EECE695D : Deep Learning Theory

Linearization Explains Fine-Tuning in Large Language Models

NeurIPS 2025

Group 1

JoonSeok Kim

Yong Jun Kim

POSTECH

Section 1 : Motivation

Motivation

- PEFT : Technique to reduce the effective number of trained parameters
    - Choices : Update only a few layers, apply rank-limited updates

- Objective of research : Proximity of the fine-tuned models to the pretrained models promotes linearity. Based on this one can predict performance of various fine-tuning decisions using the properties of NTK kernel

- Idea 1. Fine-tuned models are encouraged to remain close to the pretrained model
- Idea 2. This restriction (regularization) can achieve linearization
- Idea 3. Through this linearization, we can learn fine-tuning using the NTK kernel properties
  → Linearized fine-tuning

- What the paper shows compared to previous works
    - Quantify the extent to which linearity is preserved during fine-tuning
    - Theoretical upper bound on the distance between the fine-tuned model and its linearized approximation

Section 2 : Problem Formulation

Problem Formulation

- Pretrained model : $f_{\theta_0}(\cdot)$
- Target task dataset (downstream) : $D_T = (x_i, y_i)_{i=1}^{n}$
- Loss function : $\mathcal{L}(\cdot,\cdot) : \mathbb{R} \times \mathbb{R} \rightarrow \mathbb{R}$
- Regularized fine-tuned model : $f_{\theta^*}(\cdot): \mathbb{R}^d \rightarrow \mathbb{R}$
- Trainable fine-tuning parameters : θ
- Parameters of the pre-trained model : θ_0
- Original objective function : $\widetilde{R}(\theta)$
- Regularization strength hyperparameter : λ

- Explicit inductive bias toward the pretrained model
  - Fine-tuned models are closed to the pretrained model
  - Regularization reduces deviation between fine-tuned and pretrained models (proximal method)
  - $\theta^* = \min_{\theta}[\widetilde{R}(\theta) + \frac{\lambda}{2}\|\theta - \theta_0\|_2^2]$
    - $\widetilde{R}(\theta) = \sum_{i=1}^{n} \mathcal{L}(f_\theta(x_i), y_i) = \|f_\theta(x^*) - y^*\|^2$ (MSE loss)

- Split optimization equation
  - $\widetilde{\theta}_t = \theta_{t-1} - \eta \nabla \widetilde{R}(\theta_{t-1})^\top$: Gradient descent on R
  - $\theta_t = \widetilde{\theta}_t - \lambda\eta(\widetilde{\theta}_t - \theta_0)$: Gradient descent on $\frac{\lambda}{2}\|\theta - \theta_0\|_2^2$

Problem Formulation : Additional Slide

- Why regularize with $\frac{\lambda}{2}\|\theta - \theta_0\|_2^2$?

→ Explicit bias that promotes proximity

- Why keep it close?

→ If they are close, we can use linear characteristics

- Why use linear characteristics?

→ Cuz you can use NTK kernels if linear

- Why use NTK kernels?

