warwick
fanghui.liu@warwick.ac.uk

Grigorios G Chrysos
LIONS, EPFL
University of Wisconsin-Madison
chrysos@wisc.edu

Volkan Cevher
LIONS, EPFL
volkan.cevher@epfl.ch

Abstract

In this paper, we aim to build the global convergence theory of encoder-only shallow Transformers under a realistic setting from the perspective of architectures, initialization, and scaling under a finite width regime. The difficulty lies in how to tackle the softmax in self-attention mechanism, the core ingredient of Transformer. In particular, we diagnose the scaling scheme, carefully tackle the input/output of softmax, and prove that quadratic overparameterization is sufficient for global convergence of our shallow Transformers under commonly-used He/LeCun initialization in practice. Besides, neural tangent kernel (NTK) based analysis is also given, which facilitates a comprehensive comparison. Our theory demonstrates the separation on the importance of different scaling schemes and initialization. We believe our results can pave the way for a better understanding of modern Transformers, particularly on training dynamics.

1 Introduction

Transformers [Vaswani et al., 2017] have demonstrated unparalleled success in influential applications [Devlin et al., 2019; Brown et al., 2020; Wang et al., 2018; Dosovitskiy et al., 2021; Liu et al., 2022b]. A fundamental theoretical topic concerns the global convergence, i.e., the training dynamics of Transformers, which would be helpful for further analysis, e.g., in-context learning [von Oswald et al., 2022; Akyürek et al., 2023], generalization [Li et al., 2023]. In fact, even within a simplified Transformer framework under certain specific regimes, the global convergence guarantees still remain an elusive challenge.

To theoretically understand this, let us first recall the exact format of the self-attention mechanism, the core ingredient of the Transformer. Given the input $X \in \mathbb{R}^{d_s \times d}$ ($d_s$ is the number of tokens and $d$ is the feature dimension of each token), a self-attention mechanism is defined as:

$$\text{Self-attention}(X) \triangleq \sigma_s \left( \tau_0 (XW_Q^\top) (XW_K^\top)^\top \right) (XW_V^\top) = \sigma_s \left( \tau_0 XW_Q^\top W_K X^\top \right) (XW_V^\top),$$

where $\sigma_s$ is the row-wise softmax function, $\tau_0 \in \mathbb{R}^+$ is the scaling factor, and the learnable weights are $W_Q, W_K, W_V \in \mathbb{R}^{d_m \times d}$ with the width $d_m$. Given $X$, the input of softmax depends on $\tau_0 W_Q^\top W_K$, including the scaling factor $\tau_0$ and initialization schemes for learnable parameters, and thus determines the output of softmax and then affects the performance of Transformers in both

*Work done at LIONS, EPFL.*

theory and practice. There are several scaling schemes in previous literature. For instance, given \( W_Q \) and \( W_K \) initialized by standard Gaussian, the scaling factor \( \tau_0 \) is chosen by

- \( \tau_0 = d_m^{-1/2} \) in the original Transformer [Vaswani et al. 2017]: each element in \( \tau_0 W_Q^T W_K \) is a random variable with mean 0 and variance 1. This scaling avoids the blow-up of value inside softmax as \( d_m \) increases [Hron et al. 2020].

- \( \tau_0 = d_m^{-1} \): This scaling stems from the neural tangent kernel (NTK) analysis [Jacot et al. 2018], a commonly used technical tool for convergence analysis of fully-connected (or convolutional) networks under an infinite-width setting \( d_m \to \infty \). However, for Transformer, if one uses this scaling under the infinite-width setting, then by the law of large numbers, we have \( \lim_{d_m \to \infty} \tau_0 W_Q^T W_K \approx 0 \). As a result, the input of softmax is zero and the softmax degenerates to a pooling layer. That means, the non-linearity is missing, which motivates researchers to carefully rethink this setting.

For instance, under the \( \tau_0 = d_m^{-1} \) setting, Yang [2020] use the same query and key matrices to prevent the softmax from degenerating into a pooling layer. Besides, to avoid the analytic difficulty of softmax due to the fact that each element of the output depends on all inputs, [Hron et al. 2020] substitute softmax with ReLU under the \( \tau_0 = d_m^{-1/2} \) and infinite width setting for simplicity.

Clearly, there exists a gap between theoretical analysis and practical architectures on the use of softmax, and accordingly, this leads to the following open question:

**How can we ensure the global convergence of Transformers under a realistic setting?**

The primary contribution of this work is to establish the convergence theory of shallow Transformer under a realistic setting. Despite its shallow and encoder-only architecture, our Transformer model captures all the fundamental components found on typical Transformers, including the self-attention mechanism with the softmax activation function, one feedforward ReLU layer, one average pooling layer, and a linear output layer, cf. Eq. (1.2). We adopt the \( \tau_0 = d_m^{-1/2} \) scaling under the finite-width setting and compare the results of LeCun/He initializations, which are commonly used in practical applications. Besides, the convergence result under the \( \tau_0 = d_m^{-1} \) setting (as well as the NTK based analysis) is also studied, which facilitates a comprehensive comparison. Our theoretical results demonstrate notable separations among scaling settings, initializations, and architectures as below:

- **Scaling:** The global convergence can be achieved under both \( \tau_0 = d_m^{-1/2} \) and \( \tau_0 = d_m^{-1} \). Nevertheless, as suggested by our theory: for a small \( d_m \), there is no significant difference for these two scaling schemes on the convergence; but for a large enough \( d_m \), the \( \tau_0 = d_m^{-1/2} \) scaling admits a faster convergence rate of Transformers than that with \( \tau_0 = d_m^{-1} \). Interestingly, under this \( \tau_0 = d_m^{-1} \) setting, our theory also demonstrates the separation on the convergence result, depending on whether the input is formed along sequence dimension (\( d = 1 \)) or embedding dimension (\( d_e = 1 \)).

- **Initialization:** Under LeCun and He initialization, our shallow Transformer admits a faster convergence rate than the NTK initialization. This could be an explanation for the seldom usage of NTK initialization for Transformer training in practice.

- **Architecture:** Quadratic over-parameterization is enough to ensure the global convergence of our shallow Transformer. As a comparison, if the self-attention mechanism is substituted by a feed-forward ReLU layer, our shallow Transformer is close to a three-layer fully-connected ReLU neural networks to some extent, requiring cubic over-parameterization for global convergence.

We firmly believe that our theoretical analysis takes a significant step towards unraveling the mysteries behind Transformers from the perspective of global convergence. We hope that our analytical framework and insights on various initialization and scaling techniques would be helpful in training modern, large-scale Transformer-based models [Radford et al. 2018, Brown et al. 2020].

### 2 Related work

**Self-attention, Transformer:** Regarding training dynamics, Snell et al. [2021] explain why single-head attention focuses on salient words by analyzing the evolution throughout training. Hron et al. [2020] show that the output of Transformer converges to Gaussian process kernel and provide the
NTK formulation of Transformer. Recently, Li et al. [2023] provide sample complexity of shallow Transformer to study its generalization property under a good initialization from pretrained model. The separation between the Transformer and CNN is recently explored. Jelassi et al. [2022] provably demonstrate that Vision Transformer (ViT) has the ability to learn spatial structure without additional inductive bias such as the spatial locality in CNN. Chen et al. [2022] study the loss landscape of ViT and find that ViT converges at sharper local minima than ResNet. Park and Kim [2022] show that ViT is a low-pass filter while CNN is a high-pass filter, thus, these two models can be complementary.

NTK, lazy training, Hessian: The NTK was introduced by Jacot et al. [2018] to connect the infinite-width neural network trained by gradient descent and the kernel regression. The roles of such kernel include analysis of the training dynamics of the neural network in the over-parameterization regime Allen-Zhu et al. [2019a], Chizat et al. [2019], Du et al. [2019a,b], Zou et al. [2020]. The global convergence, generalization bound, and memorization capacity largely depend on the minimum eigenvalue of the NTK Cao and Gu [2019], Zhu et al. [2020], Nguyen et al. [2021], Bombari et al. [2022]. Even though the NTK is extended from FCNN to several typical networks including Transformer Tirer et al. [2020], Huang et al. [2020], Arora et al. [2019b], Alemohammad et al. [2021], Nguyen and Mondelli [2020], it has not been used to analyze the global convergence of Transformer. On the other hand, the stability of the tangent kernel during training is required when connecting to kernel regression, but such stability cannot be explained by the phenomenon of lazy training Chizat et al. [2019], which indicates a small change of the parameters from initialization. The hessian spectral bound is the main reason for the stability of kernel, as mentioned in Liu et al. [2020].

Over-parameterization for convergence analysis: Due to over-parameterization, neural networks (NNs) can fit arbitrary labels with zero training loss when trained with (stochastic) gradient descent (SGD), both theoretically Li and Liang [2018], Du et al. [2019b] and empirically Zhang et al. [2018]. This leads to an interesting question in theory: how much overparameterization is enough to ensure global convergence of NNs? A common recipe for the proof of global convergence relies on the variant of Polyak-Lojasiewicz condition Polyak [1963], Liu et al. [2022a], NTK Du et al. [2019b], Zou and Gu [2019], Allen-Zhu et al. [2019a], or the minimum eigenvalue of the gram matrix Nguyen [2021], Bombari et al. [2022]. In Appendix B.3, we provide a comprehensive overview of a recent line of work that improves the over-parameterization condition for ensuring the convergence of NNs. However, the over-parameterization condition for Transformer to achieve global convergence remains elusive from existing literature and we make an initial step towards this question.

3 Problem setting

This section includes the problem setting with notations and model formulation of the shallow Transformer that is studied in this paper.

3.1 Notation

Vectors (matrices) are symbolized by lowercase (upercase) boldface letters, e.g., \( \mathbf{w}, \mathbf{W} \). We use \( \| \cdot \|_F \) and \( \| \cdot \|_2 \) to represent the Frobenius norm and the spectral norm of a matrix, respectively. The Euclidean norm of a vector is symbolized by \( \| \cdot \|_2 \). The superscript with brackets is used to represent the element of a vector/matrix, e.g., \( w^{(i)} \) is the \( i \)th element of \( w \). The superscript without brackets symbolizes the parameters at different training step, e.g., \( \theta^t \). We denote by \( [N] = \{1, \ldots, N\} \) for short. We use \( \sigma_{\min}(-) \) and \( \lambda_{\min}(-) \) to represent the minimum singular value and minimum eigenvalue of a matrix. The NTK matrix and hessian matrix of the network are denoted by \( \mathbf{K} \) and \( \mathbf{H} \), respectively. The order notation, e.g., \( \mathcal{O}, \Omega \), omits the logarithmic factor. More detailed notation can be found in Table 2 of the appendix.

Let \( X \subseteq \mathbb{R}^{d_x \times d} \) be a compact metric space and \( Y \subseteq \mathbb{R} \), where \( d \) is the dimension of each token, \( d_x \) is the total sequence length of the input. The training set \( \{(X_n, y_n)\}_{n=1}^{N} \) is assumed to be iid sampled from an unknown probability measure on \( X \times Y \). In this paper, we focus the regression task by employing the squared loss. The goal of our regression task is to find a hypothesis, i.e., a Transformer \( f : X \to Y \) in our work, such that \( f(X; \theta) \) parameterized by \( \theta \) is a good approximation of the label \( y \in Y \) corresponding to a new sample \( X \in X \). We use a vector \( \theta \) to denote the collection of all learnable parameters.

3
We consider standard gradient descent (GD) training of Transformer, as illustrated in Algorithm 1, with width $d_f$ where the output is $y$. Throughout this work, we consider the encoder of Transformer, which can be applied to both regression and classification tasks [Yuksel et al., 2019, Dosovitskiy et al., 2021]. Given an input $X \in \mathbb{R}^{d_{\text{in}} \times d}$, the model is defined as below:

$$A_1 = \text{Self-attention}(X) \triangleq \sigma_s \left( \tau_0 (X W_Q \top) (X W_K \top) \right) (X W_V \top),$$

$$A_2 = \tau_1 \sigma_r (A_1, W_H), \quad A_3 = \varphi(A_2), \quad f(X; \theta) = a_3 \top w_O,$$

where the output is $f(X; \theta) \in \mathbb{R}$, $\tau_0$ and $\tau_1$ are two scaling factors. The ingredients of a Transformer with width $d_m$ are defined as follows:

- A self-attention mechanism (Eq. (1.1)): $\sigma_s$ is the row-wise softmax function; the learnable parameters are $W_Q, W_K, W_V \in \mathbb{R}^{d_m \times d}$. We employ Gaussian initialization $W_Q^{(ij)} \sim \mathcal{N}(0, \eta_Q)$, $W_K^{(ij)} \sim \mathcal{N}(0, \eta_K)$, $W_V^{(ij)} \sim \mathcal{N}(0, \eta_V)$ with $i \in [d_m]$ and $j \in [d]$. Refer to Table 1 for typical initialization examples.

- A feed-forward ReLU layer (in Eq. (1.2)): $\sigma_r$ is the ReLU activation function; the learnable parameter is $W_H \in \mathbb{R}^{d_m \times d_m}$. Following [Yang et al., 2022], we combine $W_V$ and $W_H$ together (by setting $W_H = I$) for ease of the analysis. Note that it does not mean its training dynamics are the same as the joint-training of these two adjacent matrices.

- An average pooling layer (in Eq. (1.2)): $\varphi$ indicates the column-wise average pooling. Note that the average pooling layer is applied along the sequence length dimension to ensure the final output is a scalar, which is commonly used in practical Vision Transformer or theoretical analysis [Dosovitskiy et al., 2021, Yang, 2020].

- An output layer (in Eq. (1.2)) with learnable parameter $w_O \in \mathbb{R}^{d_m}$, initialized by $w_O^{(ij)} \sim \mathcal{N}(0, \eta_O)$.

Remarks: Proper initialization and scaling are required to ensure the convergence and learnability, as seen in previous work [Jacot et al., 2018, Tirer et al., 2022, Lee et al., 2019]. For our convergence analysis, we consider standard Gaussian initialization with different variances and different scaling factor that includes three typical initialization schemes in practice. In Table 1 we detail the formula of LeCun initialization, He initialization, and NTK initialization.

Given $N$ input samples $\{X_n\}_{n=1}^N$, the corresponding ground truth label, the final output of network, and the output of the last hidden layer, are denoted by:

$$y \triangleq \{y_n\}_{n=1}^N \in \mathbb{R}^N, \quad f(\theta) \triangleq \{f(X_n; \theta)\}_{n=1}^N \in \mathbb{R}^N, \quad F_{\text{pre}}(\theta) \triangleq \{a_3(X_n; \theta)\}_{n=1}^N \in \mathbb{R}^{N \times d_m}.$$  

We consider standard gradient descent (GD) training of Transformer, as illustrated in Algorithm 1. Here the squared loss is expressed as $\ell(\theta) = \frac{1}{2} \| f(\theta) - y \|^2$.

### 4 Main results

In this section, we study the convergence guarantee of Transformer training by GD under the squared loss. Firstly, we provide a general analytical framework in Section 4.1 covering different initialization.
schemes, where we identify the condition for achieving global convergence. Next, we validate these conditions for several practical initialization schemes under \( r_0 = d_m^{-1/2} \) in Section 4.2 and \( r_0 = d_m^{-1} \) in Section 4.3, respectively. We include NTK-based results for self-completeness. Discussion on the convergence results with different scalings, initializations, and architectures is present in Section 4.4.

### 4.1 General framework for convergence analysis

Before presenting our result, we introduce a basic assumption.

**Assumption 1.** The input data is bounded \( \|X\|_F \leq \sqrt{d_x C_x} \) with some positive constant \( C_x \).

**Remark:** This assumption is standard as we can always scale the input. The embedding of token is usually assumed to have a unit norm [Li et al., 2023], which is unrelated to \( d \).

Now we are ready to present our proposition, with the proof deferred to Appendix C.2. Notice that our proposition holds with high probability under some conditions, which depend on certain initialization schemes and scaling factors. Since our proposition is devoted to a unifying analysis framework under various initialization schemes, we do not include specific probabilities here. The validation of the required conditions and probability is deferred to Sections 4.2 and 4.3, respectively.

**Proposition 1.** Consider the Transformer with \( d_m \geq N \). Let \( C_Q, C_K, C_V, C_O \) be some positive constants, and define the following quantities at initialization for simplification:

- The norm of the parameters:
  \[
  \lambda_Q \triangleq \|W_Q^0\|_2 + C_Q, \quad \lambda_K \triangleq \|W_K^0\|_2 + C_K, \quad \lambda_V \triangleq \|W_V^0\|_2 + C_V, \quad \lambda_O \triangleq \|w_0\|_2 + C_O.
  \]

- Two auxiliary terms: \( \rho \triangleq N^{1/2}d_x^{1/2}\tau_1 C_z \), \( z \triangleq \lambda_Q^2 \left( 1 + 4\tau_0^2 C_4^2 d_x^2 \lambda_Q^2 \left( \lambda_Q^2 + \lambda_K^2 \right) \right) \).

Under Assumption 1 we additionally assume that the minimum singular value of \( F_\text{pre}^0 \), i.e., \( \alpha = \sigma_{\text{min}}(F_\text{pre}^0) \) satisfies the following condition at initialization:

\[
\alpha^2 \geq 8\rho \max \left( \lambda_V C_O^{-1}, \lambda_O C_V^{-1}, 2\tau_0 C_2 d_x \lambda_K \lambda_V \lambda_O C_Q^{-1}, 2\tau_0 C_2 d_x \lambda_Q \lambda_V \lambda_O C_K^{-1} \right) \sqrt{2\ell(\theta^0)}, \quad (2)
\]

\[
\alpha^3 \geq 32\rho^2 z \sqrt{2\ell(\theta^0)}/\lambda_O. \quad (3)
\]

If the step size satisfies \( \gamma \leq 1/C \) with a constant \( C \) depending on \( (\lambda_Q, \lambda_K, \lambda_V, \lambda_O, \ell(\theta^0), \rho, \tau_0) \), then GD converges to a global minimum as follows:

\[
\ell(\theta^t) \leq \left( 1 - \frac{t \alpha^2}{2} \right)^t \ell(\theta^0), \quad \forall t \geq 0. \quad (4)
\]

**Remark:** The parameter \( \alpha \) in Eqs. (2) and (3) controls the convergence rate of global convergence, and the condition will be verified in the next subsection. The step-size \( \gamma \) is inversely proportional to \( N^{1/2} \) due to the construction of \( C \) in Appendix C.2.

**Proof sketch:** The main idea of our convergence analysis is based on the variant of Polyak-Lojasiewicz (PL) inequality [Polyak, 1963; Nguyen, 2021], i.e., \( \|\nabla \ell(\theta)\|_2^2 \geq 2\min\{K\} \ell(\theta) \geq 2\min \{ F_\text{pre}^0 F_\text{pre}^T \} \ell(\theta) \). Thus, if the minimum singular value of \( F_\text{pre}^0 \) is strictly greater than 0, then minimizing the gradient on the LHS will drive the loss to zero. To this end, Proposition 1 can be split into two parts. First, by induction, at every time step, each parameter in the Transformer can be bounded w.h.p; the minimum singular value of \( F_\text{pre} \) is bounded away for some positive quality at the initialization point. Secondly, we prove that the Lipschitzness of the network gradient, which means the loss function is almost smooth. Combining the above two results, the global convergence can be achieved. Furthermore, based on different initialization and scaling schemes, we are able to validate Eqs. (2) and (3) via the spectral norm estimation of \( \lambda_Q, \lambda_K, \lambda_V, \lambda_O \) and a positive lower bound for \( F_\text{pre}^0 \) in the following section.

