The Implicit Bias of Scale Factor in Attention Layer

Shuailong Zhu

Abstract

The attention structure is essential for the success of large language models nowadays. Many works are devoted to a better understanding of attention. In this report, we focus on the implicit bias of the scaling factor in attention. Specifically, we analyze the influence of scaling on the distance of training dynamics bewteen two parameterizations, KQ-param and W-param, and search for the scaling scheme that will lead to a "correct" linearization for KQ-param, through the ideas and tools from lazy training.

1 Introduction

Transformers have demonstrated superior success in influential applications, especially in the area of large language models[Bro+20; Ope+23; Tou+23]. A fundamental theoretical understanding of the implicit bias of the transformer can better facilitate the utilization and design of different transformer structures. In this report, we will study the simplified attention model as [Tar+23]. Given input sequences $X, Z \in \mathbb{R}^{T\times d}$ with length T and embedding dimension d, the cross attention model is formulated as

$$f_{cross}(X,Z) := V^T X^T \mathbb{S}(\tau X K Q^T Z^T) \tag{1}$$

where $K, Q \in \mathbb{R}^{d\times m}, V \in \mathbb{R}^{d\times v}$, τ is the scaling factor in softmax, which is originally chosen as $\dfrac{1}{\sqrt{m}}$ in [Vas+17]. And $\mathbb{S}(\cdot)$ denotes the softmax nonlinearity, which is applied column-wise.

To focus on attention mechanism, [Tar+23] assumes (1) $z \in \mathbb{R}^d$; (2) feed-forward network is linear and fixed; (3) V fixed. Thus, we can define the single-layer transformer

$$f_{K,Q,\tau}(X,z) = h(X^T \mathbb{S}(\tau X K Q^T z)) = v^T X^T \mathbb{S}(\tau X K Q^T z) \tag{2}$$

where $h : \mathbb{R}^d \to \mathbb{R}$ is a fixed linear function that can be written as $h(\cdot) = \langle v, \cdot\rangle$ with $v \in \mathbb{R}^d$. If z is a token in sequence X, then f could also represent the self-attention layer. The author also propose another parameterization to better understand the implicit bias of this simple model,

$$f_{W,\tau}(X,z) = h(X^T \mathbb{S}(\tau X W z)) = v^T X^T \mathbb{S}(\tau X W z) \tag{3}$$

Consider a classification task with dataset $(Y_i, X_i, z_i)_{i=1}^n$ where $Y_i \in \{+1, -1\}$, the goal is to compare ERM problem using two different parameterization under the separable setting:

$$L(W) = \frac{1}{n}\sum_{i=1}^n l(f_{W,\tau}(X,z), Y_i) \tag{W-ERM}$$

$$L(K,Q) = \frac{1}{n}\sum_{i=1}^n l(f_{K,Q,\tau}(X,z), Y_i) \tag{KQ-ERM}$$

where $l(\cdot,\cdot)$ is logistic loss function. The author shows that W-ERM and KQ-ERM correspond to two different SVM problems through Regular Path Analysis. In our project, we focus on the relationship between training dynamics of different parameterizations under different scale factors $\tau = \frac{1}{\sqrt{m}}$ or $\tau = \frac{1}{m}$ when $m \to \infty$. Specifically, we focus on the following two problems,

- Will the distance between training dynamics of W-ERM and KQ-ERM converge when $m \to \infty$? With which scaling scheme, this convergence might occur?
- Which scaling scheme will lead to convergence of training dynamics between KQ-parameterization and its linearization? What is the induced NTK, and what is the implicit bias of the linearized model?

2 Literature Review

2.1 Implicit Bias of Transformer

Over-Smoothing. In transformer (or GNN), each layer repeatedly aggregates information from neighborhoods. As the model goes deeper, token (node) representations tend to converge and become indistinguishable, which can cause low-rank issues for the output hidden states. [DCL21] shows that without skip-connection or MLP, the output hidden states of the last layer of the transformer would be a rank-1 matrix. The output hidden states of multi-head transformer can be decomposed into a sum of single-attention paths. Each path would lead to its output (of the path) converging to a rank-1 matrix double exponentially. [Ges+23] analyzes the influence of depth from the perspective of Neural ODE, showing that even with skip-connection, the low-rank issue still occurs. For the Neural ODE model, we can have a probabilistic measure for tokens (or hidden states) of each "layer t" now.

Norm Growth and Saturation (Training) The saturation phenomenon means that the softmax would tend to be a one-hot vector or a vector (whose sum is 1) concentrated on several tokens during training, which can be viewed as an implicit bias of GD with transformer. During pertaining of T5 model, there is indeed the growth of weights norm and directional convergence trend of weights as exhibited in Fig 1, which coincides with the "separation theory" in [Tar+23]. Also, most heads exhibit saturation property in experiments [Mer+21]. Interestingly, the depth also exhibits a similar impact on softmax output. From Theorem 2.1 in [Ges+23], when depth goes large, self-attention matrix after softmax converges to a low-rank Boolean matrix.

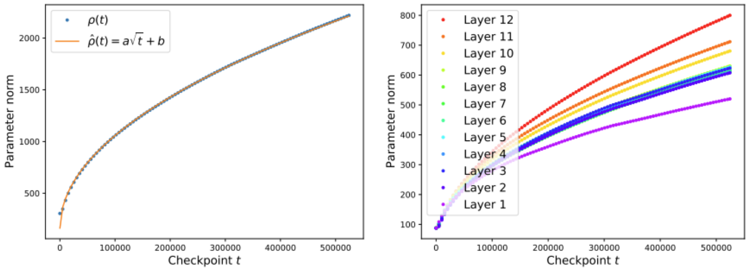


Figure 1: The norm growth phenomenon in T5 Training [Mer+21]

2.2 SVM Equivalence for One-Layer Attention

[Tar+23] establishes the equivalence between two ERM problems and the following two SVM formulations through Regular Path Analysis. Though $\tau = 1$ is used in [Tar+23], however, the scaling doesn't influence the equivalence in Section 2.2.1.

