Can the Spectrum of the Neural Tangent Kernel Anticipate Fine-Tuning Performance?

Zahra Rahimi Afzal, Tara Esmaeilbeig, Mojtaba Soltanalian and Mesrob I Ohannessian

UNIVERSITY OF ILLINOIS CHICAGO

Introduction

Given the pre-trained model $f_{\boldsymbol{\theta}_0}$ and the target dataset $\mathcal{D}_T = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ for the downstream task, we look at fine-tuning as an NTK regression problem.

- In SGD, the update to parameters at step t is given by

$$\begin{aligned}\boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t &= \eta \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_T}[\nabla_{\boldsymbol{\theta}} \mathcal{L}(f_{\boldsymbol{\theta}_t}(\mathbf{x}), \mathbf{y})] \\ &= \eta \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_T}[\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_t}(\mathbf{x}) \nabla_f \mathcal{L}(f_{\boldsymbol{\theta}_t}(\mathbf{x}), \mathbf{y})]. \end{aligned} \tag{1}$$

- Using the first-order Taylor expansion

$$\begin{aligned} f_{\boldsymbol{\theta}_{t+1}}(\mathbf{x}') - f_{\boldsymbol{\theta}_t}(\mathbf{x}') &\approx \langle \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_t}(\mathbf{x}'), \boldsymbol{\theta}_{t+1} - \boldsymbol{\theta}_t \rangle \\ &= \eta \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_T}\left[\nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_t}(\mathbf{x}')^\top \cdot \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_t}(\mathbf{x}) \nabla_f \mathcal{L}(f_{\boldsymbol{\theta}_t}(\mathbf{x}), \mathbf{y})\right] \\ &= \eta \mathbb{E}_{\mathbf{x}\sim\mathcal{D}_T}\left[\mathbf{k}_t(\mathbf{x}, \mathbf{x}') \nabla_f \mathcal{L}(f_{\boldsymbol{\theta}_t}(\mathbf{x}), \mathbf{y})\right]. \end{aligned} \tag{2}$$

This indicates that the learning dynamics of SGD is equivalent to NTK regression when the kernel is chosen to be the NTK, i.e., $\mathbf{k}_t(\mathbf{x}, \mathbf{x}') = \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_t}(\mathbf{x}')^\top \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_t}(\mathbf{x})$.

- We call the model linearized or in the lazy regime if $\mathbf{k}_t(\mathbf{x}, \mathbf{x}') \approx \mathbf{k}_0(\mathbf{x}, \mathbf{x}')$.
- Looking at fine-tuning through the lens of Neural Tangent Kernel (NTK) regression:

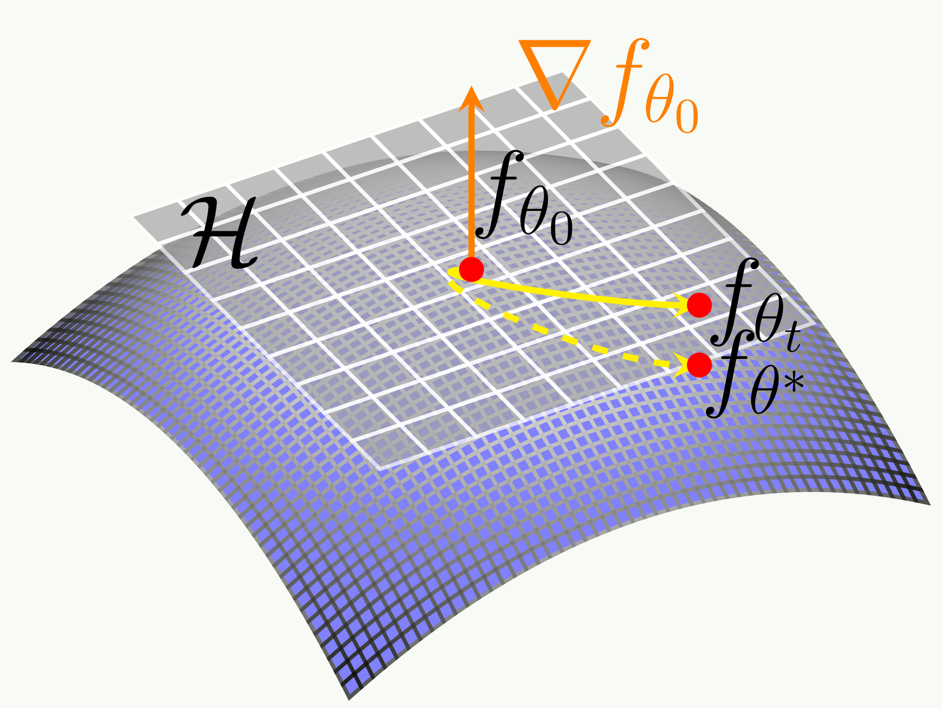


Figure 1. Fine-tuning in the lazy regime is close to kernel regression on the tangent space. $f_{\theta^*}(\mathbf{x})$ is the fine-tuned model obtained by empirical risk minimization. If fine-tuning remains in the linearized regime, then after T steps of training $f_{\theta^*}(\mathbf{x}) \approx f_{\boldsymbol{\theta}_0}(\mathbf{x}) + \langle \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{x}), \boldsymbol{\theta}_T - \boldsymbol{\theta}_0 \rangle$.

- $\boldsymbol{\theta}^l \to$ the parameters of layer l from the pretrained model.
- The NTK matrix is defined as $[\mathbf{K}]_{i,j} = \sum_{l=1}^L \nabla_{\boldsymbol{\theta}^l} f_{\boldsymbol{\theta}}(\mathbf{x}_i)^\top \nabla_{\boldsymbol{\theta}^l} f_{\boldsymbol{\theta}}(\mathbf{x}_j)$.

Neural Tangent Kernel regression

The fine-tuned model is denoted by $f_{\boldsymbol{\theta}^*}(\cdot): \mathbb{R}^d \to \mathbb{R}^c$ which is obtained by minimizing the typical empirical risk minimization problem

$$\boldsymbol{\theta}^* = \underset{\boldsymbol{\theta}}{\text{minimize}}\ \mathcal{R}(\boldsymbol{\theta}), \tag{3}$$

where

$$\mathcal{R}(\boldsymbol{\theta}) = \frac{1}{n} \sum_{i=1}^n \mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i). \tag{4}$$

Neural Tangent Kernel regression

Let $\mathcal{H}$ be the reproducing kernel Hilbert space (RKHS) endowed with a positive definite kernel function $\mathbf{k}(\cdot, \cdot)$, i.e.,

$$\mathcal{H} = \left\{ f(\cdot) = \sum_{i=1}^n \alpha_i \mathbf{k}(\cdot, \mathbf{x}_i) \right\}.$$

Assuming the solution lies in or close to this Hilbert space, then as an alternative to (3), we solve

$$\underset{f\in\mathcal{H}}{\text{minimize}}\ \frac{1}{n} \sum_{(\mathbf{x},\mathbf{y})\in\mathcal{D}_T} \|f(\mathbf{x}) - \mathbf{y}\|_2^2 + \sigma \|f\|_{\mathcal{H}}^2, \tag{5}$$

$f^*(\cdot) = \mathbf{K}(\cdot, \mathbf{X}) [\mathbf{K}(\mathbf{X}, \mathbf{X}) + \sigma \mathbf{I}]^{-1} \mathbf{y}$.

Main Theorem

The empirical risk is bounded as

$$\frac{\sigma \|\mathbf{y}\|_2^2}{\sigma + \lambda_{\max}(\mathbf{K})} \leq \mathcal{R}(\boldsymbol{\theta}) \leq \frac{\sigma \|\mathbf{y}\|_2^2}{\sigma + \lambda_{\min}(\mathbf{K})} \tag{6}$$

where $\lambda_{\min}(\mathbf{K})$ and $\lambda_{\max}(\mathbf{K})$ are the minimum and maximum eigenvalues of $\mathbf{K}(\mathbf{X}, \mathbf{X})$, respectively.

How does layer selection change the Eigenvalue spectrum of the NTK?

Let $\mathbf{K}$ be the NTK with respect to the set of selected fine-tuning parameters and $\mathbf{S}$ be the kernel with respect to the parameters of the candidate layers, to add to fine-tuning parameters. Then

$$(1-\eta)\lambda_i(\mathbf{K}) \leq \lambda_i(\mathbf{K} + \mathbf{S}) \leq (1+\eta)\lambda_i(\mathbf{K}), \tag{7}$$

where $\eta = \|\mathbf{K}^{-1/2} \mathbf{S}\, \mathbf{K}^{-1/2}\|$.