### 4.2 LeCun and He initialization under the \( d_m^{-1/2} \) setting

Here we aim to show that, under the LeCun/He initialization with the setting of \( d_m^{-1/2} \), the conditions in Eqs. (2) and (3) will be satisfied with high probability and scaling schemes in Table 1 and hence...
lead to global convergence. To derive our result, we need the following assumptions on the input data regarding the rank and dissimilarity of data.

**Assumption 2.** We assume that the input data $X$ has full row rank.

**Remark:** This assumption requires that each row $X^{(i,:)}$ is linearly independent for any $i \in [d_x]$, which is fair and attainable in practice. For example, for language tasks, even though there might be some repeated token in $X$, each row in $X$ can be uncorrelated when added with positional embedding. Similarly, in image tasks with Visual Transformer, the raw image is grouped by patch and mapped via a linear layer to construct $X$, thus each row in $X$ can be uncorrelated.

**Assumption 3.** For any data pair $(X_n, X_{n'})$, with $n \neq n'$ and $n, n' \in [N]$, then we assume that $\mathbb{P} \left( \| \langle X_n^\top, X_{n'} \rangle \| \geq t \right) \leq \exp(-t^\hat{c})$ with some constant $\hat{c} > 0$.

**Remark:** We discuss the rationale behind this assumption:

The idea behind Assumption 3 is that different data points admit a small similarity. To be specific, for two data points $X_n$ and $X_{n'}$ with $n \neq n'$, their inner product reflects the similarity of their respective (empirical) covariance matrix. We expect that this similarity is small with a high probability. The spirit of this assumption also exists in previous literature, e.g., Nguyen et al. [2021]. The constant $\hat{c}$ determines the decay of data dissimilarity. A larger $\hat{c}$ results in less separable data. Our assumption has no requirement on $\hat{c}$ such that $\hat{c}$ can be small enough, which allows for a general data distribution.

**Verification of assumption.** Here we experimentally validate this assumption under a standard language IMDB dataset [Maas et al., 2011]. We randomly sample 100 (normalized) sentences with embedding and plot the probability (frequency) of $\mathbb{P} \left( \| \langle X_n^\top, X_{n'} \rangle \| \geq t \right)$ as $t$ increases. We repeat it over 10 runs and plot the result in Figure 1. We can observe an exponential decay as $t$ increases, which implies that our assumption is fair. Besides, the validation of Assumption 3 on image data, e.g., MNIST by ViT, is deferred to Appendix C.3.

Now we are ready to present our main theorem under the $d_m^{-1/2}$ setting. The proof is deferred to Appendix C.3.

**Theorem 1.** Under the setting of LeCun/He initialization in Table 1 and Assumptions 1 to 3, suppose $d_m \geq d$ and given $\tau_0 = d_m^{-1/2}$, $d_m = \Omega(N^3)$, then with probability at least $1 - 8 e^{-d_m/2} - \delta - \exp(-\Omega((N-1)^{-\hat{c}} d_m^{-1}))$ for proper $\delta$, the GD training of Transformer converges to a global minimum with sufficiently small step size $\gamma$ as in Eq. 4.

**Remark:** The probability relates to several randomness sources, e.g., data sampling, dissimilarity of data, and parameter initialization. The quantity $\exp(-\Omega((N-1)^{-\hat{c}} d_m^{-1}))$ can be small for a small enough $\hat{c}$ as discussed in Assumption 3. Further discussion on our result refers to Section 4.4 for details. Besides, our proof framework is also valid for the $\tau_0 = d_m^{-1}$ setting. We demonstrate that $d_m = \Omega(N^2)$ suffices to achieve global convergence, see Appendix C.3 for details.

**Proof sketch:** To check whether the conditions in Eqs. 2 and 3 hold, the key idea is to provide the lower bound of $\alpha$. Then we upper bound $\lambda_Q, \lambda_K, \lambda_V, \lambda_O$ based on concentration inequalities to upper bound the initial loss, one key step is to utilize Gershgorin circle theorem [Gershgorin [1931]] to provide a lower bound for $\alpha$. Lastly, we plug these bound into the condition Eqs. 2 and 3 in order to obtain the requirement for the width $d_m$.

### 4.3 NTK initialization under the $d_m^{-1}$ setting

The NTK theory, as a representative application of the $\tau_0 = d_m^{-1}$ setting, can be also used for analysis of training dynamics. We also include the NTK results in this section for self-completeness. In this section, we first derive the limiting NTK formulation of Transformer under the $d_m^{-1}$ scaling scheme, and then show the global convergence of Transformers.
We compare the derived results under different scaling schemes, initializations, and architectures.

Theorem 2. Under the setting of NTK initialization in Table 1 and Assumptions 1 to 3, suppose \( \tau \) under the over-parameterization regime, the scaling \( \tau \) setting. Nevertheless, we do not claim that \( \tau = d_m^{-1} \) is better than \( \tau = d_m^{-1/2} \). This is because, under the over-parameterization regime, the scaling \( \tau = d_m^{-1} \) makes the self-attention layer close to a pooling layer. This analysis loses the ability to capture the key characteristics of the self-attention mechanism in Transformers.

Remark: The formulation of \( \Phi^* \) implies that the self-attention layer degenerates as \( \frac{1}{\alpha} \mathbf{1}_{d_s \times d_s} X W V^\top \), i.e., the dimension missing effect as mentioned before.

Now we are ready to present our convergence result under the \( d_m^{-1} \) scaling with the proof deferred to Appendix C.4.

Theorem 3. Under the setting of NTK initialization in Table 1 and Assumptions 1 to 3, suppose \( d_m = \Omega(N) \), then with probability at least \( 1 - 8e^{-d_m/2} - \delta - \exp\left(-\Omega((N-1)^{-c}d_m^{-1})\right) \), the GD training of Transformer converges to a global minimum with sufficiently small \( \gamma \) as in Eq. (4).

Remark: By [Liu et al., 2020, Proposition 2.3], the small Hessian norm is a sufficient condition for small change of NTK. Thus, Theorem 3 can be an indicator for the stability of NTK. Besides, Theorem 3 supplements the result in [Park and Kim, 2022] which exhibits empirically a relationship between the Hessian norm and the width but lacks theoretical proof.

4.4 Discussion on convergence results

We compare the derived results under different scaling schemes, initializations, and architectures.

Comparison of scaling schemes: The global convergence can be achieved under both \( \tau_0 = d_m^{-1/2} \) and \( \tau_0 = d_m^{-1} \). Nevertheless, as suggested by our theory, for a small \( d_m \), there is no significant difference between these two scaling schemes on the convergence; but for a large enough \( d_m \), the \( \tau_0 = d_m^{-1/2} \) scaling admits a faster convergence rate of Transformers than that of \( \tau_0 = d_m^{-1} \) due to a more tight estimation of the lower bound of \( \alpha \), see Appendix C.6 for details. The intuition behind the lower convergence rate under the setting of large width and \( \tau_0 = d_m^{-1} \) is that the input of softmax is close to zero such that softmax roughly degenerates as a pooling layer, losing the ability to fit data. This can be also explained by Lemma 1 from the perspective of dimension missing: self-attention \((X)\) degenerates as \( \frac{1}{\alpha} \mathbf{1}_{d_s \times d_s} X W V^\top \).

The result under the \( \tau_0 = d_m^{-1} \) setting requires weaker over-parameterization than the \( \tau_0 = d_m^{-1/2} \) setting. Nevertheless, we do not claim that \( \tau_0 = d_m^{-1} \) is better than \( \tau_0 = d_m^{-1/2} \). This is because, under the over-parameterization regime, the scaling \( \tau_0 = d_m^{-1} \) makes the self-attention layer close to a pooling layer. This analysis loses the ability to capture the key characteristics of the self-attention mechanism in Transformers.

Note that the lower bound of the minimum eigenvalue is in the constant order, which is tight (since it matches the upper bound). Based on this, by studying the relationship between \( d_m \) and \( \lambda_0 \), we can prove that quadratic (cubic) over-parameterization is required for \( d_m^{-1} (d_m^{-1/2}) \) scaling. This quadratic over-parameterization requirement could be relaxed if a better relationship is given while it is still unclear and beyond our proof technique.
**Comparison on initializations:** Though our results achieve the same convergence rate under these initialization schemes, we can still show the *separation* on $\alpha$ that affects the convergence in Eq. (4) among these initialization schemes. To be specific, we verify that under LeCun and He initialization, the lower bound of $\alpha^2$ is tighter than that of NTK initialization, and hence admits faster convergence, see Appendix C.5 for further details. This can be an explanation of the seldom usage of NTK initialization in practice. Besides, the NTK initialization scheme allows for a larger step size than LeCun/He initialization for training. That means, if $\alpha$ is the same in these three initialization schemes, we usually choose a large step size under the NTK initialization scheme, see Appendix E.1.

**Comparison on architectures:** Note that the Transformer defined in Eq. (1.2) includes a self-attention layer, a feed-forward ReLU layer, and an output layer. Our result proves that a cubic (quadratic) over-parameterization condition is required for $d_m^{-1/2}$ ($d_m^{-1}$) under LeCun initialization. As a comparison, a three-layer fully-connected ReLU network under LeCun initialization requires $d_m = \Omega(N^3)$ [Nguyen, 2021].

The aforementioned result holds for matrix input. Although not as frequent, some data inputs are naturally in vector form. Two ways to feed the input into Transformer are either formulating along sequence dimension ($d = 1$) or along embedding dimension ($d_s = 1$). The following result shows the separation under these two settings to understand when the Transformer works well or not regarding the input.

**Corollary 1 (Convergence of vector input).** Consider LeCun initialization with $\tau_0 = d_m^{-1}$ scaling, given vector input $x \in \mathbb{R}^d$, if one feeds the input to Transformer by setting $d_s = 1, d = \tilde{d}$, then training with GD can converge to a global minimum. On the contrary, if one sets $d_s = \tilde{d}, d = 1$, the conditions in Eqs. (2) and (3) do not hold, the convergence cannot be guaranteed by our theory.

**Remark:** Such a result is motivated by considering least squares. Specifically, given input $X \in \mathbb{R}^{N \times 1}$, then the NTK $XX^\top$ is a rank-one matrix. As a result, when the augmented matrix $[X, y]$ is not rank-one, then there is no solution to the linear system so that GD training cannot converge to zero loss. The empirical validation can be found in Figure 5.

**Technical difficulty.** The technical difficulty of our analysis includes handling the softmax function and scaling schemes beyond NTK initialization. Previous convergence analysis, e.g., [Du et al., 2019b; Nguyen, 2021], cannot be applied to our setting because of the following issues. First, different from classical activation functions, e.g., ReLU, in softmax each element of the output depends on all input. To tackle the interplay between dimensions, we build the connection between the output of softmax and $XX^\top$ to disentangle the nonlinear softmax function. By doing so, a lower bound on the minimum singular value of $F_{\text{pre}}$ in Proposition 1 can be well controlled by $XX^\top$ and the output of softmax.

Regarding different initializations and scaling, previous NTK-based analysis is only valid under the $d_m^{-1}$ setting (the softmax degenerates to an all-one vector) but is inapplicable to the realistic $d_m^{-1/2}$ setting, as discussed in the introduction. To tackle this issue, we analyze the input/output of softmax under LeCun/He initialization and identify the optimization properties of the loss function for global convergence under the finite-width setting.

## 5 Experimental validations

Our experiments are organized as follows: In Section 5.1, we conduct experiments with the model Eq. (1.2) on synthetic data and study the training dynamics. Next, we show convergence results on ViT [Dosovitskiy et al., 2021] on the standard MNIST dataset in Section 5.2. Additional results and detail on the experimental setup are deferred to Appendix E.

### 5.1 Fitting synthetic data

In this section, we corroborate our theoretical findings on the synthetic data. We generate 100 data points where the input $X \in \mathbb{R}^{10 \times 100}$ is sampled from standard Gaussian distribution. The corresponding label $y \in \tilde{F}$ is generated from standard Gaussian distribution. Squared loss is selected as the criterion. We apply gradient descent on the shallow Transformer defined in Eq. (1.2) with LeCun initialization and $\tau_0 = d_m^{-1/2}$ for 400 epochs with a fixed step size $\gamma = 1$. We test different widths of the network including $d_m = \{10, 100, 1000, 4000\}$ and plot the training progress in
Figure 2: Visualization of the training process of Transformers with LeCun initialization and $\tau_0 = d_m^{-1/2}$ scaling on synthetic data. (a) Linear convergence. (b) Rate of change of the weights during training. Observe that the weights change very slowly after the 50th epoch. (c) Evolution of the NTK during the training. The result mirrors the plot (b) and demonstrates how the kernel varies significantly at the beginning of the training and remains approximately constant later. As the width increases, the empirical NTK becomes more stable.

Figure 2. The result shows that the training can converge except for the case with sub-linear width, see the linear convergence rate in Figure 2a and the small movement of the weight in Figure 2b. For these cases, the weights do not change much after 50 epochs while the losses are still decreasing. In Figure 2c, we keep track of the evolution of NTK along each epoch. Specifically, the kernel distance is given by:

$$\text{Distance} \left( K, \tilde{K} \right) = 1 - \frac{\text{Tr} \left( K \tilde{K}^\top \right)}{\sqrt{\text{Tr} \left( K K^\top \right)} \sqrt{\text{Tr} \left( \tilde{K} \tilde{K}^\top \right)}} ,$$

which quantitatively compares two kernels by the relative rotation, as used in Fort et al. [2020]. Figure 2c shows that the kernel changes rapidly at the beginning of training while being approximately constant at later stages. The experiment with $\tau_0 = d_m^{-1}$, which is deferred to Appendix E.1, obtains similar results. Additionally, the experiment with other initialization schemes is deferred to Appendix E.1, where we observe that NTK initialization indeed yields slower convergence than LeCun/He initialization, which is consistent with our theoretical finding.

Next, in order to verify corollary 1, we feed vector input $x_n \in \mathbb{R}^{100}$ into Transformer by setting either $d_s = 1$, $d = 100$ or $d_s = 100$, $d = 1$ under LeCun initialization with $\tau_0 = d_m^{-1}$. Figure 3 shows that the training of Transformer with $d_s = 1$, $d = 100$ is similar to that of two-layer FCNN with the same width $d_m = 100$. However, the case of $d_s = 100$, $d = 1$ fails to converge, which is consistent with our theoretical finding.

Figure 3: Convergence result on synthetic data with vector input.

5.2 Fitting real-world dataset

Beyond synthetic data, in this section, we examine the convergence performance of Vision Transformer (ViT) on classification task on MNIST dataset [LeCun et al., 1998], which includes 10 classes of images with size $28 \times 28$. We use a single layer and single head ViT. The dimension of $d$ is 64. We change the dimension of the query, key, and value from 16 to 1024 and 16384. The network is optimized with SGD with step size 0.1, and momentum 0.9 for 50 epochs. We repeat the experiment for three runs. We present the convergence results over 3 runs in Figure 4a, which shows that when the width is smaller, e.g., 16, both $\tau_0 = d_m^{-1}$ and $\tau_0 = d_m^{-1/2}$ scaling have similar
convergence results. However, as the width increases to 1024 and 16384, the $\tau_0 = d_m^{-1}$ setting admits a slower rate than that of $\tau_0 = d_m^{-1/2}$, especially a extremely slow rate under $d_m = 16384$. This is consistent with our theoretical result on the dimension missing effect such that the $\tau_0 = d_m^{-1}$ setting makes Transformer difficult to fit data. Additionally, we visualize the attention map in Figure 4b, i.e., the output of softmax in the self-attention layer before training and after training under $\tau_0 = d_m^{-1/2}$ setting. We could see that the self-attention layer changes a lot during training.

6 Conclusion

We present a comprehensive analysis of the global convergence properties of shallow Transformer under various scaling and initialization schemes. Regarding scaling schemes, for a large width setting, the difference on convergence between $\tau_0 = d_m^{-1/2}$ and $\tau_0 = d_m^{-1}$ can be demonstrated both theoretically and empirically. Our theory is able to explain this in a dimension missing view. Regarding initialization schemes, our theory prefers to using LeCun and He initialization in Transformer training, which allows for a faster convergence rate than NTK initialization. Though our analysis is limited to shallow Transformers, we believe our framework can be extended to deep Transformers. We provide further discussion into this extension in Appendix C.7. Exploring the convergence properties of deep Transformers is indeed an intriguing avenue for future research.

Acknowledgements

We thank the reviewers for their constructive feedback. We thank Zhenyu Zhu for the discussion and help in this work. This work has received funding from the Swiss National Science Foundation (SNSF) under grant number 200021_205011. This work was supported by Hasler Foundation Program: Hasler Responsible AI (project number 21043). Corresponding author: Fanghui Liu.
References


Ze Liu, Han Hu, Yutong Lin, Zhuliang Yao, Zhenda Xie, Yixuan Wei, Jia Ning, Yue Cao, Zheng Zhang, Li Dong, et al. Swin transformer v2: Scaling up capacity and resolution. In Conference on Computer Vision and Pattern Recognition (CVPR), 2022b.


Contents of the Appendix

The Appendix is organized as follows:

• In Appendix A we summarize the symbols and notation used in this work.

• In Appendix B.1 we provide a theoretical overview for neural tangent kernel (NTK). The background in Sub-Exponential random variables is elaborated in Appendix B.2. More detailed related work on the convergence analysis of neural networks can be found in Appendix B.3.

• The proofs for the convergence of Transformer are included in Appendix C.

• The derivations and the proofs for the NTK are further elaborated in Appendix D including the formulation of NTK, minimum eigenvalue of NTK, and the relationship between the Hessian and the width.

• Further details on the experiments are developed in Appendix E.

• Limitations and societal impact of this work are discussed in Appendix F and Appendix G respectively.
A Symbols and Notation

We include the core symbols and notation in Table 2 for facilitating the understanding of our work.

Table 2: Core symbols and notations used in this paper.