2.2.1 "Global" SVM Formulations

Given a prediction head $v \in \mathbb{R}^d$ the score of a token x_{it} of input X_i is defined as

$$\gamma_{it} = Y_i v^T x_{it}, \forall i \in [n], t \in [T]$$

and $opt_i = argmax_{t \in [T]} \gamma_{it}, \forall i \in [n]$[1]

Definition 1 (SVM for W-ERM)**.**

$$W^{mm} = argmin ||W||_F \ s.t. \ (x_{iopt_i} - x_{it})^T W z_i \geq 1, \forall t \neq opt_i, i \in [n] \tag{Att-SVM}$$

The intuition is, let $P_i = \mathbb{S}(X_i W z_i) \in \mathbb{R}^T$,

$$L(W) = \frac{1}{n} \sum_{i=1}^{n} l(\gamma_{it} P_{it}) \geq \frac{1}{n} \sum_{i=1}^{n} l(\gamma_{iopt_i})$$

if minimizing the training loss involves choosing the optimal token x_{iopt_i}, the softmax similarities should eventually converge to a one-hot vector, which corresponds to the norm growth phenomenon in [Mer+21]

Definition 2 (SVM for KQ-ERM)**.**

$$W_*^{mm} \in \underset{rank(W) \leq m}{argmin} \ ||W||_* \ s.t. \ (x_{iopt_i} - x_{it})^T W z_i \geq 1, \forall t \neq opt_i, i \in [n] \tag{Att-SVM$_*$}$$

where $\| \cdot \|_*$ denotes nuclear norm and when $m > d$, the constraint can be removed, which is always the case [Tar+23] considers.

Theorem 1 (Equivalence through RP Analysis, Informal)**.** Given $R > 0$, find,

$$\bar{W}_R = \underset{||W||_F \leq R}{argmin} \ L(W) \tag{W-RP}$$

$$(\bar{K}_R, \bar{Q}_R) = \underset{\|K\|_F^2 + \|Q\|_F^2 \leq 2R}{argmin} \ L(K, Q) \tag{KQ-RP}$$

We have,

$$\lim_{R \to \infty} \frac{\bar{W}_R}{R} = \frac{W^{mm}}{||W^{mm}||_F}, \lim_{R \to \infty} dist(\frac{\bar{K}_R \bar{Q}_R}{R}, \frac{\mathcal{W}_*^{mm}}{\|W_*^{mm}\|_*}) = 0 \tag{4}$$

This theorem indicates the equivalence between the ERM problem and the SVM problems.

2.2.2 Local Convergence for W-ERM

Definition 3 ("Locally Optimal" SVM for W-ERM)**.** Fix token indices $\alpha = (\alpha_i)_{i=1}^n$. Solve (Att-SVM) with $(opt_i)_{i=1}^n$ replaced by $(\alpha_i)_{i=1}^n$

$$W_\alpha^{mm} = argmin ||W||_F \ s.t. \ (x_{i\alpha_i} - x_{it})^T W z_i \geq 1, \forall t \neq \alpha_i, i \in [n] \tag{Att-SVM-α}$$

If (Att-SVM-α) admits a solution, then consider the set $\mathcal{T}_i \subseteq [T]$ such that $(x_{i\alpha_i} - x_{it})^T W_\alpha^{mm} z_i = 1$ for all $t \in \mathcal{T}_i$. We refer to $(\mathcal{T}_i)$ as support vectors of α. If

$$\gamma_{i\alpha_i} > \gamma_{it}, \forall i \in [n], t \in \mathcal{T}_i$$

indices $\alpha = (\alpha_i)_{i=1}^n$ are called locally-optimal and W_α^{mm} is called a locally-optimal direction.

[1] More precisely, it should be $opt_i \in argmax_t \in [T] \gamma_{it}$.

Theorem 2 (Local Convergence, Informal)**.** Under some conditions, GD with suitable step size η will converge to a specific locally-optimal solution,

$$\lim_{k} ||W(k)||_F = \infty, \lim_{k\to\infty} \frac{W(k)}{||W(k)||_F} = \frac{W_\alpha^{mm}}{||W_\alpha^{mm}||_F} \tag{5}$$

where $W(k)$ follows gradient descent, $W(k+1) = W(k) - \eta\nabla L(W(k))$.

In our report, we focus on the relationship of KQ-parameterization (KQ-param) with other parameterizations like W-parameterization (W-param) and the linearization of KQ-param under different scalings. The implicit bias of KQ-param is already established similar to Formula 4 and there is also the similarly defined locally optimal solution in Att-SVM$_*$ of KQ-param, $W_{\alpha,*}^{mm}$, as W_α^{mm} in W-param.

3 A Revisit of Lazy Framework on Two-Layer NN

As we concentrate the relationship between KQ-param and W-param under different scalings and we search for a valid scaling that leads to the ”correct” linearization model for the infinite-width attention layer with KQ-param, we have a brief review of [COB20] here, which is a useful ”framework” to analyze problems of this prototype.