→ We can use its properties to predict performance of fine-tuned models

Target Dataset
$D_T = (x_i, y_i)_{i=1}^n$

1. Pre-trained Model
$f_{\theta_0}(\cdot)$

2. Fine-tuned Model
$f_{\theta^*}(\cdot)$

Optimize Loss Function for θ^*

3. Original Objective Function (MSE)
$\widetilde{R}(\theta) = \|f_\theta(x^*) - y^*\|^2$

4. Regularization
$\frac{\lambda}{2}\|\theta - \theta_0\|_2^2$

Split

5. Gradient Descent on R
$\widetilde{\theta}_t = \theta_{t-1} - \eta\nabla\widetilde{R}(\theta_{t-1})^\top$

6. Gradient Descent on Regularization
$\theta_t = \widetilde{\theta}_t - \lambda\eta(\widetilde{\theta}_t - \theta_0)$

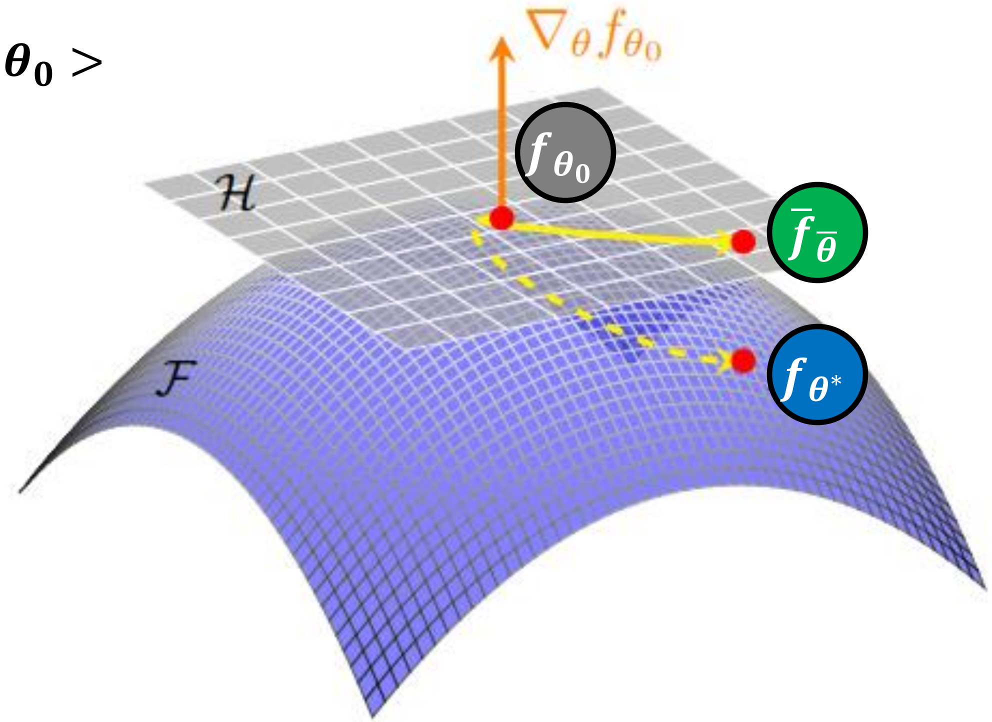
Afzal, Zahra Rahimi, et al. "Linearization Explains Fine-Tuning in Large Language Models." *The 39th Annual Conference on NeurIPS*.

Section 3 : Proximity Promotes Linearity

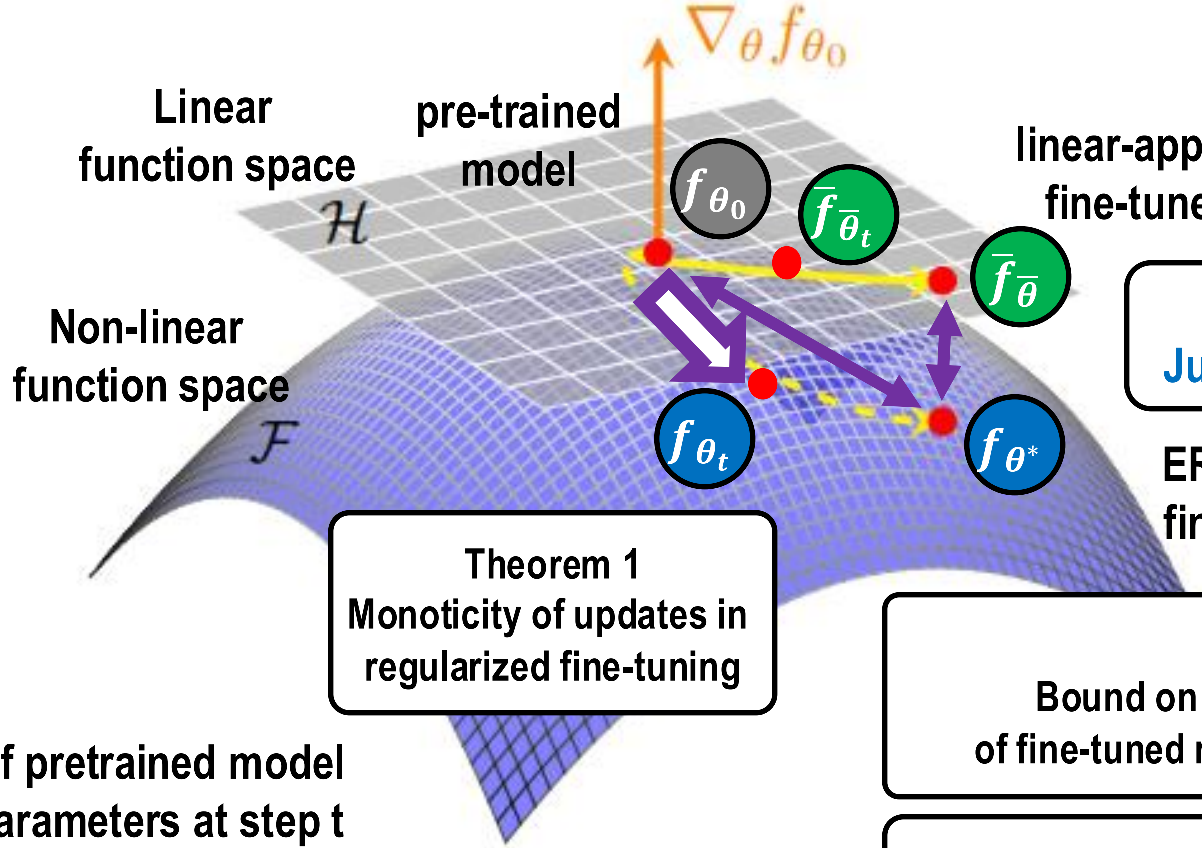
Proximity to the Pretrained Model Promotes Linearity