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Dimension(s)</th>
<th>Definition</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\mathcal{N}(\mu, \sigma)$</td>
<td>-</td>
<td>Gaussian distribution of mean $\mu$ and variance $\sigma$</td>
</tr>
<tr>
<td>$[\mathbf{w}]_i$</td>
<td>-</td>
<td>Euclidean norm of vectors $\mathbf{w}$</td>
</tr>
<tr>
<td>$|\mathbf{w}|_F$</td>
<td>-</td>
<td>Frobenius norm of matrix $\mathbf{W}$</td>
</tr>
<tr>
<td>$\lambda_{\max}(\mathbf{W})$, $\lambda_{\min}(\mathbf{W})$</td>
<td>-</td>
<td>Minimum and Maximum eigenvalues of matrix $\mathbf{W}$</td>
</tr>
<tr>
<td>$\sigma_{\max}(\mathbf{W})$, $\sigma_{\min}(\mathbf{W})$</td>
<td>-</td>
<td>Minimum and Maximum singular values of matrix $\mathbf{W}$</td>
</tr>
<tr>
<td>$\mathbf{W}^{(i)}$</td>
<td>-</td>
<td>$i$-th element of matrices $\mathbf{w}$</td>
</tr>
<tr>
<td>$\mathbf{W}^t$</td>
<td>-</td>
<td>Matrix $\mathbf{W}$ at time step $t$</td>
</tr>
<tr>
<td>$\odot$</td>
<td>-</td>
<td>Hadamard product</td>
</tr>
<tr>
<td>$\sigma_i(\mathbf{W})$</td>
<td>-</td>
<td>Row wise Softmax activation for matrix $\mathbf{W}$</td>
</tr>
<tr>
<td>$\sigma_i(\mathbf{W})$</td>
<td>-</td>
<td>Element-wise ReLU activation for matrix $\mathbf{W}$</td>
</tr>
<tr>
<td>$\mathcal{N}$</td>
<td>-</td>
<td>Size of the dataset</td>
</tr>
<tr>
<td>$d_m$</td>
<td>-</td>
<td>Width of intermediate layer</td>
</tr>
<tr>
<td>$d_s$</td>
<td>-</td>
<td>Sequence length of the input</td>
</tr>
<tr>
<td>$d$</td>
<td>-</td>
<td>Dimension of each token</td>
</tr>
<tr>
<td>$\eta_Q, \eta_K, \eta_Y, \eta_D$</td>
<td>-</td>
<td>Variance of Gaussian initialization of $\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_Y, \mathbf{W}_D$</td>
</tr>
<tr>
<td>$\tau$</td>
<td>-</td>
<td>Step size</td>
</tr>
<tr>
<td>$\mathbf{X}_n$</td>
<td>$\mathbb{R}^{d \times d}$</td>
<td>The $n$-th data point</td>
</tr>
<tr>
<td>$y_n$</td>
<td>$\mathbb{R}$</td>
<td>The $n$-th target</td>
</tr>
<tr>
<td>$\mathbf{D}_X, \mathbf{D}_Y$</td>
<td>-</td>
<td>Input data distribution, Target data distribution</td>
</tr>
<tr>
<td>$\beta_{i,n} := \sigma_x(\mathbf{W}_O \mathbf{X}_n^i \mathbf{W}_h \mathbf{X}_n^i)^\top$</td>
<td>$\mathbb{R}^d$</td>
<td>The $i$-th row of the output of softmax of the $n$-th data point</td>
</tr>
<tr>
<td>$\mathbf{y} \triangleq [y_1, \ldots, y_N]^\top$</td>
<td>$\mathbb{R}^N$</td>
<td>Ground truth label of the data samples ${\mathbf{X}<em>n}</em>{n=1}^N$</td>
</tr>
<tr>
<td>$\mathbf{f}(\theta) \triangleq [f(\mathbf{X}_1, \theta), \ldots, f(\mathbf{X}_N, \theta)]^\top$</td>
<td>$\mathbb{R}^N$</td>
<td>Network output given data samples ${\mathbf{X}<em>n}</em>{n=1}^N$</td>
</tr>
<tr>
<td>$\mathbf{f}<em>{\theta</em>{\text{me}}} \triangleq [f_{\theta_1}(\mathbf{X}<em>1, \theta), \ldots, f</em>{\theta_N}(\mathbf{X}_N, \theta)]$</td>
<td>$\mathbb{R}^N \times d_m$</td>
<td>Network output given data samples $\mathbf{X}$</td>
</tr>
<tr>
<td>$\mathbf{F}<em>{\theta</em>{\text{me}}} \triangleq [f_{\theta_1}(\mathbf{X}<em>1, \theta), \ldots, f</em>{\theta_N}(\mathbf{X}_N, \theta)]$</td>
<td>$\mathbb{R}^N \times d_m$</td>
<td>Output of the hidden layer given $\mathbf{X}$</td>
</tr>
<tr>
<td>$\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_F, \mathbf{W}_O$</td>
<td>$\mathbb{R}^{d_{\text{in}} \times d_m}$</td>
<td>Learnable parameters</td>
</tr>
<tr>
<td>$\mathbf{W}_Q, \mathbf{W}_K, \mathbf{W}_F, \mathbf{W}_O$</td>
<td>$\mathbb{R}^{d_{\text{in}} \times d_m}$</td>
<td>Learnable parameters</td>
</tr>
<tr>
<td>$\lambda_Q \triangleq |\mathbf{W}_Q|_2 + C_Q, \lambda_K \triangleq |\mathbf{W}_K|_2 + C_K$</td>
<td>$\mathbb{R}$</td>
<td>Parameters norm</td>
</tr>
<tr>
<td>$\lambda_V \triangleq |\mathbf{W}_V|_2 + C_V, \lambda_D \triangleq |\mathbf{W}_D|_2 + C_D$</td>
<td>$\mathbb{R}$</td>
<td>Parameters norm</td>
</tr>
<tr>
<td>$\rho \triangleq N^{1/2} d_{\text{in}}^{1/2} \gamma_C \gamma_{\text{scale}}$, $z \triangleq \lambda_d^2 (1 + 4 \gamma_C^2 \gamma_{\text{scale}}^2 \lambda_d^2 \gamma_C^2)$</td>
<td>$\mathbb{R}$</td>
<td>Auxiliary terms</td>
</tr>
<tr>
<td>$f_{\theta}$</td>
<td>$\mathbb{R}$</td>
<td>Output of network for input $\mathbf{X}_n$</td>
</tr>
<tr>
<td>$\gamma, \delta, \omega$ and $\theta$</td>
<td>-</td>
<td>Standard Bachmann–Landau order notation</td>
</tr>
</tbody>
</table>

* The superscript with bracket represents the element of a vector/matrix, e.g., $\mathbf{w}^{(i)}$ is the $i$-th element of $\mathbf{w}$.  
* The superscript without bracket symbolizes the parameters at different training steps, e.g., $\theta^t$.  
* The subscript without bracket symbolizes the variable associated to the $n$-th data sample, e.g., $\mathbf{X}_n$. 

B Theoretical background

B.1 Preliminary on NTK

In this section, we summarize how training a neural network by minimizing squared loss, i.e., $\ell(\theta^t) = \frac{1}{2} \sum_{n=1}^N (f(\mathbf{X}_n; \theta^t) - y_n)^2$, via gradient descent can be characterized by the kernel regression predictor with NTK.

By choosing an infinitesimally small learning rate, we can obtain the following gradient flow:

$$\frac{d\theta^t}{dt} = -\nabla \ell(\theta^t).$$
By substituting the loss into the above equation and using the chain rule, we can find that the network outputs \( f(\theta^t) \in \mathbb{R}^N \) admit the following dynamics:

\[
\frac{df(\theta^t)}{dt} = -K^t(f(\theta^t) - y),
\]

where \( K^t = \left( \frac{\partial f(\theta^t)}{\partial \theta} \right) \left( \frac{\partial f(\theta^t)}{\partial \theta} \right)^T \in \mathbb{R}^{N \times N} \). Jacot et al. [2018], Arora et al. [2019b] have shown that for fully-connected neural networks, under the infinite-width setting and proper initialization, \( K^t \) will be stable during training and \( K^0 \) will converge to a fixed matrix \( K^* \in \mathbb{R}^{N \times N} \), where \( (K^*)(ij) = K^*(X_i, X_j) \) is the NTK value for the inputs \( X_i \) and \( X_j \). Then, we rewrite Eq. (5) as:

\[
\frac{df(\theta^t)}{dt} = -K^*(f(\theta^t) - y).
\]

This implies the network output for any \( X \in \mathbb{R}^{d_i \times d} \) can be calculated by the kernel regression predictor with the associated NTK:

\[
f(X) = (K^*(X, X_1), \cdots, K^*(X, X_N)) \cdot (K^*)^{-1} y,
\]

where \( K^*(X, X_n) \) is the kernel value between test data \( X \) and training data \( X_n \).

### B.2 Preliminary on Sub-Exponential random variables

Below, we overview the definition of a sub-exponential random variable and few related lemma based on Wainwright [2019]. A random variable \( X \) with mean \( \mu \) is called sub-exponential random variable if there exist non-negative parameters \( (\nu, \alpha) \) such that

\[
\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\lambda^2 \nu^2}{2}} \quad \text{for all } |\lambda| < \frac{1}{\alpha},
\]

and we denote by \( X \sim SE(\nu, \alpha) \).

**Lemma 2.** The multiplication of a scalar \( s \in \mathbb{R}^+ \) and a sub-exponential random variable \( X \sim SE(\nu, \alpha) \) is still a sub-exponential random variable: \( sX \sim SE(s\nu, s\alpha) \).

**Lemma 3.** Given a set of independent sub-exponential random variables \( X_i \sim (\nu_i, \alpha_i) \) for \( i \in 1, \ldots, N \), then \( \sum_{i=1}^{N} X_i \sim (\sqrt{\sum_{i=1}^{N} \nu_i^2}, \max_i \alpha_i) \).

**Lemma 4.** Given a sub-exponential random variable \( X \sim SE(\nu, \alpha) \) with mean \( \mu \), the following inequality holds with probability at least \( 1 - \delta \):

\[
|X - \mu| < \max \left( \nu \sqrt{2 \log \frac{2}{\delta}}, 2\alpha \log \frac{2}{\delta} \right).
\]

### B.3 Related work on over-parameterization for convergence analysis

Recent empirical observation shows that neural networks can fit arbitrary labels with zero training loss when applying the first-order methods, e.g., gradient descent (GD) [Zhang et al. 2017]. Due to the highly non-convex and non-smooth instinct of the neural network, a large body of work have attempted to explain such a phenomenon. Early work studied the convergence of stochastic gradient descent (SGD) for training two-layer over-parameterized ReLU network with cross-entropy loss [Li and Liang 2018], Du et al. [2019b] show that only training the first layer of two-layer ReLU network with square loss by GD can lead to global convergence under the assumption that the Gram matrix is positive definite. Du et al. [2019a] extend the result to deep neural network with smooth activation function and shows that the convergence is guaranteed when the widths of all the hidden layers scale in \( \Omega(N^2) \), where \( N \) is the number of data points. Meanwhile, Zou and Gu [2019] prove that the condition for the convergence of GD for deep ReLU network is \( \Omega(N^6) \), which improves upon Allen-Zhu et al. [2019a] that show the result of \( \Omega(N^{4}) \). Allen-Zhu et al. [2019a] also provide several convergence analyses under the setting of SGD and various loss functions. The assumption regarding the positive definiteness of the Gram matrix made in Du et al. [2019b] has been rigorously proved in Nguyen et al. [2021]. This facilitates Nguyen [2021] to demonstrate that deep ReLU network under LeCUN initialization with width in the order \( \Omega(N^4) \) is enough for global convergence. Recent breakthrough [Bombari et al. 2022] improves previous results by showing that sub-linear layer widths suffice for deep neural network with smooth activation function.
Table 3: Over-parameterization conditions for the convergence analysis of neural network under gradient descent training with squared loss. $L$ is the depth of the network.

<table>
<thead>
<tr>
<th>Model</th>
<th>Depth</th>
<th>Initialization</th>
<th>Activation</th>
<th>Width</th>
</tr>
</thead>
<tbody>
<tr>
<td>Allen-Zhu et al. [2019a]</td>
<td>FCNN/CNN</td>
<td>Deep</td>
<td>NTK</td>
<td>ReLU $\Omega(N^{24L^{12}})$</td>
</tr>
<tr>
<td>Du et al. [2019a]</td>
<td>FCNN/CNN</td>
<td>Deep</td>
<td>NTK</td>
<td>Smooth $\Omega(N^{42O(L)})$</td>
</tr>
<tr>
<td>Oymak and Soltanolkotabi [2020]</td>
<td>FCNN</td>
<td>Shallow</td>
<td>Standard Gaussian</td>
<td>ReLU $\Omega(N^2)$</td>
</tr>
<tr>
<td>Zou and Gu [2019]</td>
<td>FCNN</td>
<td>Deep</td>
<td>He</td>
<td>ReLU $\Omega(N^8L^{12})$</td>
</tr>
<tr>
<td>Du et al. [2019b]</td>
<td>FCNN</td>
<td>Shallow</td>
<td>NTK</td>
<td>ReLU $\Omega(N^4)$</td>
</tr>
<tr>
<td>Nguyen [2021]</td>
<td>FCNN</td>
<td>Deep</td>
<td>LeCun</td>
<td>Smooth $\Omega(N^{3/2})$</td>
</tr>
<tr>
<td>Chen et al. [2021]</td>
<td>FCNN</td>
<td>Shallow</td>
<td>He/LeCun</td>
<td>Smooth $\Omega(N^{3/2})$</td>
</tr>
<tr>
<td>Song et al. [2021]</td>
<td>FCNN</td>
<td>Deep</td>
<td>He/LeCun</td>
<td>Smooth $\Omega(N^c)$</td>
</tr>
<tr>
<td>Bombari et al. [2022]</td>
<td>FCNN</td>
<td>Deep</td>
<td>LeCun</td>
<td>$\Omega(N^{3/2})$</td>
</tr>
<tr>
<td>Allen-Zhu et al. [2019b]</td>
<td>RNN</td>
<td>-</td>
<td>NTK</td>
<td>ReLU $\Omega(N^c)$, $c &gt; 1$</td>
</tr>
<tr>
<td>Hron et al. [2020]</td>
<td>Transformer</td>
<td>Deep</td>
<td>NTK</td>
<td>ReLU -</td>
</tr>
<tr>
<td>Yang [2020]</td>
<td>Transformer</td>
<td>Deep</td>
<td>NTK</td>
<td>Softmax+ReLU -</td>
</tr>
<tr>
<td>Our</td>
<td>Transformer</td>
<td>Shallow</td>
<td>Table 1</td>
<td>Softmax+ReLU $\Omega(N)$</td>
</tr>
</tbody>
</table>

C Proof for convergence analysis

This section is developed for the proof of the convergence result and we outline the flowchart below: Specifically, in Appendix C.1, we provide some auxiliary lemmas. Lemma 5 and Lemma 6 show that the norm of the parameters can be bounded with high probability at initialization. Lemmas 7, 9 and 10 present that the network output and the output of softmax between two adjacent time steps can be upper bounded. The Lipschitzness of network gradient and its norm is bounded in Lemmas 13 and 14. In Appendix C.2, we prove the convergence of the general cases, i.e., Proposition 1. In Appendix C.3 and C.4 we present the proof for $d_m^{-1/2}$ and $d_m^{-1}$ scaling.

C.1 Auxiliary lemmas

**Lemma 5** (Corollary 5.35 of Vershynin [2012]). For a weight matrix $W \in \mathbb{R}^{d_1 \times d_2}$ where each element is sampled independently from $N(0,1)$, for every $\zeta \geq 0$, with probability at least $1 - 2 \exp(-\zeta^2/2)$ one has:

$$\sqrt{d_1} - \sqrt{d_2} - \zeta \leq \sigma_{\text{min}}(W) \leq \sigma_{\text{max}}(W) \leq \sqrt{d_1} + \sqrt{d_2} + \zeta,$$

where $\sigma_{\text{max}}(W)$ and $\sigma_{\text{min}}(W)$ represents the maximum and minimum singular value of $W$, respectively.

**Lemma 6** (Upper bound of spectral norms of initial weight). For a weight matrix $W \in \mathbb{R}^{d_m \times d}$ where $d_m > d$, each element is sampled independently from $N(0,1)$, with probability at least $1 - 2 \exp(-d_m/2)$, one has:

$$\|W\|_2 \leq 3\sqrt{d_m}.$$  

**Proof of Lemma 6** Following Lemma 5 one has:

$$\|W\|_2 \leq \sqrt{d_m} + \sqrt{d} + \zeta.$$  

Letting $\zeta = \sqrt{d_m}$ and using the fact that $d_m > d$ complete the proof.

**Lemma 7.** Recall from Table 2 $\beta_i = \sigma_s(\tau_0 X^{(i,:)}W_Q^T W_K X^T)_{i,:}$ is the i-th row of the output of Softmax, if $\max(\|W_Q\|_2, \|W_Q^T\|_2) \leq \lambda_Q, \max(\|W_K\|_2, \|W_K^T\|_2) \leq \lambda_K$, then its difference
Proof. Between $t'$ step and $t$ step has the following upper bound:

$$\|\beta'_t - \beta'_t\|_2 \leq 2\tau_0 C_2^2 d_s \left(\bar{\lambda}_Q \|W'_{K} - W'_K\|_2 + \bar{\lambda}_K \|W'_{Q} - W'_Q\|_2\right).$$

Proof. Let $\lambda_{\bar{\tau}} = \sqrt{\frac{\tau_0}{\tau}}$, we first have:

$$\|\beta'_t - \beta'_t\|_2 \leq \|\beta'_t - \beta'_t\|_1 \leq 2 \|\sigma_0 X^{(i,:)} W^T_{Q} W'_{K} X^T - \tau_0 X^{(i,:)} W^T_{Q} W'_{K} X^T\|_\infty$$

(Because Corollary A.7 in Edelman et al. [2022])

$$= 2 \max_j \|\sigma_0 X^{(i,:)} W^T_{Q} W'_{K} X^T - \tau_0 X^{(i,:)} W^T_{Q} W'_{K} X^T\|_2$$

(Because Cauchy-Schwarz inequality)

$$\leq 2 \max_j (\tau_0 \|\sigma_0 X^{(i,:)} W^T_{Q} W'_{K} - W'_{K} X^T\|_2 \|X^{(i,:)}\|_2)$$

(Because Cauchy-Schwarz inequality, Triangle inequality)

$$\leq 2 \tau_0 C_2^2 d_s \|W^T_{Q} W'_{K} - W'_{Q} W'_{K}\|_2$$ (By Assumption[1])

$$\leq 2 \tau_0 C_2^2 d_s \left(\|W'^T_{Q} W'_{K} - W'^T_{Q} W'^T_{K}\|_2 + \|W'^T_{Q} W'^T_{K}\|_2\right)$$ (By Cauchy-Schwarz inequality, Triangle inequality)

$$\leq 2 \tau_0 C_2^2 d_s \left(\bar{\lambda}_Q \|W'_{K} - W'_{K}\|_2 + \bar{\lambda}_K \|W'_{Q} - W'_{Q}\|_2\right).$$

$$\lambda_{\bar{\tau}} = \sqrt{\frac{\tau_0}{\tau}}$$

Lemma 8 (Upper bound the Euclidean norm of the output of softmax). Suppose $v = \text{Softmax}(u) \in \mathbb{R}^{d_s}$, then one has: $1/\sqrt{d_s} \leq \|v\|_2 \leq 1$.