3.1 Lazy Training Framework

We consider a parameterized predictor $\theta \in \mathbb{R}^p \to f_\theta \in \mathcal{F}$, where $f_\theta : \mathbb{R}^d \to \mathbb{R}$. Assume we have n data samples $x_1, ..., x_n \in \mathbb{R}^d$ and a smooth objective function $\mathcal{R} : \mathbb{R}^n \to \mathbb{R}$. This leads to the following parameterized objective[2] $F : \mathbb{R}^p \to \mathbb{R}$: $F(\theta) = R(f_\theta)$. Let us define a rescaled parameterization,

$$F_\alpha(\theta) = \frac{1}{\alpha^2} R(\alpha f_\theta) \tag{6}$$

and its linearization model at initialization,

$$\bar{F}_\alpha(\bar{\theta}) = \frac{1}{\alpha^2} R(\alpha \bar{f}_\theta) \tag{7}$$

where the linearization is, $\bar{f}_\theta = f_{\theta_0} + Df(\theta_0)(\bar{\theta} - \theta_0)$.

One thing to note, the scale factor in Formula 6, 7 is to ”synchronize time scale” for different α, which means if we apply GD on $F_\alpha(\theta)$, the loss difference is asymptotically the same for large α after each step. Then we give a theorem from [COB20] about the difference between two gradient flows.

Theorem 3. Given a fixed time horizon $T > 0$ and $f_{\theta_0} = 0$,

$$sup_{t\in[0,T]} \|\alpha f_{\theta_t} - \alpha \bar{f}_{\bar{\theta}_t}\| = O(1/\alpha)$$
$$sup_{t\in[0,T]} \|\theta_t - \bar{\theta}_t\| = O(1/\alpha^2)$$

This theorem means that the increase of α will make the model behave like its linearization at initialization.

Experiments. We start a small one-step/few-step GD experiment to show this convergence rate in the distance between two predictor spaces and also we will show the importance of condition $f_{\theta_0} = 0$. The

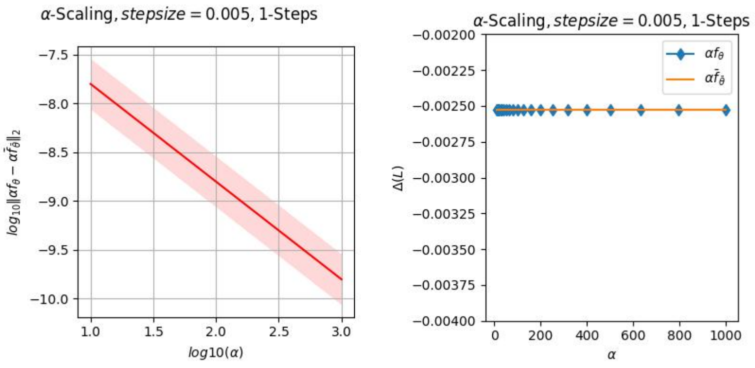
[2] We might sometimes simply regard $f_\theta \in \mathbb{R}^n$ as the outputs for n data samples

reason we do this experiment is to help us better understand and design the one-step GD experiment on infinite-width two-layer NN and infinite-width attention layer with different scale factors This is more like a record of solving the problem of previous inconsistency between my experiment results and theory, which might not be that related to the main concern of this project.. We define a 2-layer NN,

$$\alpha f_\theta(x) = \alpha \sum_{j=1}^{m} \phi(w_j, x) \tag{8}$$

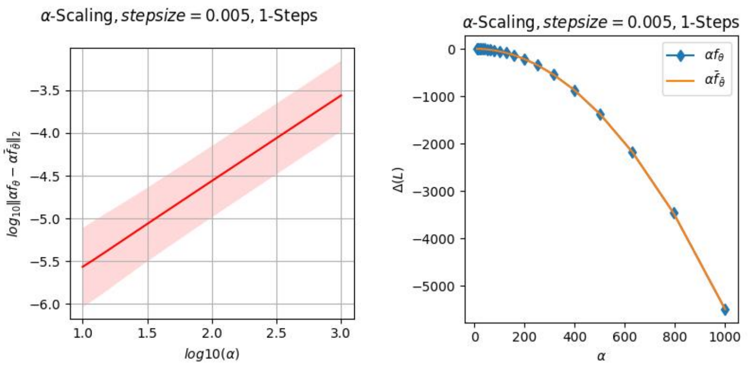
where $w_j = (a_j, b_j)$, $a_j \in \mathbb{R}$ and $b_j \in \mathbb{R}^d$, $\phi(w_j, x) = a_j \sigma(b_j^T x)$. We can assume $\sigma(\cdot)$ can be a ReLU activation function (non-smooth), identity function, or sigmoid function and $w_j \sim \mu_0 \in \mathcal{P}(\mathbb{R}^{d+1})$. We consider a dataset $(x_i, y_i)_n$, where $x_i \in \mathbb{R}^d, y_i \in \mathbb{R}$. In our experiment, we set $n = 5, d = 10$, set R mean square loss, set $\mu_0 = \mathcal{N}(0, I_{d+1})$ as our defualt setting and we would replace f_θ by $f_\theta - f_{\theta_0}$ to make sure zero initialization if we don't mention explicitly.

Experiement - $f_{\theta_0} = 0$ matters. In this experiment, we fix $m = 20$, and see the influence of α.



(a) (b)

Figure 2: One-step GD for two-layer NN with zero initialization. We set step size to $\eta_\alpha = \eta_0 \dfrac{1}{\alpha^2}$ with $\eta_0 = 0.005$ in our implementation. $\Delta(L)$ in (b) refers to the change of $R(\alpha f_\theta)$ (not $F_\alpha(\theta)$) after one-step GD.