Interaction Between eigenvalue spectrum and risk Bounds

Let $\mathbf{K}$ be the NTK induced by the trainable parameters in $\boldsymbol{\theta}$, then if $\kappa(\mathbf{K} + \sigma\mathbf{I}) \leq c$, we have

$$\frac{\lambda_{\max}(\mathbf{K} + \mathbf{S} + \sigma\mathbf{I})}{a \lambda_{\max}(\mathbf{K} + \sigma\mathbf{I})} \leq \frac{\mathcal{R}(\boldsymbol{\theta} \cup \hat{\boldsymbol{\theta}})}{\mathcal{R}(\boldsymbol{\theta})} \leq \frac{a \lambda_{\max}(\mathbf{K} + \mathbf{S} + \sigma\mathbf{I})}{\lambda_{\max}(\mathbf{K} + \sigma\mathbf{I})}, \tag{8}$$

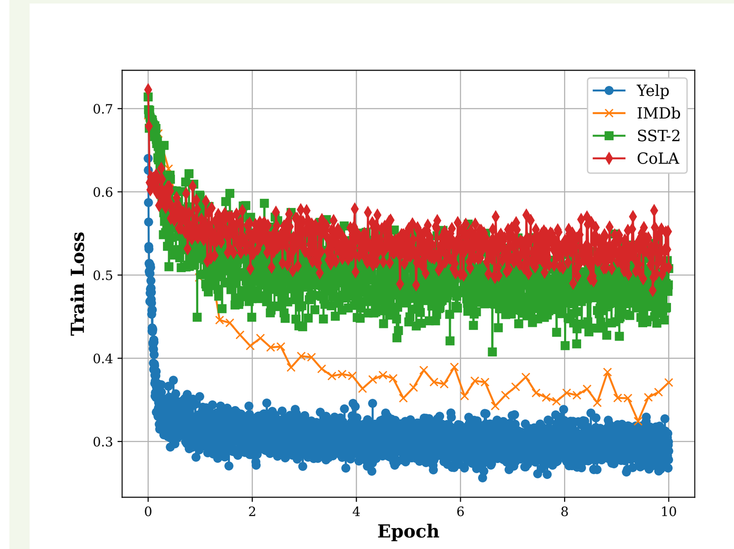
where $a = \frac{c}{(1-\eta)^2}$, $\eta = \|\mathbf{K}^{-1/2} \mathbf{S} \mathbf{K}^{-1/2}\|$ and $\mathbf{S}$ is the kernel induced by $\hat{\boldsymbol{\theta}}$ with $[\mathbf{S}]_{i,j} = \nabla_{\hat{\boldsymbol{\theta}}} f_{\boldsymbol{\theta}}(\mathbf{x}_i)^\top \nabla_{\hat{\boldsymbol{\theta}}} f_{\boldsymbol{\theta}}(\mathbf{x}_j)$.

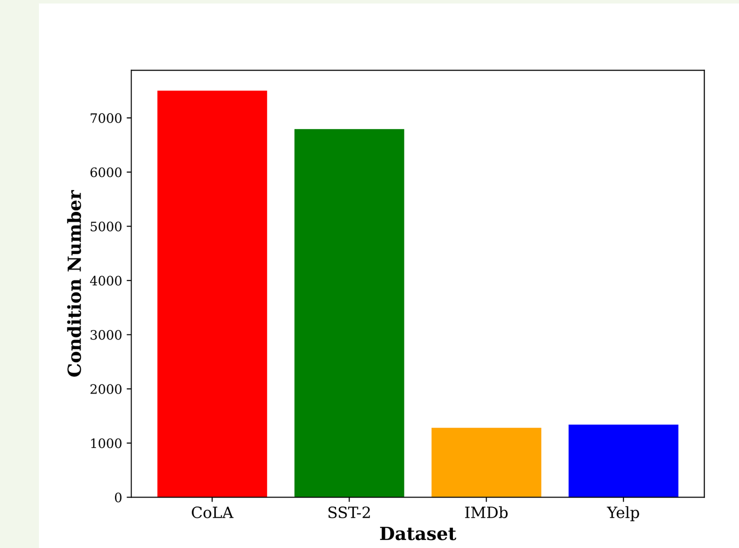
- Time(s) for calculating the NTK on 32 random samples from the training set:

Dataset	Fine-tuning Time	NTK Calculation Time
CoLA	187	33
SST-2	794	63
Yelp	46,096	245