Proof. By the inequality of arithmetic and geometric mean:

$$\frac{1}{\sqrt{d_s}} = \frac{1}{\sqrt{d_s}} \leq \|v\|_2 = \sqrt{\sum_{i=1}^{d_s} \left(\begin{smallmatrix} v(i) \end{smallmatrix}\right)^2} = \sqrt{\frac{d_s}{\sum_{i=1}^{d_s} \left(\begin{smallmatrix} v(i) \end{smallmatrix}\right)^2}} = 1.$$

Lemma 9. If at $t'$ and $t$ step, max $\left(\|W_{Q}\|_2, \|W_{K}\|_2\right) \leq \bar{\lambda}_Q$, max $\left(\|W_{Q}\|_2, \|W_{K}\|_2\right) \leq \bar{\lambda}_L$, then the difference between the output of the last hidden layer at $t'$ step and $t$ can be upper bounded by:

$$\|F'_{pre} - F'_{pre}\|_F \leq \rho \left(\|W'_V - W'_V\|_2 + \bar{\lambda}_V 2\tau_0 C_2^2 d_s \left(\bar{\lambda}_Q \|W'_{K} - W'_{K}\|_2 + \bar{\lambda}_K \|W'_{Q} - W'_{Q}\|_2\right)\right).$$

Proof. Note that $F_{pre} \in \mathbb{R}^{N \times d_m}$ is the output of the last hidden layer given $\{X_n\}_{n=1}^{N}$ sample as defined in Table 2. We firstly analyze each sample, i.e., $f_{pre} \in \mathbb{R}^{d_m}$ and we drop the index $n$ for simplification.
where the last equality is by the definition of $\rho$.

**Lemma 10.** If at $t'$ and $t$ step, 

$$\max (\|W_Q^t\|_2, \|W_Q^{t'}\|_2) \leq \bar{\lambda}_Q, \max (\|W_K^t\|_2, \|W_K^{t'}\|_2) \leq \bar{\lambda}_K, \max (\|W^t\|_2, \|W^{t'}\|_2) \leq \bar{\lambda}_W, \max (\|W^t_O\|_2, \|W^{t'}_O\|_2) \leq \bar{\lambda}_O,$$

then the difference between the network output at $t'$ step and $t$ step can be upper bounded by:

$$\|f^{t'} - f^t\|_2 \leq \rho \bar{\lambda}_W \|w^t_O - w^{t'}_O\|_2 + \bar{\lambda}_O \|F^{t'}_{\text{pre}} - F^t_{\text{pre}}\|_F.$$

**Proof.**

$$\|f^{t'} - f^t\|_2 = \|F^{t'}_{\text{pre}} w^{t'}_O - F^t_{\text{pre}} w^t_O\|_2 \leq \|F^{t'}_{\text{pre}}\|_F \|w^{t'}_O - w^t_O\|_2 + \|w^t_O\|_2 \|F^{t'}_{\text{pre}} - F^t_{\text{pre}}\|_2,$$

where we use triangle inequality. Then, the first term of the RHS can be bounded by:

$$\|F^{t'}_{\text{pre}}\|_F \leq \|F^{t'}_{\text{pre}}\|_F \leq \tau_1 \sum_{i=1}^{d_x} \sigma_r \left( W_{V_i} X^T \beta_i^{t'} \right) \leq \tau_1 \sum_{i=1}^{d_x} \sigma_r \left( W_{V_i} X^T \beta_i^t \right) \leq \tau_1 \sum_{i=1}^{d_x} \sigma_r \left( W_{V_i} X^T \beta_i^{t'} \right),$$

where the last equality is by the definition of $\rho$.

**Lemma 11** (Jacobi of Softmax). Suppose $v = \text{Softmax}(u) \in \mathbb{R}^{d_v}$, then $rac{\partial v}{\partial u} = \text{diag}(v) - vv^T$.

**Proof.** We can reformulate $v$ as:

$$v = \begin{bmatrix} \exp(u^{(1)}) \\ \sum_{i=1}^{d_x} \exp(u^{(i)}) \\ \vdots \\ \exp(u^{(d_x)}) \\ \sum_{i=1}^{d_x} \exp(u^{(i)}) \end{bmatrix}.$$

20
Then, we have

\[
\frac{\partial v^{(j)}}{\partial u^{(k)}} = \frac{\partial}{\partial u^{(k)}} \frac{\exp(u^{(j)})}{\sum_{i=1}^d \exp(u^{(i)})} = \begin{cases} \frac{-\exp(u^{(j)}) - \exp(u^{(k)})}{\sum_{i=1}^d \exp(u^{(i))})^2} & \text{if } j \neq k \\ \frac{-\exp(u^{(j)}) - \exp(u^{(k)})}{\sum_{i=1}^d \exp(u^{(i))})^2} & \text{if } j = k \end{cases}
\]

Thus

\[
\frac{\partial v}{\partial u} = \text{diag}(v) - vv^T.
\]

Lemma 12 (Upper bound for the loss gradient norm). If \(\|W_Q^0\|_2 \leq \lambda_Q, \|W_K^1\|_2 \leq \lambda_K, \|W_V^1\|_2 \leq \lambda_V, \|w_O^0\|_2 \leq \lambda_O\), then the gradient norm with respect to \(W_Q, W_K, W_V, w_O\) can be upper bounded by:

\[
\|\nabla w_Q \ell(\theta^\top)\|_F \leq 2\rho_0 \lambda_K \lambda_V \lambda_O d_s C_x^2 \|f - y\|_2, \quad \|\nabla w_K \ell(\theta^\top)\|_F \leq 2\rho_0 \lambda_Q \lambda_V \lambda_O d_s C_x^2 \|f - y\|_2, \quad \|\nabla w_v \ell(\theta^\top)\|_F \leq \rho \lambda_O \|f - y\|_2, \quad \|\nabla w_o \ell(\theta^\top)\|_2 \leq \rho \lambda_V \|f - y\|_2.
\]

Proof. To simplify the notation, in the proof below, we hide the index \(t\). Firstly, consider the gradient w.r.t \(W_V\),

\[
\|\nabla w_v \ell(\theta^\top)\|_F = \left\| -\sum_{n=1}^N (f(X_n) - y_n) \frac{\partial f(X_n)}{\partial W_v} \right\|_F = \tau_1 \left\| \sum_{n=1}^N (f(X_n) - y_n) \sum_{i=1}^d (w_O \circ \sigma_r(W_V X_n^\top \beta_{i,n})) \beta_{i,n}^\top X_n \right\|_F \leq \tau_1 \sum_{n=1}^N \|f(X_n) - y_n\| d_s^{3/2} \lambda_O C_x \leq \tau_1 \lambda_V \lambda_O d_s^{3/2} \lambda_O C_x \sqrt{N} \|f - y\|_2 = \rho \lambda_O \|f - y\|_2,
\]

where the last equality is by the definition of \(\rho\) in Proposition. \(\square\) Next, consider the gradient w.r.t \(w_O\),

\[
\|\nabla w_o \ell(\theta^\top)\|_2 = \left\| -\sum_{n=1}^N (f(X_n) - y_n) \frac{\partial f(X_n)}{\partial w_o} \right\|_2 = \tau_1 \left\| \sum_{n=1}^N (f(X_n) - y_n) \sum_{i=1}^d \sigma_r(W_V X_n^\top \beta_{i,n}) \right\|_2 \leq \tau_1 \sum_{n=1}^N \|f(X_n) - y_n\| d_s^{3/2} \lambda_V C_x \leq \tau_1 \lambda_V \lambda_O d_s^{3/2} \lambda_V C_x = \rho \lambda_V \|f - y\|_2,
\]

(8)
Then following the step as in Eqs. (7) to (10), each term can be bounded as follows:

Next, consider the gradient w.r.t $W_Q$,

$$
\|\nabla_w \ell (\theta)\|_F = \left\| - \sum_{n=1}^{N} (f(X_n) - y_n) \frac{\partial f(X_n)}{\partial W_Q} \right\|_F
= \tau_0 \tau_1 \left\| \sum_{n=1}^{N} (f(X_n) - y_n) \sum_{i=1}^{d_x} W_k X_n^\top \left( \text{diag}(\beta_{i,n}) - \beta_{i,n} \beta_{i,n}^\top \right) X_i W_V^\top (w_O \circ \sigma_r (W_V X_n \beta_{i,n})) X_n^{(i,:)} \right\|_F
\leq 2\tau_0 \tau_1 \sum_{n=1}^{N} \left| (f(X_n) - y_n) \right| d_s \lambda_K \lambda_V \lambda_O (C_x \sqrt{d_s})^3
\leq 2\tau_0 \tau_1 N \sum_{n=1}^{N} \left| (f(X_n) - y_n) \right|^2 d_s \lambda_K \lambda_V \lambda_O (C_x \sqrt{d_s})^3
= 2\tau_0 \lambda_K \lambda_V \lambda_O d_s C_x^2 \|f - y\|_2.
$$

Similarly, for the gradient w.r.t $W_K$, we have:

$$
\|\nabla_w \ell (\theta)\|_F \leq 2\tau_0 \lambda_Q \|W_K\|_2 \leq \lambda_K, \|W_V\|_2 \leq \lambda_V, \|w_O\|_2 \leq \lambda_O, \text{then one has:}
\|\nabla_w f^{(i)}\|_2 \leq c_2.
$$

where

$$
c_2 \triangleq \rho \sqrt{\lambda_o^2 + \lambda_v^2} + (2\tau_0 \lambda_K \lambda_V \lambda_O d_s C_x^2)^2 + (2\tau_0 \lambda_Q \lambda_V \lambda_O d_s C_x^2)^2.
$$

**Proof.** To simplify the notation, in the proof below, we hide the index $t$. Firstly, note that:

$$
\|\nabla_{\theta} f\|_2 \leq \|\nabla_{\theta} f\|_F
= \sqrt{\sum_{n=1}^{N} \left( \|\nabla_{w_0} f(X_n; \theta)\|_F^2 + \|\nabla_{w_0} f(X_n; \theta)\|_F^2 + \|\nabla_{w_k} f(X_n; \theta)\|_F^2 + \|\nabla_{w_k} f(X_n; \theta)\|_F^2 \right)}.
$$

Then following the step as in Eqs. (7) to (10), each term can be bounded as follows:

$$
\sum_{n=1}^{N} \left( \|\nabla_{w_0} f(X_n; \theta)\|_F^2 \right) \leq (\rho \lambda_o)^2, \quad \sum_{n=1}^{N} \left( \|\nabla_{w_0} f(X_n; \theta)\|_2 \right) \leq (\rho \lambda_V)^2,
\sum_{n=1}^{N} \left( \|\nabla_{w_0} f(X_n; \theta)\|_F^2 \right) \leq (2\rho \lambda_K \lambda_V \lambda_O d_s C_x^2)^2, \quad \sum_{n=1}^{N} \left( \|\nabla_{w_k} f(X_n; \theta)\|_F \right) \leq (2\rho \lambda_Q \lambda_V \lambda_O d_s C_x^2)^2.
$$

Plugging these bounds back Eq. (12) finishes the proof.

**Lemma 14** (Upper bound for the Lipschitzness of the network gradient). Suppose

$$
\max (\|W_Q\|_2, \|W_V\|_2) \leq \lambda_Q, \max (\|W_K\|_2, \|W_V\|_2) \leq \lambda_K, \max (\|W_V\|_2, \|W_V\|_2) \leq \lambda_V, \max (\|w_O\|_2, \|w_O\|_2) \leq \lambda_O, \text{and define } z \triangleq 2\tau_0 C_x^2 d_s (\lambda_Q + \lambda_K), \text{then one has}
\|\nabla_{\theta} f^{(i')} - \nabla_{\theta} f^{(i')}\|_2 \leq c_3 \|\theta^{(i')} - \theta^{(i')}\|_2,
$$

\(c_3\) and \(c_2\) are constants.

22
where
\[(c_3)^2 \triangleq N(\tau_1 C_x d_x^{3/2}(1 + \bar{\lambda}_V z))^2 + N \left\{ \tau_1 C_x d_x^{3/2} \left[ \bar{\lambda}_O z + \bar{\lambda}_O C_x \sqrt{d_x(1 + \bar{\lambda}_V z)} + 1 \right] \right\}^2 + N \left\{ \tau_0 \tau_1 C_x d_x \left\{ 2 \bar{\lambda}_K d_x C_x^2 \left[ \bar{\lambda}_V \left( \bar{\lambda}_O C_x \sqrt{d_x(1 + \bar{\lambda}_V z)} + 1 \right) + \bar{\lambda}_O \right] + \bar{\lambda}_V \bar{\lambda}_O [C_x^2 d_x + 3 \bar{\lambda}_K z] \right\}^2 \right\}^2.
\]

Proof. Firstly, note that:
\[
\| \nabla \theta f' - \nabla \theta f \|_2^2 \\
\leq \sum_{n=1}^N \left( \left\| \nabla_{w_o} f(X_n; \theta') - \nabla_{w_o} f(X_n; \theta) \right\|_2 + \left\| \nabla_{w_V} f(X_n; \theta') - \nabla_{w_V} f(X_n; \theta) \right\|_2 \\
+ \left\| \nabla_{w_Q} f(X_n; \theta') - \nabla_{w_Q} f(X_n; \theta) \right\|_2 + \left\| \nabla_{w_K} f(X_n; \theta') - \nabla_{w_K} f(X_n; \theta) \right\|_2 \right).
\]

Then, we will prove each term separately. Regarding the first term in Eq. (15), we have
\[
\left\| \nabla_{w_o} f(X_n; \theta') - \nabla_{w_o} f(X_n; \theta) \right\|_2 \\
= \left\| f'_{\text{pre}} - f'_{\text{pre}} \right\|_2 \\
\leq \tau_1 C_x d_x^{3/2} \left( \left\| W_{\theta'}' - W_{\theta} \right\|_2 + \bar{\lambda}_V 2 \tau_0 C_x^2 d_x \left( \bar{\lambda}_O \left\| W_{\theta'}' - W_{\theta} \right\|_2 + \bar{\lambda}_K \left\| W_{\theta'}' - W_{\theta} \right\|_2 \right) \right) \\
\leq \tau_1 C_x d_x^{3/2} \left( 1 + \bar{\lambda}_V 2 \tau_0 C_x^2 d_x \left( \bar{\lambda}_O + \bar{\lambda}_K \right) \right) \| \theta' - \theta \|_2 \\
= \tau_1 C_x d_x^{3/2} \left( 1 + \bar{\lambda}_V z \right) \| \theta' - \theta \|_2.
\]

where the first inequality is by Eq. (6), and in the last equality is by the definition of \( z \) in the lemma. Regarding the second term in Eq. (15), we have:
\[
\left\| \nabla_{w_V} f(X_n; \theta') - \nabla_{w_V} f(X_n; \theta) \right\|_F \\
= \tau_1 \sum_{i=1}^{d_x} \left( w_{\theta'}' \circ \sigma_r \left( W_{\theta'}' X_n \beta_{i,n}' \right) \right) \beta_{i,n}'^T X_n - \sum_{i=1}^{d_x} \left( w_{\theta} \circ \sigma_r \left( W_{\theta} X_n \beta_{i,n} \right) \right) \beta_{i,n}^T X_n \\
\leq \tau_1 C_x \sqrt{d_x} \sum_{i=1}^{d_x} \left\| w_{\theta'}' \circ \sigma_r \left( W_{\theta'}' X_n \beta_{i,n}' \right) \right\|_2 \| \beta_{i,n}' - \beta_{i,n} \|_2 + \left\| \beta_{i,n} \right\|_2 \left\| w_{\theta} \circ \sigma_r \left( W_{\theta} X_n \beta_{i,n} \right) \right\|_2 \\
\leq \tau_1 C_x \sqrt{d_x} \sum_{i=1}^{d_x} \left[ \left\| w_{\theta'}' \circ \sigma_r \left( W_{\theta'}' X_n \beta_{i,n}' \right) \right\|_2 + \left\| \beta_{i,n}' - \beta_{i,n} \right\|_2 \right] + \left\| \beta_{i,n} \right\|_2 \left\| w_{\theta} \circ \sigma_r \left( W_{\theta} X_n \beta_{i,n} \right) \right\|_2 \\
\leq \tau_1 C_x \sqrt{d_x} \sum_{i=1}^{d_x} \left[ \bar{\lambda}_O \left\| \beta_{i,n}' - \beta_{i,n} \right\|_2 + \left\| \beta_{i,n}' \circ \sigma_r \left( W_{\theta'}' X_n \beta_{i,n}' \right) - w_{\theta} \circ \sigma_r \left( W_{\theta} X_n \beta_{i,n} \right) \right\|_2 \right].
\]

The term \( \| \beta_{i,n}' - \beta_{i,n} \|_2 \) can be bounded by Lemma 7 as follows:
\[
\| \beta_{i,n}' - \beta_{i,n} \|_2 \leq 2 \tau_0 C_x^2 d_x \left( \bar{\lambda}_O \left\| W_{\theta'}' - W_{\theta} \right\|_2 + \bar{\lambda}_K \left\| W_{\theta'}' - W_{\theta} \right\|_2 \right) \leq z \| \theta' - \theta \|_2.
\]
The term \( \| w'_i \circ \sigma_r \left( W'_i X_n^\top \beta'_i \right) - w'_i \circ \sigma_r \left( W'_i X_n^\top \beta_{i,n} \right) \|_2 \) can be bounded by the triangle inequality, Cauchy–Schwarz inequality, and the same method in Eq. (6) as follows:

\[
\begin{align*}
\| w'_i \circ \sigma_r \left( W'_i X_n^\top \beta'_i \right) - w'_i \circ \sigma_r \left( W'_i X_n^\top \beta_{i,n} \right) \|_2 & \leq \| w'_i \|_2 \| \sigma_r \left( W'_i X_n^\top \beta'_i \right) - \sigma_r \left( W'_i X_n^\top \beta_{i,n} \right) \|_2 + \| \sigma_r \left( W'_i X_n^\top \beta'_i \right) \circ \left( w'_i - w'_i \right) \|_2 \\
& \leq \lambda_o \left( \| W'_i X_n^\top \beta'_i - W'_i X_n^\top \beta_{i,n} \|_2 + \| w'_i - w'_i \|_2 \right) \\
& \leq \lambda_o C_x \sqrt{d_x} \left( \| W'_i - W'_i \|_2 + \lambda_v 2 \pi_n C^2_{d_x} \left( \lambda_Q \| W'_i - W'_i \|_2 + \lambda_K \| W'_i - W'_i \|_2 \right) \right) + \| w'_i - w'_i \|_2 \\
& \leq \left( \lambda_o C_x \sqrt{d_x} (1 + \lambda_v z) + 1 \right) \| \theta' - \theta \|_2.
\end{align*}
\]

Plugging Eq. (18) and Eq. (19) back Eq. (17), we obtain:

\[
\begin{align*}
\| \nabla W f(X_n; \theta') - \nabla W f(X_n; \theta) \|_F & \leq \tau_1 C_x d_{i/2} \left( \lambda_o z + \lambda_o C_x \sqrt{d_x} (1 + \lambda_v z) + 1 \right) \| \theta' - \theta \|_2.
\end{align*}
\]

Regarding the third term in Eq. (15), let us denote by:

\[
\begin{align*}
U'_t_{i,n} &= W'_t X_n^\top \left( \text{diag}(\beta'_i_{i,n}) - \beta'_i_{i,n} \beta_{i,n}^\top \right) X_n, \quad h'_t_{i,n} = W'_t \left( w'_i \circ \sigma_r \left( W'_i X_n^\top \beta_{i,n} \right) \right), \\
U^t_{i,n} &= W'_t X_n^\top \left( \text{diag}(\beta'_i_{i,n}) - \beta'_i_{i,n} \beta_{i,n}^\top \right) X_n, \quad h^t_{i,n} = W'_t \left( w'_i \circ \sigma_r \left( W'_i X_n^\top \beta_{i,n} \right) \right).
\end{align*}
\]