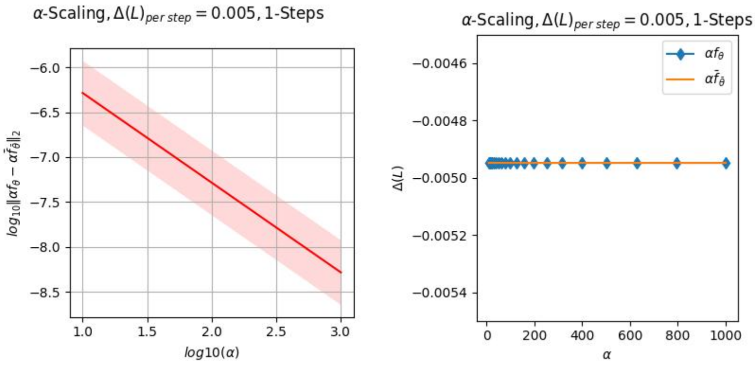


(a) (b)

Figure 3: One-step GD for two-layer NN with non-zero initialization.

We can see from Fig 2 that with zero initialization, $\|\alpha f_{\theta_1} - \alpha \bar{f}_{\bar{\theta}_1}\| \sim O(1/\alpha)$, which is consistent with Theorem 3. However, if the initialized output is not zero, the Theorem 3 doesn't hold anymore. This is useful for our infinite-width experiment in Section 3.2 and Section 4. Although $E_{w\sim\mu_0}[\phi(w,x)] = 0$, symmetric initialization strategy from [COB20] indeed removes the randomness at initialization and helps to show a more consistent result when m is not that large.

Experiment - $\eta_\alpha = \dfrac{\eta_1}{\|\nabla R(\alpha f_\theta)\|^2}$. This step size choice can approximately make the change of loss equaling to η_1 after each step for different parameterizations[3], which is consistent with the experiment result in Fig 4(b). And it is clear that after one-step GD, $\|\alpha f_{\theta_1} - \alpha \bar{f}_{\bar{\theta}_1}\| \sim O(1/\alpha)$. It is easy to prove that $\|\nabla R(\alpha f_{\theta_0})\| \sim \Theta(\alpha)$, which is under our expectation. And that is why this step size will lead to similar performance as $\eta_\alpha = \dfrac{\eta_0}{\alpha^2}$. If we want to observe the $\|\theta_t - \bar{\theta}_t\|$, we can do a t-step GD with $t > 1$ as one-step GD will lead to $\|\theta_1 - \bar{\theta}_1\| = 0$. The results of multi-step GD are in Appendix A.1.



(a)

(b)

Figure 4: One-Step GD for two-layer NN with zero initialization with $\eta_\alpha = \dfrac{\eta_1}{\|\nabla R(\alpha f_\theta)\|^2}$, where $\eta_1 = 0.005$ is the $\Delta(L)_{per\ step}$ in the figure.

Take away from the experiments. Zero initialization matters, and we will use step size of $\eta_\alpha = \dfrac{\eta_1}{\|\nabla R(\alpha f_\theta)\|^2}$ in the remaining part of the report. The reason why we use this step size is explained in Appendix A.2. Then we want to use this lazy training framework to analyze our attention layer in the following.

3.2 Lazy Training Analysis on Infinite Two-Layer NN

Before moving to the analysis of the infinite-width attention layer, we first revisit the infinite-width two-layer NN first. As claimed in [COB20], we consider a two-layer NN with the following scheme,

$$f_{\theta^m}(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \phi(w_j, x) = \sqrt{m} \left(\frac{1}{m} \sum_{j=1}^{m} \phi(w_j, x) \right)$$

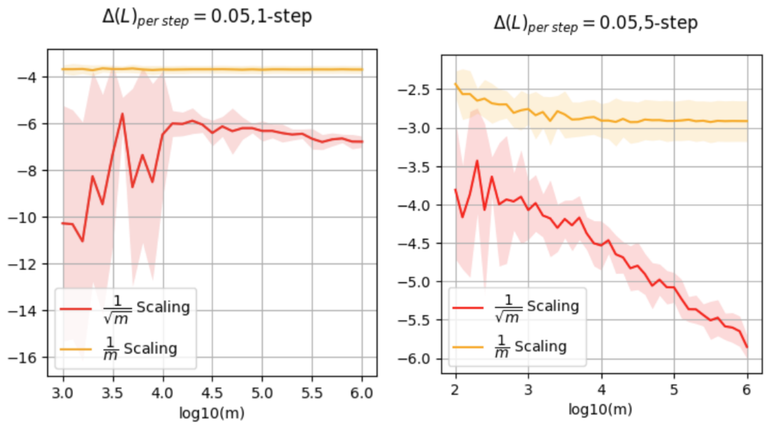
[3] This is a general result for any function [CN23]

where $\left(\dfrac{1}{m}\sum_{j=1}^{m}\phi(w_j,x)\right) \to \mathbb{E}_{w\sim\mu_0}\left[\phi(w,x)\right]$. Asymptotically, $\mathbb{E}_{w\sim\mu_0}\left[\phi(w,x)\right]$ is equivalent to the f_θ in lazy training framework, and f_{θ^m} is equivalent to αf_θ with $\alpha = \sqrt{m}$. Therefore, we have the following corollary on two-layer NN.

Corollary 1. Given a fixed time horizon $T > 0$,

$$sup_{t\in[0,T]}\|f_{\theta_t^m} - \bar{f}_{\bar{\theta}_t^m}\| = O(1/\sqrt{m})$$
$$sup_{t\in[0,T]}\|\theta_t^m - \bar{\theta}_t^m\| = O(1/m)$$

Experiments. We use the symmetric initialization strategy from [COB20] to make sure that $f_{\theta_0^m} = 0$ for any m that is even.