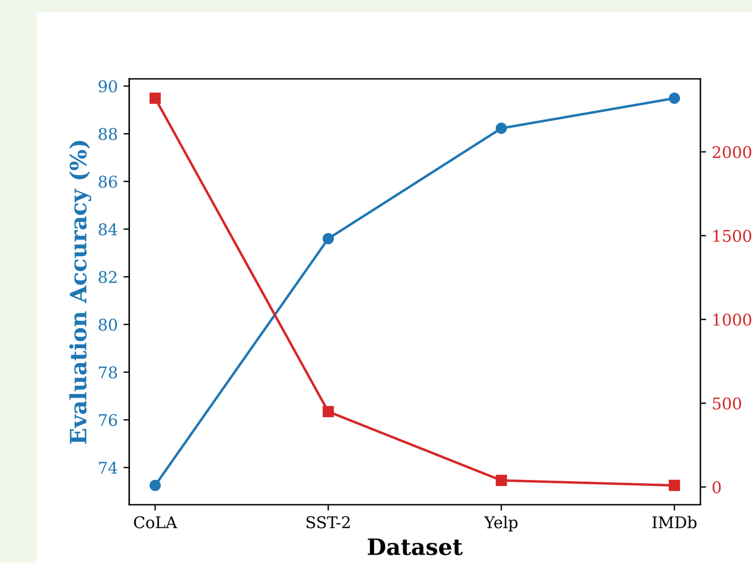
Numerical Result

- There is a positive correlation between the convergence rate of optimization steps of LoRA over 10 epochs and condition number $\kappa(\mathbf{K} + \sigma\mathbf{I})$ of NTK at initialization:

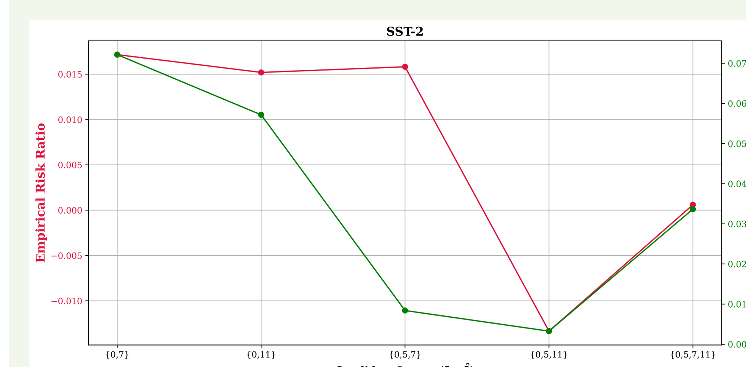


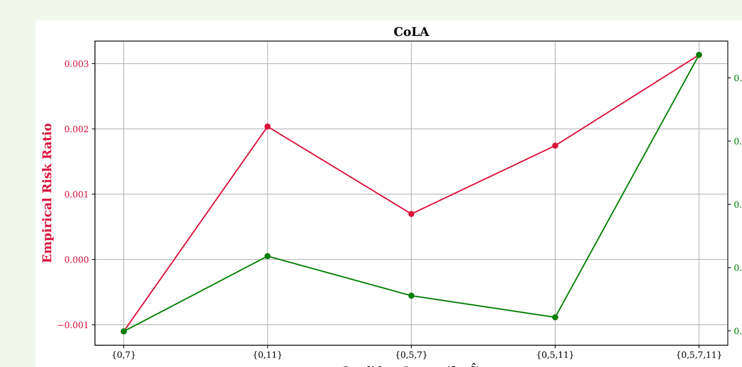


- There is a negative correlation between evaluation accuracy and the condition number of NTK. LoRA with $r = 8$ is used to fine-tune $\{\mathbf{W}_k\}$ of the layers $\{0, 5, 11\}$. In our experiments we observed that $\lambda_{\min}(\mathbf{K}) \approx 0 \to$ the regularized condition number, $\kappa(\mathbf{K} + \sigma\mathbf{I})$, is tracing $\lambda_{\max}(\mathbf{K})$.



- Empirical risk ratio $\log\left(\frac{\mathcal{R}(\boldsymbol{\theta} \cup \hat{\boldsymbol{\theta}})}{\mathcal{R}(\boldsymbol{\theta})}\right)$ and maximum eigenvalue ratio $\log\left(\frac{\lambda_{\max}(\mathbf{K}+\mathbf{S}+\sigma\mathbf{I})}{\lambda_{\max}(\mathbf{K}+\sigma\mathbf{I})}\right)$ are used to evaluate the impact of candidate layers on the model. Here, $\boldsymbol{\theta}$ is fixed as the weights $\{\mathbf{W}_k\}$ of layer $\{0\}$, while $\hat{\boldsymbol{\theta}}$ represents the candidate layers:





References

- Malladi, S., Wettig, A., Yu, D., Chen, D. and Arora, S., 2023, July. A kernel-based view of language model fine-tuning. In International Conference on Machine Learning (pp. 23610-23641). PMLR.
- Jacot, A., Gabriel, F. and Hongler, C., 2018. Neural tangent kernel: Convergence and generalization in neural networks. Advances in neural information processing systems, 31.

zesmae2@uic.edu