Then:

\[
\begin{align*}
\| \nabla W f(X_n; \theta') - \nabla W f(X_n; \theta) \|_F & = \tau_0 \tau_1 \left| \sum_{i=1}^{d_x} \sum_{n=1}^{d_x} \left( U'_t_{i,n} h^t_{i,n} X_n^{(i,:)} - U^t_{i,n} h^t_{i,n} X_n^{(i,:)} \right) \right|_F \\
& \leq \tau_0 \tau_1 C_x \left( \sum_{i=1}^{d_x} \sum_{n=1}^{d_x} \| U'_t_{i,n} h^t_{i,n} \|_2^2 - \sum_{i=1}^{d_x} \sum_{n=1}^{d_x} \| U^t_{i,n} h^t_{i,n} \|_2^2 \right) \leq \tau_0 \tau_1 C_x \sum_{i=1}^{d_x} \left( \| U'_t_{i,n} \|_2 \| h^t_{i,n} - h^t_{i,n} \|_2 + \| h^t_{i,n} \|_2 \| U'_t_{i,n} - U^t_{i,n} \|_2 \right).
\end{align*}
\]

We bound each term separately:

\[
\begin{align*}
\| U'_t_{i,n} \|_2 & \leq \| W'_t \|_2 \| X_n \|_2 \| \text{diag}(\beta'_i_{i,n}) - \beta'_i_{i,n} \beta_{i,n}^\top \|_2 \leq 2 \lambda_K d_x C_x. \\
\| h^t_{i,n} \|_2 & \leq \| W'_t \|_2 \| w'_i \|_2 \leq \lambda_v \lambda_o.
\end{align*}
\]

\[
\begin{align*}
\| h^t_{i,n} - h^t_{i,n} \|_2 & \leq \| W'_t \|_2 \| w'_i \circ \sigma_r \left( W'_i X_n^\top \beta_{i,n} \right) - w'_i \circ \sigma_r \left( W'_i X_n^\top \beta_{i,n} \right) \|_2 + \| W'_t - W'_t \|_2 \| w'_i \|_2 \\
& \leq \left( \lambda_o \sqrt{d_x} (1 + \lambda_v z) + 1 \right) \| \theta' - \theta \|_2.
\end{align*}
\]
where the second inequality uses the result in Eq. (19). 
\[
\left\| U_{i,n}^t - U_{i,n}^0 \right\|_2 \\
\leq \left\| W_K^t X_n^T \left( \text{diag}(\beta_{i,n}^t) - \beta_{i,n}^t \beta_{i,n}^\top \right) - W_K^t X_n^T \left( \text{diag}(\beta_{i,n}^t) - \beta_{i,n}^t \beta_{i,n}^\top \right) \right\|_2 \left\| X_n \right\|_2 \\
\leq C_x \sqrt{d_s} \left[ \left\| W_K^t - W_K^0 \right\|_2 \left\| X_n \right\|_2 \left\| \text{diag}(\beta_{i,n}^t) - \beta_{i,n}^t \beta_{i,n}^\top \right\|_2 \\
+ \left\| W_K^t \right\|_2 \left\| X_n \right\|_2 \left\| \text{diag}(\beta_{i,n}^t) - \beta_{i,n}^t \beta_{i,n}^\top \right\|_2 \right] \\
\leq C_x^2 ds \left[ \| \theta^t - \theta^0 \|_2 + \lambda_K \left( \left\| \text{diag}(\beta_{i,n}^t) - \text{diag}(\beta_{i,n}^t) \right\|_2 + \left\| \beta_{i,n}^t \beta_{i,n}^\top - \beta_{i,n}^t \beta_{i,n}^\top \right\|_2 \right) \right] \tag{22} \\
\leq C_x^2 ds \left[ \| \theta^t - \theta^0 \|_2 + \lambda_K \left( \left\| \beta_{i,n}^t - \beta_{i,n}^t \right\|_\infty + \left\| \beta_{i,n}^t \beta_{i,n}^\top \right\|_2 \right) \right] \tag{23} \\
\leq C_x^2 ds \left[ \| \theta^t - \theta^0 \|_2 + 3 \lambda_K \left\| \beta_{i,n}^t - \beta_{i,n}^t \right\|_2 \right] \\
\leq C_x^2 ds \left[ \| \theta^t - \theta^0 \|_2 + 3 \lambda_K \tau_0 C_d^2 ds \left( \lambda_Q \left\| W_K^t - W_K^t \right\|_2 + \lambda_K \left\| W_K^t - W_K^t \right\|_2 \right) \right] \\
\leq \left[ C_x^2 ds + 3 \lambda_K \tau_0 \right] \| \theta^t - \theta^0 \|_2 ,
\]

where the last second inequality is by Lemma 7. Plugging back Eq. (21), we obtain:
\[
\left\| \nabla w_Q f(X_n; \theta^t) - \nabla w_Q f(X_n; \theta^0) \right\|_F \\
\leq \tau_0 \tau_1 C_x ds \left\{ 2 \lambda_K \tau_0 C_d^2 \left( \lambda_V \left( \lambda_Q \sqrt{d_s} (1 + \lambda_V z) + 1 \right) + \lambda_Q \lambda_O \left( C_d^2 ds + 3 \lambda_K z \right) \right) \right\} \| \theta^t - \theta^0 \|_2 . \tag{24}
\]

Similarly, the fourth term in Eq. (15) can be bounded by:
\[
\left\| \nabla w_K f(X_n; \theta^t) - \nabla w_K f(X_n; \theta^0) \right\|_F \\
\leq \tau_0 \tau_1 C_x ds \left\{ 2 \lambda_Q \tau_0 C_d^2 \left( \lambda_V \left( \lambda_Q \sqrt{d_s} (1 + \lambda_V z) + 1 \right) + \lambda_Q \lambda_O \left( C_d^2 ds + 3 \lambda_K z \right) \right) \right\} \| \theta^t - \theta^0 \|_2 . \tag{25}
\]

Plugging the upper bound for these four terms back Eq. (15) finishes the proof. \qed

C.2 Proof of Proposition 1

\textbf{Proof.} We can reformulate Eq. (1.2) as:
\[
f(X) = \tau_1 w_Q^T \sum_{i=1}^{d_s} \sigma_s \left( W_i X^\top \beta_i \right) = w_Q^T f_{\text{pre}},
\]
with \( \beta_i := \sigma_s \left( \tau_0 X^{(i:)} W_Q^\top W_K X^\top \right)^\top \in \mathbb{R}^{d_s} \). We show by induction for every \( t \geq 0 \)
\[
\begin{align*}
\left\| W_s^t \right\|_2 &\leq \lambda_Q, \left\| W_K^t \right\|_2 \leq \lambda_K, \quad s \in [0, t], \\
\left\| W_F^t \right\|_2 &\leq \lambda_V, \left\| w_Q^t \right\|_2 \leq \lambda_O, \quad s \in [0, t], \\
\sigma_{\min} \left( F_{\text{pre}}^t \right) &\geq \frac{1}{2} \sigma_s, \quad s \in [0, t], \\
\ell(\theta^s) &\leq \left( 1 - \gamma \alpha^2 z^2 \right)^s \ell(\theta^0), \quad s \in [0, t].
\end{align*}
\tag{25}
\]

It is clear that Eq. (25) holds for \( t = 0 \). Assume that Eq. (25) holds up to iteration \( t \). By the triangle inequality, we have
\[
\left\| W_Q^{t+1} - W_Q^0 \right\|_F \leq \sum_{s=0}^{t} \left\| W_Q^{s+1} - W_Q^s \right\|_F = \gamma \sum_{s=0}^{t} \left\| \nabla w_Q \ell(\theta^s) \right\|_F \\
\leq 2 \gamma \rho_0 \lambda_K \lambda_V \lambda_Q d_s C_d^2 \sum_{s=0}^{t} \sqrt{2\ell(\theta^s)} \leq 2 \gamma \rho_0 \lambda_K \lambda_V \lambda_Q d_s C_d^2 \sum_{s=0}^{t} \left( 1 - \frac{\gamma \alpha^2 z^2}{2} \right)^{s/2} \sqrt{2\ell(\theta^0)}, \tag{26}
\]

25
where the second inequality follows from the upper bound of the gradient norm in Lemma 12, and the last inequality follows from induction assumption. Let \( u := \sqrt{1 - \gamma \alpha^2/2} \), we bound the RHS of the previous expression:

\[
\rho 2r \lambda K \lambda V \lambda O d_s C_x^2 2 \alpha^2 \frac{1}{u^2} \frac{1 - u^{t+1}}{1 - u} \sqrt{2f(\theta^0)} \leq \rho 2r \lambda K \lambda V \lambda O d_s C_x^2 \frac{4}{\alpha^2} \sqrt{2f(\theta^0)}, \quad (\text{since } u \in (0, 1))
\]

\[
\leq C_Q. \quad (\text{by Eq. (27)}).
\]

By Weyl’s inequality, this implies:

\[
\|W_Q^{t+1}\|_2 \leq \|W_Q^0\|_2 + C_Q = \tilde{\lambda}_Q.
\]

Similarly,

\[
\|w_O^{t+1} - w_O^0\|_2 \leq \sum_{s=0}^t \|w_s^{t+1} - w_s^0\|_2 = \gamma \sum_{s=0}^t \|\nabla w_s \Phi(W_s)\|_2 \leq \gamma \rho \lambda V \sum_{s=0}^t \sqrt{2f(\theta^0)}
\]

\[
\leq \gamma \rho \lambda V \frac{\alpha^2}{2} \sqrt{2f(\theta^0)} \leq \rho \lambda V \frac{2}{\alpha^2} (1 - u^2) \frac{1 - u^{t+1}}{1 - u} \sqrt{2f(\theta^0)} \leq \rho \lambda V \frac{4}{\alpha^2} \sqrt{2f(\theta^0)} \leq C_O.
\]

By Weyl’s inequality, this implies

\[
\|w_O^{t+1}\|_2 \leq \|w_O^0\|_2 + C_O = \tilde{\lambda}_O.
\]

Similarly, we can obtain that

\[
\|W_K^{t+1} - W_K^0\|_F \leq 2r \rho \lambda Q \lambda V \lambda O d_s C_x^2 \frac{4}{\alpha^2} \sqrt{2f(\theta^0)}, \quad \|W_K^{t+1}\|_2 \leq \tilde{\lambda}_K,
\]

\[
\|W_V^{t+1} - W_V^0\|_F \leq \rho \lambda V \frac{4}{\alpha^2} \sqrt{2f(\theta^0)}, \quad \|W_V^{t+1}\|_2 \leq \tilde{\lambda}_V.
\]

Next, we will prove the fifth inequality in Eq. (25):

\[
\|F_{pre}^{t+1} - F_{pre}^0\|_F
\]

\[
\leq \rho \left\{ \|W_V^{t+1} - W_V^0\|_2 + \|W_V^0\|_2 \right\} 2r \rho \lambda V \frac{4}{\alpha^2} \sqrt{2f(\theta^0)} \left\{ \|W_K^{t+1}\|_2 + \|W_K^0\|_2 \right\}
\]

\[
\leq \sum_{s=0}^t \gamma \rho \left\{ \rho \lambda O \sqrt{2f(\theta^0)} + \lambda V 2r \lambda O d_s C_x^2 \right\} \left\{ \|W_K^{t+1}\|_2 + \|W_K^0\|_2 \right\}
\]

\[
\leq \rho^2 \lambda O \left[ 1 + 4 r_0^2 C_x^2 d_s \lambda V \right] \frac{16}{\alpha^2} \sqrt{2f(\theta^0)}
\]

\[
= \rho^2 \frac{z}{2} \frac{16}{\alpha^2} \sqrt{2f(\theta^0)}
\]

\[
\leq \frac{1}{2} \alpha, \quad (\text{by Eq. (31)}).
\]

The first inequality holds by Lemma 9. This result further implies that \( \sigma_{\min}(F_{pre}^{t+1}) \geq \frac{1}{2} \alpha \) by Weyl’s inequality. It remains to prove the last inequality in Eq. (25) holds for step \( t + 1 \).

We start by proving the Lipschitz constant for the gradient of the loss restricted to the interval \([\theta^t, \theta^{t+1}]\). We define \( \theta^{t+\phi} := \theta^t + \phi(\theta^{t+1} - \theta^t) \), for \( \phi \in [0, 1] \). Then, by triangle inequality and Cauchy–Schwarz inequality, we have:

\[
\|\nabla f(\theta^{t+\phi}) - \nabla f(\theta^t)\|_2
\]

\[
= \|\nabla f(\theta^{t+\phi}) - \nabla f(\theta^t)\|_2 \leq \|f^{t+\phi} - f^t\|_2 \|\nabla f(\theta^{t+\phi})\|_2 + \|f^{t+\phi} - f^t\|_2 \|\nabla f(\theta^{t+\phi}) - \nabla f^t\|_2
\]

\[
\leq \|f^{t+\phi} - f^t\|_2 \|\nabla f(\theta^{t+\phi})\|_2 + 2 \|\nabla f(\theta^{t+\phi}) - \nabla f^t\|_2 \|f^t - y\|_2
\]
where the last inequality is by induction rule. Then, we bound each term separately. Note that:

\[
\|W^{t+\phi}_Q - W^0_Q\|_F \leq \|W^{t+\phi}_Q - W^0_Q\|_F + \sum_{s=0}^{t-1} \|W^{s+1}_Q - W^s_Q\|_F
\]

and

\[
= \phi \gamma \|\nabla W Q \ell(\theta^T)\|_F + \gamma \sum_{s=0}^{t-1} \|\nabla W Q \ell(\theta^s)\|_F \leq \gamma \sum_{s=0}^t \|\nabla W Q \ell(\theta^s)\|_F.
\]

Then following exact the same step as in Eqs. (26), (27) and (30), we have: \(\|W^{t+\phi}_Q\|_2 \leq \lambda_Q\). By the same method, we have: \(\|W^t_{K}\|_2 \leq \lambda_K\), \(\|W^t_{V}\|_2 \leq \lambda_V\), \(\|w^t_{O}\|_2 \leq \lambda_O\). Now we proceed to bound the first term in Eq. (35).

\[
\|f^{t+\phi} - f^t\|_2 \leq \rho \|W^{t+\phi}_V\|_2 \|w^{t+\phi}_O - w^t_O\|_2 + \|w^t_O\|_2 \|F^{t+\phi}_{\text{pre}} - F^{t}_{\text{pre}}\|_2
\]

\[
\leq \rho \|W^{t+\phi}_V\|_2 \|w^{t+\phi}_O - w^t_O\|_2
\]

\[
+ \rho \|w^t_O\|_2 \left( \|W^{t+\phi}_V - W^t_V\|_2 + \|W^t_V\|_2 \right)
\]

\[
\leq \rho \lambda_V \|\theta^{t+\phi} - \theta^t\|_2 + \rho \lambda_O (1 + \lambda_V^2 C_2^2 d (\lambda_Q + \lambda_K)) \|\theta^{t+\phi} - \theta^t\|_2
\]

\[
= \rho (\lambda_V + \lambda_O + \lambda_V^2 C_2^2 d (\lambda_Q + \lambda_K)) \|\theta^{t+\phi} - \theta^t\|_2
\]

\[
\triangleq c_1 \|\theta^{t+\phi} - \theta^t\|_2,
\]

(36)

where the first inequality is by Lemma [10], the second inequality is by Lemma [9]. Next, the second term in Eq. (35) can be bounded by Lemma [13]. The third term in Eq. (35) can be bounded by Lemma [14]. As a result, Eq. (35) has the following upper bound:

\[
\|\nabla \ell(\theta^{t+\phi}) - \nabla \ell(\theta^T)\|_2 \leq c_1 c_2 \|\theta^{t+\phi} - \theta^T\|_2 + 2 c_3 \|\theta^{t+\phi} - \theta^T\|_2 \triangleq C \|\theta^{t+\phi} - \theta^T\|_2,
\]

(37)

where \(c_1, c_2, c_3\) are defined at Eqs. (11), (14) and (36), and we further define \(C \triangleq c_1 c_2 + 2 c_3 \ell(\theta^0)\).

Lastly, by applying Lemma 4.3 in Nguyen and Mondelli [2020] and Eq. (37), we have:

\[
\ell(\theta^{t+1}) \leq \ell(\theta^t) + \langle \nabla \ell(\theta^t), \theta^{t+1} - \theta^t \rangle + \frac{C}{2} \|\theta^{t+1} - \theta^T\|^2_F
\]

\[
= \ell(\theta^t) - \gamma \|\nabla \ell(\theta^t)\|_F^2 + \frac{C}{2} \gamma^2 \|\nabla \ell(\theta^t)\|_F^2
\]

\[
\leq \ell(\theta^t) - \frac{1}{2} \gamma \|\nabla \ell(\theta^t)\|_F^2\quad \text{(By the condition on } \gamma\text{)}
\]

\[
\leq \ell(\theta^t) - \frac{1}{2} \gamma \|\nabla w_c \ell(\theta^t)\|_2^2
\]

\[
= \ell(\theta^t) - \frac{1}{2} \gamma \|\nabla \ell(\theta^t)\|_F^2
\]

\[
\leq \ell(\theta^t) - \frac{1}{2} \gamma \|\nabla \ell(\theta^t)\|_F^2
\]

\[
\leq \ell(\theta^t) - \frac{1}{2} \gamma \|\nabla \ell(\theta^t)\|_F^2
\]

\[
\leq (1 - \frac{\gamma}{2}) \ell(\theta^t)\quad \text{(By induction assumption)}
\]

which concludes the proof.

\(\square\)

**C.3 Proof of Theorem**

**Lemma 15.** Let \(\Phi = [X_1^T \beta_{1,1}, ..., X_N^T \beta_{1,N}]^T \in \mathbb{R}^{N \times d}\), where \(\beta_{1,n} = \sigma_s (\tau_0 X_n^{(1:1)} W_Q X_n^T W_K X_n^T)\), \(n \in [N]\), then under Assumptions 2 and 3, with probability at
least $1 - \exp(-\Omega((N-1)^{-\epsilon}d_x^{-1}))$, one has: $\lambda_0 := \lambda_{\min} \left( \mathbb{E}_w \mathbb{E}_X \left[ \sigma_r(\Phi w)\sigma_r(\Phi w)^T \right] \right) \geq \Theta(\eta_V / d_s)$.

\textbf{Proof.} Due to Assumption 2 for any data $X$, the matrix $XX^\top$ is positive definite and thus has positive minimum eigenvalue. We denote it as $\lambda_{\min}(XX^\top) \geq C_\lambda$.

According to [Nguyen et al., 2021] Lemma 5.3, using the Hermite expansion of $\sigma_r$, one has:

$$\lambda_0 \geq \eta_V \mu(\sigma_r)^2 \lambda_{\min}(\Phi \Phi^\top),$$

where $\mu(\sigma_r)$ is the 1-st Hermite coefficient of ReLU satisfying $\mu(\sigma_r) > 0$.