(a) (b)

Figure 5: Few-Step GD for two-layer ReLU NN with symmetric initialization. The y-axis refers to $log_{10}\|f_{\theta^m} - \bar{f}_{\bar{\theta}^m}\|$ for $\frac{1}{\sqrt{m}}$ scaling and log_{10} value of the corresponding distance in predictor space under $\frac{1}{m}$ scaling. (a) represents the 1-step GD experiment, in which $\|f_{\theta_1^m} - \bar{f}_{\bar{\theta}_1^m}\| \sim O(\frac{1}{\sqrt{m}})$ when m is large enough. (b) represents a 5-step GD experiment, in which $\|f_{\theta_5^m} - \bar{f}_{\bar{\theta}_5^m}\| \sim O(\frac{1}{\sqrt{m}})$ is more clear.

Take away from this experiment. Sometimes multiple-step GD experiments show the result (the rate) more clearly, as Fig 5 suggests. Also, we need multiple-step GD to observe $\|\theta_t - \bar{\theta}_t\|$ as in Appendix A.1. [4]

4 KQ-param and W-param of Attention Layer

Theoretically, the local convergence of GD on W-ERM is guaranteed under some ”strong” conditions; Empirically, when $d \gg n$, with high probability, the local convergence of W-param can be achieved showed in [Tar+23]. In this section, we would like to see the distance between W-param and KQ-param during training, to see which scaling scheme will lead to a converge phenomenon between the two training dynamicses.

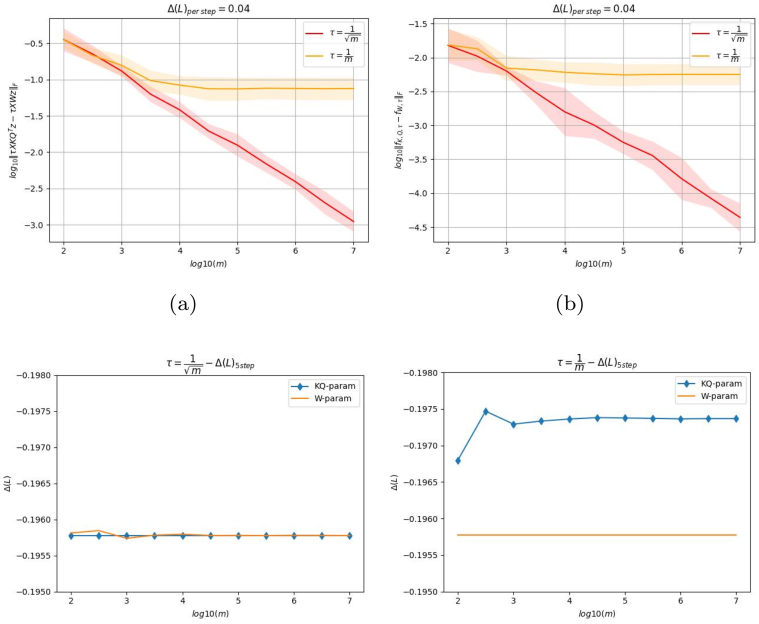
[4] In addition, the symmetric initialization is important for the effect of the simulation from my experience.

4.1 Few-Step GD

Since V of the original attention model as Formula 1 is not inclulded in the trainable parameters in our model, the only trainable part is within the softmax $\mathbb{S}(\cdot)$. Also, assume we have a fixed dataset $(Y_i, X_i, z_i)_n$, we simplify identify the ERM problem as

$$l(f_{K,Q,\tau}(X,z),Y) = l(v^T X^T \mathbb{S}(\tau XKQ^T z, Y) = \tilde{l}(\tau XKQ^T z) \tag{9}$$

in which a structure similar to two-layer NN with different scale factors occurs. With experience in lazy training frameworks on two-layer NN, we will design a similar few-step GD experiment on the W-param and KQ-param of the two models in this section. We will use the step size scheme with $\eta = \frac{\eta_1}{\|\nabla l(f_{K,Q,\tau})\|^2}$ and $\eta = \frac{\eta_1}{\|\nabla l(f_{W,\tau})\|^2}$ to control the loss difference to be the same after each step for each parameterization, which achieves the synchronization of "time scale". From Fig 6, we observe a clear convergence between two different dynamics after "fixed Δt" under the scaling of $\tau = \frac{1}{\sqrt{m}}$.



(a)

(b)

(c)

(d)

Figure 6: 5-Step GD for attention layer with zero initialization in softmax with $\eta_\alpha = \frac{\eta_1}{\|\nabla_{\theta_0} R(\alpha f_\theta)\|^2}$.

4.2 Training Trajectory

Also, we do a simple experiment to see whether the increased m would lead to the same training dynamics between W-param and KQ-param. According to [CB20], when there is a large initialization with an exponential tail in a classification setting, the training dynamics will detour to the solution of the linearized model first, demonstrating the characteristic of kernel regime which cannot be seen in small initialization.

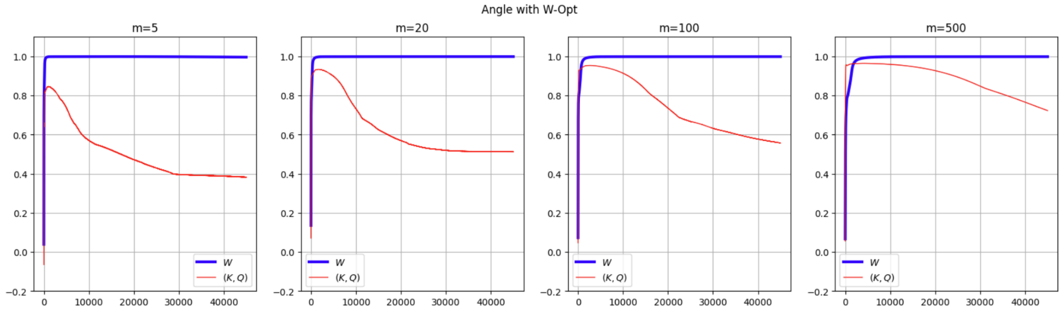


Figure 7: Training Trajectory with W-param and KQ-param under $\frac{1}{\sqrt{m}}$ scaling. The y-axis corresponds to the cosine similarity with W_{α}^{mm}, the optimal solution of (Att-SVM-α), where α is determined by the directional convergent solution of training dynamics under W-param. In this experiment, $(n, T, d) = (5, 6, 8)$.

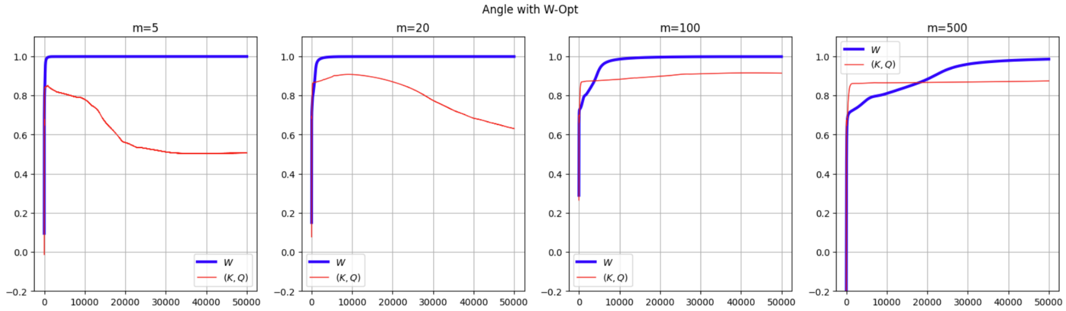


Figure 8: Training Trajectory with W-param and KQ-param under $\frac{1}{m}$ scaling. Although there is a detour phenomenon, no clear phenomenon of "approaching" to W_{α}^{mm} when m increases is observed.

In our experiment, we observe a clear detour phenomenon in the training of KQ-param to W_{α}^{mm}, and the detour weight is "approaching" W_{α}^{mm} when $m \to \infty$, which is not demonstrated under $\frac{1}{m}$ scaling[5] in Fig 8.

Conclusion. Empirical experiments suggest that under $\tau = \frac{1}{\sqrt{m}}$ scaling, there might be a convergence of distance between the training trajectory of W-param and KQ-param. However, the analysis of the distance between these two parameterizations might be untractable. Thus we will turn to the distance between KQ-param and its linearization model first, for which there are already some guarantees, and there are more tools to analyze it.

5 KQ-Param and its Linearization

Linearization inside softmax. Similar as Formula 9, we still focus the KQ-param here and combine $\mathbb{S}(\cdot)$ into $\tilde{l}$, and thus we define

$$g_{K,Q,\tau}(X, z) = \tau XKQz \tag{10}$$

with which we can define its linear model $\bar{g}_{\bar{K},\bar{Q},\tau}$ through first-order expansion. We implement a similar experiment on KQ-param as infinite two-layer NN here.

[5] We already make sure that the dataset and v are fixed.

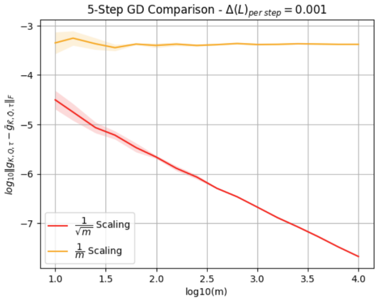


Figure 9: 5-step GD for KQ-param in attention layer

We observe that, similar to two-layer NN, $\tau = \dfrac{1}{m}$ will lead to $\|g_{K,Q,\tau} - \bar{g}_{\bar{K},\bar{Q},\tau}\| \sim \Theta(1)$ within a fixed time of training, while $\tau = \dfrac{1}{\sqrt{m}}$ will lead to $\|g_{K,Q,\tau} - \bar{g}_{\bar{K},\bar{Q},\tau}\| \sim \Theta(1/m)$[6]. This suggests that $\tau = \dfrac{1}{\sqrt{m}}$ can lead to a kernel regime.

Combined with softmax. As [Hro+20; Wu+23] shows, with $\tau = \dfrac{1}{\sqrt{m}}$, each element of τKQ^T will be a random variable with mean 0 and variance 1 at initialization. However, $\tau = \dfrac{1}{\sqrt{m}}$ will cause each element τKQ^T convergent to 0 when $m \to \infty$, which might force the softmax to be an average pooling layer. However, [Hro+20; Wu+23] both state that with $\dfrac{1}{m}$, the attention structure can lead to a valid NTK, which seems very controversial to the current understanding. However, it is probably because in the linearization of [Hro+20; Wu+23] the softmax term is treated as a whole and it is not the output layer in their setting. Considering the saturation property, the scaling inside softmax doesn't bring a large initialization effect, which will change the analysis.

Remained doubts. For this part, I have some doubts remained for future exploration,

- When can we apply the lazy training analysis? In the two-layer attention layer model, we can formulate it in an expectation term (mean-field parameterization) multiplied by $\sqrt{m}$. Now assume we have another Neural Network in the following formula and we still use $\dfrac{1}{\sqrt{m}}$ scaling,

$$g_\theta(x) = \begin{pmatrix} g_{\theta,1(x)} \\ \vdots \\ g_{\theta,d_{out}(x)} \end{pmatrix}, \; g_{\theta,i}(x) = \frac{1}{\sqrt{m}} \sum_{j=1}^{m} V_{i,j}\sigma(U_j x) \tag{11}$$

  where $V \in \mathbb{R}^{d_{out} \times m}, U_j \in \mathbb{R}^d, x \in \mathbb{R}^d$. Now, we can still write each element of $g_\theta(\cdot)$ as "a scale factor α times an expectation term". However, $g_{\theta,i}(\cdot)$ is not independent with each other as they share $U = \begin{pmatrix} U_1^T \\ \vdots \\ U_m^T \end{pmatrix}$. What is the influence of losing independence on lazy training analysis? (We care about this because the independence is also missing in KQ-param.)