Now we proceed to provide a lower bound for $\lambda_{\min}(\Phi \Phi^\top)$. For notational simplicity, define:

$$B_{ij} = \beta_{1,i} \beta_{1,j}^\top \in \mathbb{R}^{d_x \times d_x}, C_{ij} = X_i X_j^\top \in \mathbb{R}^{d_x \times d_x}.$$  

Then we can rewrite $\Phi \Phi^\top$ as follows:

$$\Phi \Phi^\top = \begin{bmatrix} \text{Trace}(B_{11}^\top C_{11}) & \text{Trace}(B_{11}^\top C_{12}) & \cdots & \text{Trace}(B_{1N}^\top C_{1N}) \\ \\
\text{Trace}(B_{21}^\top C_{21}) & \text{Trace}(B_{22}^\top C_{22}) & \cdots & \text{Trace}(B_{2N}^\top C_{2N}) \\ \\
\vdots & \vdots & \ddots & \vdots \\ \\
\text{Trace}(B_{N1}^\top C_{N1}) & \text{Trace}(B_{N1}^\top C_{N2}) & \cdots & \text{Trace}(B_{NN}^\top C_{NN}) \end{bmatrix}.$$  

By Gershgorin circle theorem [Gershgorin, 1931], there exists $k \in [N]$ such that:

$$\lambda_{\min}(\Phi \Phi^\top) \geq \text{Trace}(B_{kk} C_{kk}) - \sum_{j \neq k} \text{Trace}(B_{kj}^\top C_{kj}).$$

Using Von Neumann’s trace inequality [Mirsky, 1955] and noting that $B_{kj}$ is a rank one matrix, one has:

$$\text{Trace}(B_{kj} C_{kj}) \leq \sigma_{\max}(B_{kj}) \sigma_{\max}(C_{kj}) = \|\beta_{1,k}\|_2 \|\beta_{1,j}\|_2 \sqrt{\lambda_{\max}(C_{kj}^\top C_{kj})}$$

$$\leq \|\beta_{1,k}\|_2 \sqrt{\text{Trace}(C_{kj} C_{kj}^\top)} = \|\beta_{1,k}\|_2 \sqrt{\langle X_k^\top X_k, X_j^\top X_j \rangle},$$

where we use the definition of the inner product between two matrices and $\|\beta_{1,j}\|_2 \leq 1$. By Assumption 2 we have $\lambda_{\min}(X_k X_k^\top) \geq C_\lambda$, where $C_\lambda$ is some positive constant. By setting $t := \|\beta_{1,k}\|_2^2 C_\lambda^2 / (N-1)^2$ in Assumption 3 with probability at least $1 - \exp(-\Omega((N-1)^{-\epsilon}))$, one has:

$$\sqrt{\langle X_k^\top X_k, X_j^\top X_j \rangle} \leq \|\beta_{1,k}\|_2 C_\lambda (N-1)^{-1}, \quad \forall j \neq k.$$  

Plugging back Eq. (39) and Eq. (40), we obtain:

$$\lambda_{\min}(\Phi \Phi^\top) \geq \text{Trace}(B_{kk} C_{kk}) - C_\lambda \|\beta_{1,k}\|_2^2 \geq \lambda_{\min}(C_{kk}) \text{Trace}(B_{kk}) - C_\lambda \|\beta_{1,k}\|_2^2$$

$$= \lambda_{\min}(X_k X_k^\top) \|\beta_{1,k}\|_2^2 - C_\lambda \|\beta_{1,k}\|_2^2 \geq \Theta(\|\beta_{1,k}\|_2^2) \geq \Theta(1/d_s),$$

where the last inequality is by the lower bound of $\beta_{1,k}$ in Lemma 8. Lastly, plugging the lower bound of $\lambda_{\min}(\Phi \Phi^\top)$ back Eq. (38) finishes the proof.

\hfill \Box

\textbf{Remark:} The estimation of $\lambda_0$ is actually tight because its upper bound is also in a constant order. To be specific, denote $G := \sigma_r(\Phi w)\sigma_r(\Phi w)^\top$, we have

$$\lambda_0 := \lambda_{\min} \left( \mathbb{E}_w G \right) \leq \frac{\text{tr}(\mathbb{E}_w G)}{N} = \frac{\sum_{n=1}^N \mathbb{E}_w \left[ \sigma_r(\Phi^{(n,:)} w) \right]^2}{N},$$

where $\Phi^{(n,:)} = \beta_{1,n}^\top X_n$. Next, by Liao and Couillet [2018] (Sec.A in Supplementary Material):

$$\mathbb{E}_w [\sigma_r(\Phi^{(n,:)}) w]^2 = \frac{\eta_V}{2\pi} \left\| \Phi^{(n,:)} \right\|^2 \arccos(-1) = \frac{\left\| \Phi^{(n,:)} \right\|^2}{2}. $$
Combine Eq. (42) and Eq. (43), we have
\[ \lambda_0 \leq \frac{\sum_{n=1}^{N} \eta V \| \Phi^{(n,:)} \|_2^2}{2N} \leq \eta V d_s C_x^2 \leq O(1). \]

That means, our estimation on \( \lambda_0 \) is tight as its upper and lower bounds match with each other.

Now we are ready to present the proof for LeCun initialization under \( \tau_0 = d_m^{-1/2} \) scaling.

**Proof.** We select \( C_Q = C_K = 1 = C_V = C_O = 1 \), then by Lemma 5 with probability at least \( 1 - 8e^{-d_m/2} \), we have:
\[ \bar{\lambda}_V = O(\sqrt{d_m/d}), \quad \bar{\lambda}_O = O(1), \quad \bar{\lambda}_Q = O(\sqrt{d_m/d}). \]
(44)

Plugging Eq. (44) into Eqs. (2) and (3), it suffices to prove the following equations.
\[ \alpha^2 \geq O\left( \sqrt{N d_s^3/2} C_x \sqrt{d_m/d} \right) \sqrt{2\ell(\theta^0)} \]
(45)
\[ \alpha^3 \geq O\left( N d_s^3 C_x^2 (1 + 4C_x^2 d_s^2 d_m^{-2}) \right) \sqrt{2\ell(\theta^0)} \]
(46)

Next, we will provide the lower bound for \( \alpha^2 = \lambda_{\text{min}}( (\mathbf{F}_{\text{pre}}^0)^T (\mathbf{F}_{\text{pre}}^0) ) \). In the following context, we hide the index 0 for simplification. One can note that \( \mathbf{F}_{\text{pre}}^0 (\mathbf{F}_{\text{pre}}^0)^\top \) is the summation of PSD matrices, thus, it suffices to lower bound: \( \lambda_{\text{min}}( (\mathbf{F}_{\text{pre}} (\mathbf{F}_{\text{pre}})^\top ) \), where we introduce the following notations:
\[ \hat{\mathbf{F}}_{\text{pre}} (\hat{\mathbf{F}}_{\text{pre}})^\top \]
\[ \hat{\mathbf{F}}_{\text{pre}} = \sigma_\tau (\mathbf{\Phi} \mathbf{W}_v^\top ) \sigma_\tau (\mathbf{\Phi} \mathbf{W}_v^\top )^\top \\
\mathbf{\Phi} = [\mathbf{X}_1 \mathbf{\beta}_1, \ldots, \mathbf{X}_N \mathbf{\beta}_1] \\
\mathbf{\beta}_1 = \mathbf{\beta}_s \left( \tau_0 \mathbf{X}_n^{(1,:)}, \mathbf{W}_Q \mathbf{K} \mathbf{X}_n^\top \right)^\top n \in [N]. \]
(47)

By Matrix-Chernoff inequality, we can obtain that (e.g. Lemma 5.2 of [Nguyen et al. 2021]) w.p at least \( 1 - \delta_1 \),
\[ \lambda_{\text{min}}( (\mathbf{F}_{\text{pre}} (\mathbf{F}_{\text{pre}})^\top ) \geq \frac{d_m \lambda_0}{4}, \]
(48)
as long as it holds \( d_m \geq \Omega(\sqrt{N/\lambda_0}) \), where \( \lambda_0 = \lambda_{\text{min}}( \mathbb{E}_{w \sim \mathcal{N}(0, \eta v \mathbb{I})} (\sigma_\tau (\mathbf{\Phi} \mathbf{w}) \sigma_\tau (\mathbf{\Phi} \mathbf{w})^\top ) \), and \( \Omega \) hides logarithmic factors depending on \( \delta_1 \). Lastly, w.p. at least \( 1 - \delta_2 \), one has \( \sqrt{2\ell(\theta^0)} \leq \hat{O}(\sqrt{N}) \). Plugging back Eqs. (2) and (3), it suffices to prove the following inequality.
\[ d_m \lambda_0/4 \geq \hat{O}(N d_s^3 C_x \sqrt{d_m/d}) \]
(49)
\[ (d_m \lambda_0/4)^{3/2} \geq \hat{O}(N^{3/2} d_s^3 C_x^2 (1 + 4C_x^2 d_s^2 d_m^{-2})) \]
(50)

By Lemma 15 with probability at least \( 1 - \exp(-\Omega((N - 1)^{-\delta} d_s^{-1})) \), one has \( \lambda_0 \geq \Theta(\eta v / d_s) = \Theta(d_m^{-1} d_s^{-1}) \). Thus, when \( d_m \geq \hat{O}(N^3) \), all of the above conditions hold. As a result, the conditions in Eqs. (45) and (46) are satisfied and the convergence of training Transformer is guaranteed as in Eq. (4). Note that one can achieve the same width requirement and probability for He initialization, and the proof bears resemblance to the LeCun initialization.

For the proof under LeCun initialization and \( \tau_0 = d_m^{-1} \) scaling, we follow the same strategy. Specifically: As the same in the proof with \( \tau_0 = d_m^{-1/2} \) scaling, we select \( C_Q = C_K = C_V = C_O = 1 \), then plugging Eq. (44) into Eqs. (2) and (3), it suffices to prove the following equations.
\[ \alpha^2 \geq O\left( \sqrt{N d_s^3/2} C_x \sqrt{d_m/d} \right) \sqrt{2\ell(\theta^0)} \]
(51)
\[ \alpha^3 \geq O\left( N d_s^3 C_x^2 (1 + 4C_x^2 d_s^2 d_m^{-2}) \right) \sqrt{2\ell(\theta^0)} \]
(52)

Next, we will provide the lower bound for \( \alpha^2 = \lambda_{\text{min}}( (\mathbf{F}_{\text{pre}} (\mathbf{F}_{\text{pre}})^\top ) \). In the following context, we hide the index 0 for simplification. One can note that \( \mathbf{F}_{\text{pre}} (\mathbf{F}_{\text{pre}})^\top \) is the summation of PSD matrices,
thus, it suffices to lower bound: $\lambda_{\min}(\tilde{F}_{\text{pre}}\tilde{F}_{\text{pre}}^\top)$, where we introduce the following notations:

$$
\tilde{F}_{\text{pre}}\tilde{F}_{\text{pre}}^\top = \sum\sigma_r(\Phi W^r)\sigma_r(\Phi W^r)^\top,
$$

$$
\Phi = [X_1^\top \beta_{1,1}, ..., X_N^\top \beta_{1,N}]^\top,
$$

$$
\beta_{1,n} = \sigma_s(\sigma(\Phi X_n^\top W_{Q}^n W_{K} X_n^\top))^\top n \in [N].
$$

By Eq. (48) w.p at least $1 - \delta_1$, $\lambda_{\min}(\tilde{F}_{\text{pre}}\tilde{F}_{\text{pre}}^\top) \geq d_m \lambda_0 / 4$, as long as it holds $d_m \geq \tilde{\Omega}(N/\lambda_0)$, where $\lambda_0 = \lambda_{\min}(E_{w \sim N(0,\eta_v I)})[\sigma_r(\Phi w)\sigma_r(\Phi w)^\top]$, and $\tilde{\Omega}$ hides logarithmic factors depending on $\delta_1$. Lastly, note that the activation function in the output layer $\sigma_r$ is 1-Lipschitz and is applied to $\sigma_r(W_V X^\top \beta_1)$, where $X^\top \beta_1$ is bounded due to the softmax’s property in Lemma 8, then by Lemma C.1 of Nguyen and Mondelli [2020], w.p. at least $1 - \delta_3$, one has $\sqrt{2/(\theta_0)} \leq \tilde{\Omega}(\sqrt{N})$. Plugging back Eqs. (2) and (3), it suffices to prove the following inequality.

$$
d_m \lambda_0 / 4 \geq \tilde{\Omega}(N d_s^2/2C_x \sqrt{d_m / d}),
$$

$$
(d_m \lambda_0 / 4)^{3/2} \geq \tilde{\Omega}(N^{3/2} d_s^2 C_x^2 (1 + 4C_x^4 d_s^2 d^{-2})).
$$

By Lemma 15 with probability at least $1 - \exp(-\Omega((N - 1)^{-\epsilon} d_s^{-1}))$, one has $\lambda_0 \geq \Theta(\eta_v / d_s) = \Theta(d^{-1} d_s^{-1})$. Thus, when $d_m \geq \tilde{\Omega}(N^2)$, all of the above conditions hold. As a result, the conditions in Eqs. (43) and (46) are satisfied and the convergence of training Transformer is guaranteed as in Eq. (4). Note that one can achieve the same width requirement and probability for He initialization, and the proof bears resemblance to the LeCun initialization.

\[\square\]

C.4 Proof of Theorem 2 (NTK analysis)

Proof. Below, we present the proof for NTK initialization. We select $C_Q = C_K = C_V = C_O = 1$, then by Lemma 5 with probability at least $1 - 8e^{-d_m / 2}$, we have:

$$
\lambda_V = O(\sqrt{d_m + \sqrt{d}}), \quad \lambda_O = O(\sqrt{d_m}),
$$

$$
\lambda_Q = O(\sqrt{d_m + \sqrt{d}}), \quad \lambda_K = O(\sqrt{d_m + \sqrt{d}}).
$$

When $d_m \geq d$, plugging Eq. (56) into Eqs. (2) and (3), it suffices to prove the following equations.

$$
\alpha^2 \geq O(\sqrt{N} d_s^3/2C_x) \sqrt{2/(\theta_0)},
$$

$$
\alpha^3 \geq O(N d_s^3 C_x^2 (1 + 4C_x^4 d_s^2)) \sqrt{2/(\theta_0)}.
$$

Next, we will provide the lower bound for $\alpha^2 = \lambda_{\min}(\tilde{F}_{\text{pre}}\tilde{F}_{\text{pre}}^\top)$. In the following context, we hide the index 0 for simplification. One can note that $\tilde{F}_{\text{pre}}\tilde{F}_{\text{pre}}^\top$ is the summation of PSD matrices, thus, it suffices to lower bound: $\lambda_{\min}(\tilde{F}_{\text{pre}}\tilde{F}_{\text{pre}}^\top)$, where we introduce the following notations:

$$
\tilde{F}_{\text{pre}}\tilde{F}_{\text{pre}}^\top = \sum\sigma_r(\Phi W^r)\sigma_r(\Phi W^r)^\top,
$$

$$
\Phi = [X_1^\top \beta_{1,1}, ..., X_N^\top \beta_{1,N}]^\top,
$$

$$
\beta_{1,n} = \sigma_s(\sigma(\Phi X_n^\top W_{Q}^n W_{K} X_n^\top))^\top n \in [N].
$$

By Eq. (48), w.p at least $1 - \delta$, $\lambda_{\min}(\tilde{F}_{\text{pre}}\tilde{F}_{\text{pre}}^\top) \geq d_m \lambda_0 / 4$, as long as it holds $d_m \geq \tilde{\Omega}(N/\lambda_0)$, where $\lambda_0 = \lambda_{\min}(E_{w \sim N(0,\eta_v I)})[\sigma_r(\Phi w)\sigma_r(\Phi w)^\top]$, and $\tilde{\Omega}$ hides logarithmic factors depending on $\delta_1$. Lastly, w.p. at least $1 - \delta_3$, one has $\sqrt{2/(\theta_0)} \leq \tilde{\Omega}(\sqrt{N})$. Plugging back Eqs. (2) and (3), it suffices to prove the following inequality.

$$
d_m \lambda_0 / 4 \geq \tilde{\Omega}(N \tilde{\Omega}(d_s^2/2C_x^2)),
$$

$$
(d_m \lambda_0 / 4)^{3/2} \geq \tilde{\Omega}(N^{3/2} d_s^2 C_x^2 (1 + 4C_x^4 d_s^2 d^{-2})).
$$

By Lemma 15 with probability at least $1 - \exp(-\Omega((N - 1)^{-\epsilon} d_s^{-1}))$, we have $\lambda_0 \geq \Omega(1)$. Thus, all of the above conditions hold when $d_m = \tilde{\Omega}(N)$. As a result, the conditions in Eqs. (57) and (58) are satisfied and the convergence of training Transformer is guaranteed. \[\square\]
C.5 Discussion for different initialization schemes

Recall that the convergence result in Theorem 2 shows:

$$\ell(\theta^\top) \leq \left(1 - \gamma \frac{\alpha^2}{2}\right) \ell(\theta^0).$$

Thus, to discuss the convergence speed for different initialization, we need to check the lower bound for $\alpha^2$. From the proofs for different initialization schemes above, we have the following lower bound for $\alpha^2$, i.e.,

$$\alpha^2 = \lambda_{\min} (\tilde{F}_{\text{pre}} \tilde{F}_{\text{pre}}^\top) \geq \tau_1^2 d_m \lambda_0 / 4 \geq \tau_1^2 \eta_\nu \lambda d_m \Omega(N/d),$$

with high probability. Plugging the value of $\tau_1$ and $\eta_\nu$, we observe that for LeCun initialization and He initialization: $\alpha^2 \geq \Omega(d_m N/d)$ while for NTK initialization: $\alpha^2 \geq \Omega(N/d)$. Thus, the convergence speed of LeCun initialization and He initialization is faster than NTK initialization. As a result, faster step-size is required for NTK initialization.

C.6 Discussion for $\tau_0 = d_m^{-1}$ and $\tau_0 = d_m^{-1/2}$

Eq. (4) indicates that the convergence speed is affected by $\alpha$, where we compare the lower bound for $\alpha$ for these two scalings. In Appendix C.3, we have proved that under the LeCun initialization, one has $\alpha^2 \geq d_m \lambda_0 / 4 \geq d_m \eta_\nu \mu (\sigma_r) \beta_2 (\beta_{1,k})^2$. Note that this bound holds for these two different scalings, which is inside $\beta$. Thus in the next, we need to see the difference between the lower bound of $\beta_{1,k}$ in the case of these two scalings. Specifically, for $\tau_0 = d_m^{-1/2}$ scaling, we have proved that $\|\beta_{1,k}\|_2^2 \geq 1 / d_s$ by Lemma 8. However, for the case of $\tau_0 = d_m^{-1}$ scaling, when the width is large enough, the value inside the softmax tends to zero, as a result, $\|\beta_{1,k}\|_2^2 \approx 1 / d_s$. Thus, we can see that as the width increases, the convergence speed of $\tau_0 = d_m^{-1/2}$ could be faster. Lastly, we remark on the difference in the step size for these two scales, which can be seen from the definition of $C$ and its corresponding $c_1, c_2, c_3$ in Appendix C.2.

C.7 Discussion for extension to deep Transformer and residual Transformer

Extension from our shallow Transformer to deep Transformer is not technically difficult as they share the same analysis framework. Nevertheless, the extension requires several tedious derivations and calculations involving the query, key, and value matrices along different layers. Here we point out the proof roadmap for this extension. The first step is following Proposition 1 to provide the sufficient condition for the convergence guarantee, e.g., Eqs. 2 and 3. The second step is similar to Theorem 2 where we need to verify the aforementioned assumptions for different initialization. In the second part, the main task is to prove the lower bound of $\alpha := \sigma_{\min}(F_{\text{pre}}^0)$, where $F_{\text{pre}}$ is the output of the last hidden layer. One can apply concentration inequality to bound the difference between $\sigma_{\min}(F_{\text{pre}}^0)$ and $\sigma_{\min}(F_{\text{pre}}^{*0})$, where the latter is the corresponding limit in infinite width. Lastly, one needs to plug the lower bound into the assumptions in order to obtain the width requirement.