[6] It might seems inconsistent with $O(\dfrac{1}{\sqrt{m}})$ as lazy training theory suggests, however, we have seen similar rate when we use identity activation in two-layer NN as Fig 12 in Appendix A.3.

- Why $\tau = \dfrac{1}{m}$ could lead to a valid NTK? The argument in [Hro+20] is based on NNGP and Gaussian Conditioning, for which I will learn some computation techniques to better understand. Generally, combining softmax and loss function might not be a good choice to analyze the attention layer as it is a very special structure of attention. However, when we do the linearization inside the softmax, it makes the infinite-width limit with $\tau = \dfrac{1}{\sqrt{m}}$ more meaningful, since $\tau = \dfrac{1}{\sqrt{m}}$ is what a practical transformer uses. How do we choose between those two?
- Once we find the "correct" linearization scheme, we can finally arrive at the problem we originally cared about. What will be the implicit bias of the linearized model at the infinite-width limit?

6 Conclusion

In this project, we conduct a comprehensive analysis of lazy training within the explicit scale format and the NTK parameterization of two-layer neural networks. This analysis is to ensure each block is experimentally and theoretically "correct", which is crucial for subsequent investigations (the selection of step size and symmetric initialization). We empirically investigate the impact of the scaling factor on the distance of training dynamics between the KQ-param and W-param of the attention layer. Our findings suggest that a scaling factor of $\frac{1}{\sqrt{m}}$ may lead to convergence of the distance between these two parameterizations. However, a more thorough examination is necessary for the implicit bias of linearization of KQ-param, which will be a focus of future research.

References

[Bro+20] Tom Brown et al. “Language Models are Few-Shot Learners”. In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle et al. Vol. 33. Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

[CB20] Lénaïc Chizat and Francis Bach. “Implicit Bias of Gradient Descent for Wide Two-layer Neural Networks Trained with the Logistic Loss”. In: *Proceedings of Thirty Third Conference on Learning Theory*. Ed. by Jacob Abernethy and Shivani Agarwal. Vol. 125. Proceedings of Machine Learning Research. PMLR, Sept. 2020, pp. 1305–1338. URL: https://proceedings.mlr.press/v125/chizat20a.html.

[CN23] Lénaïc Chizat and Praneeth Netrapalli. *Steering Deep Feature Learning with Backward Aligned Feature Updates*. 2023. arXiv: 2311.18718 [cs.LG].

[COB20] Lenaic Chizat, Edouard Oyallon, and Francis Bach. *On Lazy Training in Differentiable Programming*. 2020. arXiv: 1812.07956 [math.OC].

[DCL21] Yihe Dong, Jean-Baptiste Cordonnier, and Andreas Loukas. “Attention is not all you need: pure attention loses rank doubly exponentially with depth”. In: *Proceedings of the 38th International Conference on Machine Learning*. Vol. 139. Proceedings of Machine Learning Research. PMLR, 18–24 Jul 2021, pp. 2793–2803. URL: https://proceedings.mlr.press/v139/dong21a.html.

[Ges+23] Borjan Geshkovski et al. *The emergence of clusters in self-attention dynamics*. 2023. arXiv: 2305.05465 [cs.LG].

[Hro+20] Jiri Hron et al. “Infinite attention: NNGP and NTK for deep attention networks”. In: *Proceedings of the 37th International Conference on Machine Learning*. Ed. by Hal Daumé III and Aarti Singh. Vol. 119. Proceedings of Machine Learning Research. PMLR, 13–18 Jul 2020, pp. 4376–4386. URL: https://proceedings.mlr.press/v119/hron20a.html.

[Mer+21] William Merrill et al. “Effects of Parameter Norm Growth During Transformer Training: Inductive Bias from Gradient Descent”. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics, Nov. 2021, pp. 1766–1781. DOI: 10.18653/v1/2021.emnlp-main.133. URL: https://aclanthology.org/2021.emnlp-main.133.

[Ope+23] OpenAI et al. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774 [cs.CL].

[Tar+23] Davoud Ataee Tarzanagh et al. *Transformers as Support Vector Machines*. 2023. arXiv: 2308.16898 [cs.LG].

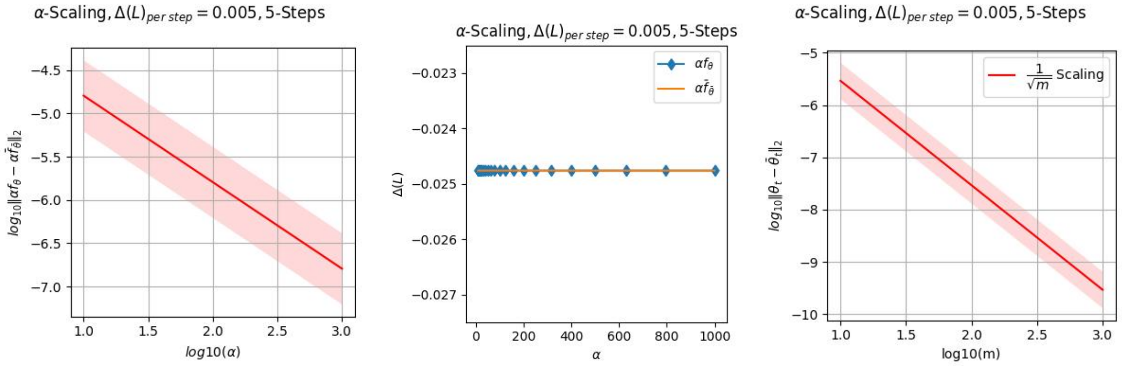
[Tou+23] Hugo Touvron et al. *LLaMA: Open and Efficient Foundation Language Models*. 2023. arXiv: 2302.13971 [cs.CL].