Our proof framework is general and can handle the following residual Transformer. Here we give a proof sketch to show how to achieve this. Specifically, we consider the residual block in the self-attention layer:

$$A_1 = \text{Self-attention}(X) \triangleq \sigma_a \left(\tau_0 (X W_Q^\top) (X W_K^\top)^\top \right) (X W_V^\top) + X.$$

As a result, the output becomes

$$f(X) = \tau_1 w^\top \sum_{i=1}^{d_s} \sigma_r \left(W_V X^\top \beta_i + (X^{(i)})^\top\right).$$

To prove the convergence of the above residual Transformer, the first part that will be modified is the proof for Proposition 1. The formula for $f_{\text{pre}}$ becomes as follows:

$$f_{\text{pre}} = \tau_1 w^\top \sum_{i=1}^{d_s} \sigma_r \left(W_V X^\top \beta_i + (X^{(i)})^\top\right).$$
where the second equality uses the law of large numbers. Secondly, we analyze

\[ W \]

where we denote by

\[ \Omega(X(i)) \]

This leads to the minimum eigenvalue in the order of

\[ \frac{1}{d} \]

In this section, we discuss the link between the linear over-parametrization and attention module behaving as a pooling layer.

In this section, we elaborate the proof for Lemma 1 in Appendix D.1, the proof for Theorem 3 in Appendix D.3 respectively.

**D.1 Proof of Lemma**

**Proof.** We will compute the inner product of the Jacobian of each weight separately. Firstly, we analyze \( w_O \). Let us denote by \( \beta_i := \sigma_s (\tau_0 X^{(i)}) W_Q W_K X^\top \top \in \mathbb{R}^{d_x} \). Then:

\[
\frac{\partial f(X)}{\partial w_O} = \tau_1 \sum_{i=1}^{d_x} \sigma_r \left( W_V X^\top \beta_i \right)
\]

The inner product of the gradient is:

\[
\lim_{d_m \to \infty} \left< \frac{\partial f(X)}{\partial w_O}, \frac{\partial f(X')}{\partial w_O} \right>
\]

\[
= \tau_1^2 \sum_{i=1}^{d_x} \lim_{d_m \to \infty} \left( \sigma_r \left( W_V X^\top \beta_i' \right) \right) \left( \sigma_r \left( W_V X^\top \beta_i \right) \right)
\]

\[
= \tau_1^2 \sum_{i=1}^{d_x} \left( \sigma_r \left( w^\top \beta_i' \right) \right) \left( \sigma_r \left( w^\top \beta_i \right) \right)
\]

where the second equality uses the law of large numbers. Secondly, we analyze \( W_Q \):

\[
\frac{\partial f(X)}{\partial W_Q^{(p,q)}} = \tau_0 \tau_1 \sum_{i=1}^{d_x} \left( w_O \circ \sigma_r \left( W_V X^\top \beta_i \right) \right) W_V X^\top \left( \sigma_r \left( \beta_i \right) \right) X W_K e_p e_q^\top X^{(i,:)}
\]

Lemmas 7 and 8 remain unchanged. In Lemma 9, only the first step in the proof changes while the remaining part does not change because the term \( X^{(i,:)} \) in two adjacent time steps cancels out. In Lemma 10, we have

\[
|| \tau_1 \sum_{i=1}^{d_x} \sigma_r \left( W_V X^\top \beta_i' \right) ||_2 \leq \tau_1 d_x \left( ||W_V^\top||_F d_x^{1/2} C_x + C_x \right)
\]

Similarly, in the remaining lemmas, we need to add the additional term \( C_x \) for the upper bound of \( ||X^{(i,:)}||_2 \).

The second part is the proof for Theorem 1 regarding the lower bound for \( \alpha_0 \). By Weyl’s inequality:

\[
\alpha_0 = \sigma_{\min} (F_{\text{pre}}) \geq \sigma_{\min} (F_{\text{pre}}^*) - ||F_{\text{pre}} - F_{\text{pre}}^*||_F
\]

where we denote by \( F_{\text{pre}} = [f_{\text{pre}}(X_1), \ldots, f_{\text{pre}}(X_N)] \), and \( F_{\text{pre}}^* \) is the corresponding one without the residual connection. Then we can upper bound the second term can be bounded as follows:

\[
\leq \sqrt{N} ||w_O^\top ||_2 \sum_{i=1}^{d_x} \sigma_r \left( W_V X^\top \beta_i \right) - \tau_1 d_x ||w_O||_2 ||W_V||_F C_x
\]

The remains step follows the same as previous analysis.

**C.8 Linear over-parametrization and attention module behaving as a pooling layer**

In this section, we discuss the link between the linear over-parametrization and attention module behaving as a pooling layer under the \( d_m^{-1} \) scaling. First, due to the \( d_m^{-1} \) scaling, the attention module degenerates to a pooling layer according to the law of large numbers. In this case, the nonlinearity on \( X \) disappears and thus the minimum eigenvalue of \( \Theta \Theta^\top \) can be estimated via \( XX^\top \). Accordingly, this leads to the minimum eigenvalue in the order of \( \Omega(N/d) \), and thus linear over-parametrization is enough.

**D Proof for NTK**

In this section, we elaborate the proof for Lemma 1 in Appendix D.1, the proof for Theorem 3 in Appendix D.3 respectively.
Thus:
\[
\frac{\partial f(X)}{\partial W_Q} = \tau_0 \tau_1 \sum_{i=1}^{d_x} W_k X^T \left( \text{diag}(\beta_i) - \beta_i \beta_i^T \right) X W_V^T \left( w_O \circ \sigma_r \left( W_V X^T \beta_i \right) \right) X^{(i,:)}. 
\]

The inner product of the gradient is:
\[
\lim_{d_m \to \infty} \left\langle \frac{\partial f(X)}{\partial W_Q}, \frac{\partial f(X')}{\partial W_Q} \right\rangle = \lim_{d_m \to \infty} \tau_0^2 \tau_1^2 \sum_{i=1,j=1}^{d_x} \text{Trace} \left( W_k X^T \left( \text{diag}(\beta_i) - \beta_i \beta_i^T \right) X W_V^T \left( w_O \circ \sigma_r \left( W_V X^T \beta_i \right) \right) \right) X^{(j,:)T} \left( W_V X^T \left( \text{diag}(\beta_j') - \beta_j' \beta_j'^T \right) X' W_k^T \right). 
\]

Thus:
\[
\tau_0 \tau_1 W^T_k W_k X^T \left( \text{diag}(\beta_j') - \beta_j' \beta_j'^T \right) X' W_V^T \left( w_O \circ \sigma_r \left( W_V X^T \beta_j' \right) \right).
\]

For the first term of the inner product, we have:
\[
\lim_{d_m \to \infty} \tau_0 \tau_1 X^T \left( \text{diag}(\beta_i) - \beta_i \beta_i^T \right) X W_V^T \left( w_O \circ \sigma_r \left( W_V X^T \beta_i \right) \right) = \lim_{d_m \to \infty} X^T \left( \text{diag}(1_{d_x}) \right) X = 0,
\]

where the second equality is by the law of large numbers and \( \mathbb{E} w = 0 \) for a random variable \( w \sim \mathcal{N}(0, 1) \). Thus: \( \lim_{d_m \to \infty} \left\langle \frac{\partial f(X)}{\partial W'_Q}, \frac{\partial f(X')}{\partial W'_Q} \right\rangle = 0 \). Similarly, \( \lim_{d_m \to \infty} \left\langle \frac{\partial f(X)}{\partial W''_Q}, \frac{\partial f(X')}{\partial W''_Q} \right\rangle = 0 \).

Lastly, we analyze \( W_V \):
\[
\frac{\partial f(X)}{\partial W_V} = \tau_1 \sum_{i=1}^{d_x} \left( w_O \circ \sigma_r \left( W_V X^T \beta_i \right) \right)^T e_p e_q^T X^T \beta_i.
\]

Thus:
\[
\frac{\partial f(X)}{\partial W_V} = \tau_1 \sum_{i=1}^{d_x} \left( w_O \circ \sigma_r \left( W_V X^T \beta_i \right) \right) \beta_i^T X.
\]

The inner product of the gradient is:
\[
\lim_{d_m \to \infty} \left\langle \frac{\partial f(X)}{\partial W_V}, \frac{\partial f(X')}{\partial W_V} \right\rangle = \lim_{d_m \to \infty} \tau_1^2 \sum_{i=1,j=1}^{d_x} \text{Trace} \left( \left( w_O \circ \sigma_r \left( W_V X^T \beta_i \right) \right)^T \beta_i^T X X'^T \beta_j' \left( w_O \circ \sigma_r \left( W_V X'^T \beta_j' \right) \right)^T \right)
\]
\[
= \lim_{d_m \to \infty} \tau_1^2 \sum_{i=1,j=1}^{d_x} \left( w_O \circ \sigma_r \left( W_V X'^T \beta_j' \right) \right)^T \left( w_O \circ \sigma_r \left( W_V X^T \beta_i \right) \right) \beta_i^T X X'^T \beta_j'
\]
\[
= \lim_{d_m \to \infty} \tau_1^2 \sum_{i=1,j=1}^{d_x} \sum_{k=1}^{d_k} \left( w_O(k) \right)^2 \sigma_r \left( W_V(k) X'^T \beta_j' \right) \sigma_r \left( W_V(k) X^T \beta_i \right) \beta_i^T X X'^T \beta_j'
\]
\[
= d_x^2 \mathbb{E} w \sim \mathcal{N}(0, 1) \left( \sigma_r \left( w^T \tau_{1} X^{(1,:)} \right) \right) \left( \sigma_r \left( w^T \tau_{1} X^{(1,:)} \right) \right) \left( 1_{d_x}^T X X'^T 1_{d_x} \right),
\]

where the last equality uses the law of large numbers.

\[\square\]
D.2 NTK minimum eigenvalue

**Lemma 16.** Given $\Phi$ defined in Eq. (53) and $\Phi^*$ defined in Lemma 1, denote $\lambda_s = \lambda_{\min}(\Phi^* \Phi^*)$, and suppose the width satisfies $d_m = \Omega\left(N^2 \sqrt{\log(2d^2 N^2/\delta)}\right)$, then with probability at least $1 - \delta$, one has $\|\Phi \Phi^T - \Phi^* \Phi^*\|_F \leq \frac{\lambda_s}{4}$ and $\lambda_{\min}(\Phi^* \Phi^*) \geq \frac{3\lambda_s}{4}$.

**Proof.** In the following content, the variable with * indicates the corresponding variable with infinite width $d_m$. According to the definition of $\Phi$ in Eq. (53) and $\Phi^*$ in Lemma 1 for each entry in $\Phi \Phi^T$ and $\Phi^* \Phi^*$, we have

$$
(\|\Phi \Phi^T - \Phi^* \Phi^*\|_F)^2 = \|\Phi \Phi^T\|_F^2 - \|\Phi^* \Phi^*\|_F^2
$$

Substituting back to Eq. (65), the following inequality holds with the same width requirement and probability at least $\delta$:

$$
\|\beta_{1,n} - \beta_{1,n}^*\|_2^2 \leq 2C^2_xd_2\|\tau_0 \Phi W_Q^T \Phi W_K - \tau_0 \Phi^* W_Q^T \Phi^* W_K\|_F^2
$$

Let first consider the absolute value of each element of $\tau_0 W_Q^T \Phi W_K - \tau_0 \Phi^* W_Q^T \Phi^* W_K$, i.e.,

$$
|\tau_0 W_Q^T \Phi W_K - \tau_0 \Phi^* W_Q^T \Phi^* W_K| = \|\tau_0 \sum_{q=1}^{d_m} W_q^{(q,i)} W^{(q,j)}_k - \tau_0 \sum_{q=1}^{d_m} W^{(q,i)}_q W^{(q,j)}_k\|.
$$

By Bernstein’s inequality, when $d_m \geq 2\log(2/\delta)$, the following inequality holds with probability at least $1 - \delta$:

$$
\|\tau_0 \sum_{q=1}^{d_m} W_q^{(q,i)} W^{(q,j)}_k - \tau_0 \sum_{q=1}^{d_m} W^{(q,i)}_q W^{(q,j)}_k\| \leq 2\tau_0 \eta Q_k \sqrt{d_m \log(2/\delta)}.
$$

Substituting $\delta = \frac{\delta'}{d_m}$ and applying union bound, we can obtain that when $d_m \geq 2\log \frac{2d^2}{\delta}$, with probability at least $1 - \delta'$:

$$
\|\beta_{1,n} - \beta_{1,n}^*\|_2 \leq 4\tau_0 \eta Q_k C^2_xd_2d\sqrt{d_m \log(2d^2/\delta')}.
$$

Substituting back to Eq. (65), the following inequality holds with the same width requirement and probability. $|(\Phi \Phi^T - \Phi^* \Phi^*)^{(n,r)}| \leq 8\tau_0 \eta Q_k C^4_xd_2d\sqrt{d_m \log(2d^2/\delta')}$ Applying the union bound over the index $(n, r)$ for $n \in [N]$ and $r \in [N]$, and substituting $\delta' = \frac{\delta''}{N^2}$, we obtain that when the width $d_m \geq 2\log \frac{2d^2 N^2}{\delta''}$, the following inequality holds with probability at least $1 - \delta''$:

$$
\|\Phi \Phi^T - \Phi^* \Phi^*\|_F^2 \leq \frac{N^2 \sqrt{\log(2d^2 N^2/\delta'')}}{\lambda_0}.
$$

In the case of LeCun initialization when $d_m = \Omega\left(N^2 \sqrt{\log(2d^2 N^2/\delta'')}\right)$, $\|\Phi \Phi^T - \Phi^* \Phi^*\|_F \leq \frac{\lambda_s}{4}$. Lastly, one has:

$$
\lambda_{\min}(\Phi^* \Phi^*) \geq \lambda_s - \|\Phi \Phi^T - \Phi^* \Phi^*\|_2 \geq \lambda_s - \|\Phi \Phi^T - \Phi^* \Phi^*\|_F \geq \frac{3\lambda_s}{4}.
$$

where the first inequality is by Weyl’s inequality. One can easily check that the same result also holds for He initialization and NTK initialization.

**Lemma 17.** Given the $\Phi^*$ defined in Lemma 1 then when $N \geq \Omega(d^4)$, with probability at least $1 - e^{-d}$ one has: $\lambda_{\min}(\Phi^* \Phi^*) \geq \Theta(N/d)$.
Proof. Firstly, let us define \( Z^* = \Phi^* \Sigma^{-1/2} \), then:

\[
\lambda_{\min}(\Phi^* \Phi^*^T) = \lambda_{\min}(\Phi^*^T \Phi^*) = \lambda_{\min}(\Sigma^{1/2} Z^* Z^* \Sigma^{1/2^T}) \geq \lambda_{\min}(Z^* \Sigma^*) \lambda_{\min}(\Sigma) \geq \lambda_{\min}(Z^* \Sigma^*) C_{\Sigma},
\]

where the last inequality is by Assumption \[4\). Note that we can reformulate \( Z^* \) by plugging \( \Phi^* \) as follows:

\[
Z^* = \Phi^* \Sigma^{-1/2} = \begin{bmatrix}
\frac{1}{d_x} \sum_{i=1}^{d_x} X_i^{(i:2)} \Sigma^{-1/2} \\
\vdots \\
\frac{1}{d_x} \sum_{i=1}^{d_x} X_N^{(i:2)} \Sigma^{-1/2}
\end{bmatrix}.
\]

Recall that in Assumption \[4\), we define the random vector \( x = X_i^{(i:2)^T} \) and write the covariance matrix for each feature of \( x \) as \( \Sigma = \mathbb{E}[xx^T] \). Here we further define the random vector \( z = \Sigma^{-1/2} x \). Clearly, the covariance matrix for each feature of \( z \) is \( \Sigma_z = \mathbb{E}[zz^T] = I_d \). Therefore, by Proposition 4 of [Zhu et al. 2022], we have \( \lambda_{\min}(Z^* \Sigma^*) \geq \Theta(N/d) \), which finishes the proof.

\[ \square \]

**Lemma 18.** Suppose the number of samples \( N \geq \Omega(d^4) \) and the width \( d_m = \tilde{\Omega}(N^2 \lambda_s^{-2}) \), where \( \lambda_s = \lambda_{\min}(\Phi^* \Phi^*^T) \). Then under Assumption \[4\) with probability at least \( 1 - \delta - e^{-d} \), one has

\[
\lambda_0 = \lambda_{\min}(\mathbb{E}_{w \sim N(0, \Sigma)}[\sigma_r(\Phi^* w) \sigma_r(\Phi^* w)^T]) \geq \Theta(\eta_N N/d).
\]

**Proof.** By the Hermite expansion of \( \sigma_r \), one has:

\[
\lambda_0 \geq \eta_N \mu(\sigma_r)^2 \lambda_{\min}(\Phi^* \Phi^T),
\]

where \( \mu(\sigma_r) \) is the 1-st Hermite coefficient of ReLU satisfying \( \mu(\sigma_r) > 0 \). By Lemma \[16\), when the width satisfies \( d_m = \Omega(N^2 \lambda_s^{-2}) \), then with probability at least \( 1 - \delta \), one has \( \lambda_{\min}(\Phi^* \Phi^T) \geq \frac{3}{4} \lambda_s \). Furthermore, by Lemma \[17\) when \( N \geq \Omega(d^4) \) w.p. at least \( 1 - e^{-d} \) one has \( \lambda_s = \Theta(N/d) \). Thus, with probability at least \( 1 - \delta - e^{-d} \), one has:

\[
\lambda_0 \geq \frac{3}{4} \eta_N \mu(\sigma_r)^2 \lambda_s \geq \Theta(\eta_N N/d).
\]

\[ \square \]

Below, we provide a lower bound for the minimum eigenvalue of NTK, which plays a key role in analyzing convergence, generalization bounds, and memorization capacity [Arora et al. 2019a, Nguyen et al. 2021, Liang and Rakhlin 2020, Vershynin 2018, Hastie et al. 2022].

To prove this, we need the following assumption.

**Assumption 4.** Let \( x = \sum_{i=1}^{d_x} (X_i^{(i:2)})^T, \Sigma = \mathbb{E}[xx^T] \), then we assume that \( \Sigma \) is positive definite, i.e., \( \lambda_{\min}(\Sigma) \geq C_{\Sigma} \) for some positive constant \( C_{\Sigma} \).

**Remark:** This assumption implies that the covariance matrix along each feature dimension is positive definite. In statistics and machine learning [Liu et al. 2021, Liang and Rakhlin 2020, Vershynin 2018, Hastie et al. 2022], the covariance of \( x \) is frequently assumed to be an identity matrix or positive definite matrix.

**Lemma 19 (Minimum eigenvalue of limiting NTK).** Under Assumptions \[7\) and \[8\) and scaling \( r_0 = d_m^{-1} \), when \( N \geq \Omega(d^4) \), the minimum eigenvalue of \( K \) can be lower bounded with probability at least \( 1 - e^{-d} \) as:

\[
\lambda_{\min}(K) \geq \mu(\sigma_r)^2 \Omega(N/d).
\]

where \( \mu(\sigma_r) \) represents the first Hermite coefficient of the ReLU activation function.

**Proof.** By the Hermite expansion of \( \sigma_r \), we have:

\[
\lambda_{\min}(K) > \lambda_{\min}(\mathbb{E}_{w \sim N(0, I)}[\sigma_r(\Phi^* w) \sigma_r(\Phi^* w)^T])
\]

\[
= \lambda_{\min}\left( \sum_{s=0}^{\infty} \mu_s(\sigma_r)^2 \Omega_{i=1}^{s} (\Phi^* \Phi^T) \right) \quad \text{[Nguyen and Mondelli 2020, Lemma D.3]}
\]

\[
\geq \mu(\sigma_r)^2 \lambda_{\min}(\Phi^* \Phi^T).
\]

35
Note that when \( N \geq \Omega(d^4) \), based on Lemma [17] and the fact that \( \Phi^* \Phi^* \) and \( \Phi^* \Phi^\top \) share the same non-zero eigenvalues, with probability at least \( 1 - e^{-d} \) one has:

\[
\lambda_{\text{min}}(\mathbf{K}) \geq \mu(\sigma_r)^2 \left( \frac{N}{d} - 9N^{2/3}d^{1/3} \right) = \mu(\sigma_r)^2 \Theta(N/d),
\]

which completes the proof. \( \Box \)

**Remark:** The minimum eigenvalue of the NTK plays an important role in the global convergence, similar to \( \alpha \) in Proposition [1]. By defining \( \alpha \triangleq \sigma_{\text{min}}(\mathbf{F}_0^{\text{pre}}) \), under the over-parameterized regime with \( d_m \geq N \), we have \( \alpha^2 = \lambda_{\text{min}}((\mathbf{F}_0^{\text{pre}})^\top \mathbf{F}_0^{\text{pre}}) \), which is the exact minimum eigenvalue of the empirical NTK with respect to the weight in the output layer.

### D.3 Proof of Theorem 3

**Proof.** Before starting the proof, we introduce some useful lemmas that are used to analyze the randomness of initialized weight.

**Lemma 20.** Given an initial weight vector \( \mathbf{w} \in \mathbb{R}^{d_m} \) where each element is sampled independently from \( \mathcal{N}(0, 1) \), for any \( \hat{\mathbf{w}} \) such that \( \|\hat{\mathbf{w}} - \mathbf{w}\|_2 \leq R \), with probability at least \( 1 - 2 \exp(-d_m/2) \), one has:

\[
\|\hat{\mathbf{w}}\|_2 \leq R + 3\sqrt{d_m}.
\]

**Proof of Lemma 20** By triangle inequality and Lemma [6]

\[
\|\hat{\mathbf{w}}\|_2 \leq \|\hat{\mathbf{w}} - \mathbf{w}\|_2 + \|\mathbf{w}\|_2 \leq R + 3\sqrt{d_m}.
\]

Now we are ready to start our proof, in the following content, we avoid the tilde symbol (\( \tilde{\cdot} \)) over the weight for simplicity. Because we aim to study the effect of width \( (d_m) \), to simplify the proof, we set the embedding dimension of the input as one, we can write the output layer of the network into the following form:

\[
f(\mathbf{x}) = \tau_0 \sum_{i=1}^{d_s} \sigma_r \left( \sigma_s \left( \tau_1 x^{(i)} w_O^T x \right) \right) w_O^T \mathbf{x}. \tag{71}
\]

The Hessian matrix \( \mathbf{H} \) of the network can be written as the following structure:

\[
\mathbf{H} = \begin{pmatrix}
\mathbf{H}_{Q,Q} & \mathbf{H}_{Q,K} & \mathbf{H}_{Q,V} & \mathbf{H}_{Q,O} \\
\mathbf{H}_{K,Q} & \mathbf{H}_{K,K} & \mathbf{H}_{K,V} & \mathbf{H}_{K,O} \\
\mathbf{H}_{V,Q} & \mathbf{H}_{V,K} & \mathbf{H}_{V,V} & \mathbf{H}_{V,O} \\
\mathbf{H}_{O,Q} & \mathbf{H}_{O,K} & \mathbf{H}_{O,V} & \mathbf{H}_{O,O}
\end{pmatrix},
\]

where we partition the Hessian \( \mathbf{H} \) of the network to each Hessian block e.g., \( \mathbf{H}_{Q,Q} = \frac{\partial^2 J}{\partial w_Q^2} \), \( \mathbf{H}_{Q,K} = \frac{\partial^2 J}{\partial w_Q \partial w_K} \), etc. Based on the triangle inequality of the norm, the spectral norm of the Hessian \( \mathbf{H} \) can be upper bounded by the sum of the spectral norm of each Hessian block.

We start by analyzing \( \mathbf{H}_{Q,Q} \). The time step \( t \) is hidden for simplification.

\[
\frac{\partial f(\mathbf{x})}{\partial w_Q^{(j)}} = \tau_0 \tau_1 \sum_{i=1}^{d_s} x^{(i)} w_K^{(j)} w_V^T (w_O \circ \sigma_r (w_V x^T \beta_1)) x^T (\text{diag}(\beta_1) - \beta_1 \beta_1^T) \mathbf{x}, \tag{73}
\]

where the Jacobian of Softmax is given by Lemma [11]. Next, we calculate the second-order derivative. Each element of the Hessian \( \mathbf{H}_{Q,Q} \) is

\[
H_{Q,Q}^{(j,p)} = \frac{\partial^2 f(\mathbf{x})}{\partial w_Q^{(j)} \partial w_Q^{(p)}} = \tau_0 \tau_1 \sum_{i=1}^{d_s} w_K^{(j)} w_K^{(p)} (x^{(i)})^2 w_V^T (w_O \circ \sigma_r (w_V x^T \beta_1)) x^T \Gamma_i \mathbf{x},
\]

36
where we denote by \( \Gamma_i = \left\{ \text{diag} \left( \beta_i \circ x - \beta_i \beta_i^\top x \right) - \text{diag} \left( \beta_i \right) \right\} \beta_i^\top - \beta_i x^\top \left( \text{diag} \left( \beta_i \right) - \beta_i \beta_i^\top \right) x, \beta_i \right\} \), \( \circ \) symbolizes Hadamard product. According to initialization of \( w_O, w_V, w_Q, \) and Lemma 20, with probability at least \( 1 - 8e^{-d_m/2} \), we have

\[
\|w_V\|_2 \leq 3\sqrt{\eta_V d_m} + R, \quad \|w_O\|_2 \leq 3\sqrt{\eta_Q d_m} + R, \quad \|w_Q\|_2 \leq 3\sqrt{\eta_Q d_m} + R.
\]

The spectral norm of \( H_{Q,Q} \) can be bounded with probability at least \( 1 - 8e^{-d_m/2} \):

\[
\|H_{Q,Q}\|_2 = \|\tau_0^2 \sigma_1 \sum_{i=1}^{d_z} (x^{(i)})^2 w_V^\top \left( w_O \circ \sigma_r \left( w_V x^\top \beta_i \right) \right) x^\top \Gamma_i x w_K^\top \|_2
\]

\[
\leq \|\tau_0^2 \sigma_1 \sum_{i=1}^{d_z} (x^{(i)})^2 w_V^\top \left( w_O \circ \sigma_r \left( w_V x^\top \beta_i \right) \right) x^\top \Gamma_i x \| \|w_K w_K^\top\|_2 \quad (\text{Homogeneity of norm})
\]

\[
\leq \|\tau_0^2 \sigma_1 \sum_{i=1}^{d_z} (x^{(i)})^2 x^\top \Gamma_i x \| \|w_O\|_2 \|w_V\|_2 \|w_K\|_2^2 \quad (\text{Cauchy–Schwarz inequality})
\]

\[
\leq \tau_0^2 \sigma_1 \sum_{i=1}^{d_z} (x^{(i)})^2 \Gamma_i x \left( 3\sqrt{\eta_V d_m} + R \right) \left( 3\sqrt{\eta_V d_m} + R \right) \left( 3\sqrt{\eta_Q d_m} + R \right) \left( 3\sqrt{\eta_K d_m} + R \right) \quad (\text{Lemma 20})
\]

\[
= O\left(1/\sqrt{d_m}\right),
\]

where the last equality is by the bound of \( \|\Gamma_i\|_2 \), by triangle inequality:

\[
\|\Gamma_i\|_2 \leq \|\text{diag} \left( \beta_i \circ x - \beta_i \beta_i^\top x \right)\|_2 + \|\text{diag} \left( \beta_i \right) \right\} \|x\|_2 + \|\beta_i^\top \text{diag} \left( \beta_i \right) \right\} \|x\|_2.
\]

For the first part of Eq. (74), we have

\[
\|\text{diag} \left( \beta_i \circ x - \beta_i \beta_i^\top x \right)\|_2 \leq \|\text{diag} \left( \beta_i \circ x - \beta_i \beta_i^\top x \right)\|_2 \leq \|\beta_i \|_2 \|x\|_2 + \|\beta_i \|_2 \|x\|_2 \leq 2C_x.
\]

For the second part of Eq. (74), we have

\[
\|\beta_i \|_2 \leq \|\beta_i \|_2 \|x\|_2 \|\beta_i \|_2 \leq 2C_x.
\]

For the third part of Eq. (74), we have

\[
\|\beta_i^\top (\text{diag} \left( \beta_i \right) \right\} \|x\|_2 \leq \|\beta_i^\top \|_2 \|x\|_2 \|\beta_i^\top \|_2 \leq 2C_x.
\]

Next, we analyze the Hessian \( H_{Q,K} \), where each element is:

\[
\frac{\partial^2 f}{\partial w_Q^2} = \left[ \frac{\partial^2 f}{\partial w_Q^2} \right] = \tau_0^2 \sigma_1 \sum_{i=1}^{d_z} w^{(i)} w^{(p)} \left( x^{(i)} \right)^2 \left( w_Q \circ \sigma_r \left( w_V x^\top \beta_i \right) \right) x^\top \Gamma_i x.
\]

Due to the symmetry, similar to \( H_{Q,Q} \), the spectral norm of \( H_{Q,K} \) can be bounded with probability at least \( 1 - 8e^{-d_m/2} \); \( \|H_{Q,K}\|_2 = O\left(1/\sqrt{d_m}\right) \).

Next, we analyze the Hessian \( H_{Q,V} \), where each element is:

\[
\frac{\partial^2 f}{\partial w_V^2} = \left[ \frac{\partial^2 f}{\partial w_V^2} \right] = \tau_0^2 \sigma_1 \sum_{i=1}^{d_z} w^{(i)} w^{(p)} \sigma_r \left( w_V x^\top \beta_i \right) x^\top \left( \text{diag} \left( \beta_i \right) \right) x.
\]
The spectral norm of $H_{Q,V}$ can be bounded with probability at least $1 - 8e^{-d_m/2}$:

$$\|H_{Q,V}\|_2 = \|\tau_1 \sum_{i=1}^{d_s} x^{(i)} \sigma_r \left( w_v^{(p)} x^\top \beta_i \right) x^\top (\text{diag}(\beta_i) - \beta_i \beta_i^\top) x w_K w_O^\top \|_2$$

$$\leq |\tau_1 \sum_{i=1}^{d_s} x^{(i)} \sigma_r \left( w_v^{(p)} x^\top \beta_i \right) x^\top (\text{diag}(\beta_i) - \beta_i \beta_i^\top) x ||w_K w_O^\top ||_2 \quad \text{(Homogeneity of norm)}$$

$$\leq |\tau_1 \sum_{i=1}^{d_s} |x_i| ||\text{diag}(\beta_i) - \beta_i \beta_i^\top||_2 ||x||_2 ||w_K||_2 ||w_O||_2 \quad \text{(Cauchy–Schwarz inequality)}$$

$$\leq |\tau_1 \sum_{i=1}^{d_s} |x_i| ||\text{diag}(\beta_i) - \beta_i \beta_i^\top||_2 ||x||_2 \left(3\sqrt{\eta_{QK} d_m} + R\right) \left(3\sqrt{\eta_{QV} d_m} + R\right) \quad \text{(Lemma 20)}$$

$$= O(1/d_m),$$

where the last step is by Weyl’s inequality and the range of the output of Softmax (from zero to one):

$$\left(\min_j (\beta_i^{(j)}) + ||\beta_i||^2\right)^2 \leq ||\text{diag}(\beta_i) - \beta_i \beta_i^\top||_2 \leq \left(\max_j (\beta_i^{(j)}) + ||\beta_i||^2\right)^2 = O(1).$$

It suffices to bound $||\beta_i||_2$, i.e., $\|\sigma_s (x^{(i)} x w_v^\top w_K) \|_2$. Since the range of each element of the output of Softmax is from 0 to 1 and the sum of is one, $||\beta_i|| \leq 1$.

Next, we analyze the Hessian $H_{Q,O}$, where each element is:

$$H_{Q,O}^{(j,p)} = \frac{\partial^2 f}{\partial u_Q^{(j)} \partial u_O^{(p)}} = \tau_1 \sum_{i=1}^{d_s} x^{(i)} \sigma_r \left( w_v^{(p)} x^\top \beta_i \right) x^\top (\text{diag}(\beta_i) - \beta_i \beta_i^\top) x.$$

The spectral norm of $H_{Q,V}$ can be bounded with probability at least $1 - 8e^{-d_m/2}$:

$$\|H_{Q,V}\|_2 = \|\tau_1 \sum_{i=1}^{d_s} x^{(i)} \sigma_r \left( w_v^{(p)} x^\top \beta_i \right) x^\top (\text{diag}(\beta_i) - \beta_i \beta_i^\top) x w_K w_V^\top \|_2$$

$$\leq |\tau_1 \sum_{i=1}^{d_s} x^{(i)} \sigma_r \left( w_v^{(p)} x^\top \beta_i \right) x^\top (\text{diag}(\beta_i) - \beta_i \beta_i^\top) x ||w_K w_V^\top ||_2 \quad \text{(Homogeneity of norm)}$$

$$\leq |\tau_1 \sum_{i=1}^{d_s} |x_i| ||\text{diag}(\beta_i) - \beta_i \beta_i^\top||_2 ||x||_2 ||w_K||_2 ||w_V||_2 \quad \text{(Cauchy–Schwarz inequality)}$$

$$\leq |\tau_1 \sum_{i=1}^{d_s} |x_i| ||\text{diag}(\beta_i) - \beta_i \beta_i^\top||_2 ||x||_2 \left(3\sqrt{\eta_{QK} d_m} + R\right) \left(3\sqrt{\eta_{QV} d_m} + R\right) \quad \text{(Lemma 20)}$$

$$= O(1/d_m).$$

Next, we analyze the Hessian $H_{O,V}$. Each element of $H_{O,V}$ is:

$$H_{O,V}^{(p,j)} = \frac{\partial^2 f}{\partial u_O^{(p)} \partial u_V^{(j)}} = \tau_1 \sum_{i=1}^{d_s} x^{(i)} \sigma_r \left( w_v^{(p)} \beta_i^\top x \right) x^\top \beta_i \mathbb{1}_{\{j=p\}}.$$

Note that $H_{O,V}$ is a diagonal matrix. Consequently, the spectral norm of $H_{O,V}$ is:

$$\|H_{O,V}\|_2 = \max_{j \in [d_m]} |H_{O,V}^{(j,j)}| \leq |\tau_1 \sum_{i=1}^{d_s} x^\top \beta_i| \leq |\tau_1 \sum_{i=1}^{d_s} ||x||_2 ||\beta_i||_2 \| = O(1/d_m).$$

Next, since each element of $H_{O,O}$ and $H_{V,V}$ is zero, the corresponding spectral norm is zero. Lastly, due to the symmetry of $w_K$ and $w_Q$, we can obtain the same bound for the corresponding Hessian block. Thus, the spectral norm of Hessian $H$ is upper bounded by $O(1/\sqrt{d_m})$, which completes the proof.
Figure 5: Comparison of convergence curve for different initialization schemes. The convergence speed under Lecun/He initialization is generally faster than that of NTK initialization with the same step size.

E Additional details on experiments

E.1 Additional result in Section 5.1

In this section, we present the convergence curve for different initialization schemes following the setup in Section 5.1. The results in Figures 5a to 5c show that when using the same step size \( \gamma = 1 \) for LeCun and He initialization exhibits similar behavior while NTK initialization leads to slower convergence. This is consistent with our theoretical finding. We also depict the result with larger step size \( \gamma = 10 \) for NTK initialization in Figure 5d, following our analysis in Appendix C.5.

E.2 Additional result and set-up in Section 5.2

We have validated Assumption 3 on language data in the main paper. Here we will validate it on image dataset. Specifically, we choose MNIST dataset and consider the embedding in the ViT as \( X \), then we plot \( P (|\langle X_n, X_n', X_n', X_n \rangle| \geq t) \) as \( t \) increases in Figure 6, where we could see exponential decay. The architecture of ViT is the same as in Section 5.2. Additionally, in Figure 7, we can see that \( d_m^{-1/2} \) setting achieves faster convergence in training loss and higher test accuracy. In Figure 8, we conduct the experiment on MNIST with regression task by using the same architecture. We take two classes \( \{0, 1\} \) in MNIST for binary classification, the output of our regression is taken by \( \text{sign}(f) \), and the training error is given by \( \sum_{n=1}^{N} |\text{sign}(f(X_n)) - y_n|^2 \). We can see that the convergence speed of \( d_m^{-1/2} \) is faster than that of \( d_m^{-1} \). Overall, these separation results on both classification and regression tasks provide good justification for the \( d_m^{-1/2} \) scaling.

Figure 6: Verification of Assumption 3 on MNIST dataset.
Figure 7: Test accuracy of classification task on MNIST dataset.

Figure 8: Regression task (with MSE loss) on MNIST dataset.

Figure 9: Experimental validation of the theoretical results on Transformers with $\tau_0 = d_m^{-1}$ scaling trained on synthetic data. (a) Linear convergence. (b) Rate of change of the weights during training. Observe that the weights change very slowly after the 50th epoch. (c) Evolution of the NTK during the training. The result mirrors the plot (b) and demonstrates how the kernel varies significantly at the beginning of the training and remains approximately constant later. As the width increases, the NTK becomes more stable.

F Limitation

Firstly, in this work, we prove the global convergence guarantee for Transformer under different initialization schemes. Nevertheless, we only consider the gradient descent training under squared loss.
Our result does not cover SGD training and other loss functions, e.g., hinge loss and cross-entropy loss. Secondly, the model architecture being analyzed in this work is the encoder of Transformer which is widely used for regression problems or image classification problems. We do not consider the entire Transformer including the decoder, which is designed for sequence-to-sequence modeling. Lastly, our model does not cover the deep Transformer. We believe our analytic framework paves a way for further analyzing large-scale Transformer.

G Societal impact

This work offers a convergence analysis for the Transformer, which is a core element within contemporary large language models. Our theoretical analysis with realistic initializations and scaling schemes lays a theoretical foundation for the interested practitioner in the ML community to further study other priorities of Transformer such as generalization and in-context learning. We do not expect any negative societal bias from this work.