[Vas+17] Ashish Vaswani et al. “Attention is All you Need”. In: *Advances in Neural Information Processing Systems*. Ed. by I. Guyon et al. Vol. 30. Curran Associates, Inc., 2017. URL: https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

[Wu+23] Yongtao Wu et al. *On the Convergence of Encoder-only Shallow Transformers*. 2023. arXiv: 2311.01575 [cs.LG].

A More Results from Analysis of Lazy Training Framework

A.1 Multiple-Step GD Experiments for α-Scaling



(a) (b) (c)

Figure 10: 5-Step GD for two-layer NN with zero initialization with $\eta_\alpha = \dfrac{\eta_1}{\|\nabla_{\theta_0} R(\alpha f_\theta)\|^2}$. We observe from (c) that $\|\theta_1 - \bar{\theta}_1\| \sim O(\frac{1}{\alpha^2})$

A.2 A More Careful Observation about the Step Size for Infinite-Width NN

For the α scaling case discussed in section 3.1, we can check easily that $\|\nabla_{\theta_0} R(\alpha f_\theta)\| \sim \Theta(\alpha)$, and that's why $\eta_\alpha = \dfrac{\eta_1}{\|\nabla_{\theta_0} R(\alpha f_\theta)\|^2}$ can lead to similar simulation effect as $\eta_\alpha = \dfrac{\eta_0}{\alpha^2}$. However when it comes to infinite-width two-layer NN, things become a bit different. Assume $f_{\theta_0^m} = 0$[7], consider for a fixed x[8],

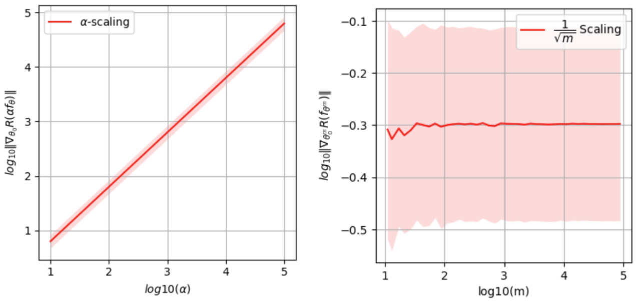
$$\|\nabla_{\theta_0^m} R(f_{\theta^m}(x))\|^2 = \sum_j \|\nabla_{w_j} f_{\theta^m}(x))\|^2 \|R'(0)\|^2 \tag{12}$$

$$= \|R'(0)\|^2 \sum_j \frac{1}{m} \|\nabla_{w_j} \phi(w_j, x)\|^2 \tag{13}$$

$$\sim \Theta(1) \neq \Theta(m) \tag{14}$$

[7] In general, $\mathbb{E}_{w_j \sim \mu_0}[\phi(w_j, \cdot)] = 0$. However, our symmetric initialization in experiments can ensure zero output for any m.

[8] Here we could treat $\mathcal{R} : \mathbb{R} \to \mathbb{R}$



(a) (b)

Figure 11: The scale of the norm of gradient under α-scaling and $\dfrac{1}{\sqrt{m}}$-scaling. (a) shows the $\|\nabla_{\theta_0} R(\alpha f_\theta)\| \sim \Theta(\alpha)$ for α-scaling in the original rescaled setting in [COB20], (b) shows the $\|\nabla_{\theta_0^m} R(f_{\theta^m}(x))\| \sim \Theta(1)$ for $\dfrac{1}{\sqrt{m}}$-scaling in infinite-width NN setting.

The step size of $\dfrac{\eta_1}{\|\nabla R\|^2}$ suggests if we want to "synchronize time scale" for $\frac{1}{\sqrt{m}}$-parameterization, we should directly use a fixed step size η_1 for any m instead of $\eta_m = \dfrac{\eta_0}{m}$ suggested by the rescaled model in Formula 6. This "inconsistency" is a bit tricky, probably stemming from the "infinite-width" and we should be careful about it. Also, if we change the NTK initialization (with an explicit $\frac{1}{\sqrt{m}}$) to Lecun initialization ($a_j \sim \mathcal{N}(0, 1/m)$), we will obtain $\|\nabla_{\theta_0^m} R(f_{\theta^m}(x))\| \sim \Theta(\sqrt{m})$, which is consistent with lazy training framework.

A.3 Two-Layer NN with Identity Activation

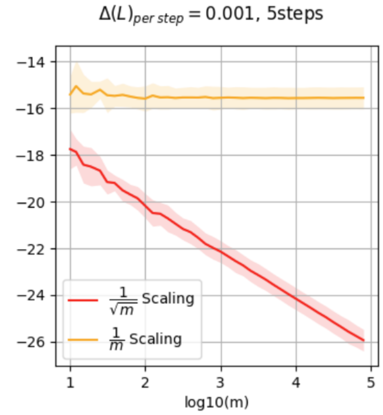


Figure 12: Few-Step GD for two-layer NN with Identity activation and symmetric initialization. The y-axis refers to $log_{10}\|f_{\theta^m} - \bar{f}_{\bar{\theta}^m}\|$ for $\frac{1}{\sqrt{m}}$ scaling and log_{10} value of the corresponding distance in predictor space under $\frac{1}{m}$ scaling.