- The authors need to justify proximity of the fine-tuned model to the pretrained model
  - Show the similarity between final fine-tuned solution (answer) & linearized counterpart (objective)

- Linearized fine-tuned model : $\bar{f}_{\bar{\theta}_t}(x) = f_{\theta_0}(x) + < \nabla f_{\theta_0}(x), \bar{\theta}_t - \theta_0 >$
  - Non-linear function space : $\mathcal{F}$
  - Linear function space : $\mathcal{H}$
    - Defined by NTK
  - Initial pretrained model : f_{θ_0}
  - Linearized fine-tuned model : $\bar{f}_{\bar{\theta}_t}$ (at step t)
    - On $\mathcal{H}$
  - Fine-tuned model by ERM : f_{θ^*}
    - On $\mathcal{F}$
- If fine-tuning remains in the linearized regime, then the following is a good approximation
  - $f_{\theta^*}(x) \approx f_{\theta_0}(x) + < \nabla_\theta f_{\theta_0}(x), \bar{\theta}_t - \theta_0 >$



Proximity : Additional Slide



Objective :
Justify that these are close

Theorem 1
Monoticity of updates in regularized fine-tuning

Theorem 2
Bound on distance between parameters (θ) of fine-tuned model (θ_t) and pretrained model (θ_0)

Theorem 3
Bound on distance between inference of fine-tuned model ($f_\theta(x^*)$) and linearized model ($\bar{f}_{\bar{\theta}}(x^*)$)

- θ_0 : Parameters of pretrained model
- θ_t : Fine-tuning parameters at step t
- θ^* : ERM obtained fine-tuned model params
- $\bar{\theta}$: Parameters of linearized model
- f : Fine-tuned model
- $\bar{f}$: Linearized model

Lemma 3. Gradient Flow

- **Let θ_t be the gradient flow limit of the regularized fine-tuning gradient descent described above. If we assume that λ switches at most countably often and denote instantaneous value by λ_t, the θ_t satisfies the following differential equation**

$$\frac{d}{dt}\theta_t = -\nabla_\theta \widetilde{R}(\theta_t)^\top - \lambda_t(\theta_t - \theta_0)$$

Theorem 1. Monotonic Updates in Regularized Fine-tuning

- **Under the squared loss, for any $t > 0$, if $\lambda > 0$ and $\nabla_\theta \widetilde{R}(\theta_t)(\theta_t - \theta_0) \geq 0$, then $\frac{d}{dt}\left\| f_{\theta_t}(x^*) - y^* \right\|^2 \leq 0$**
- **Moreover, if $\nabla_\theta \widetilde{R}(\theta_t)(\theta_t - \theta_0) < 0$, then $\lambda = 0$ is a sufficient condition for the above equation to hold.**

- Monotonicity of updates → Training fine-tuned models decreases loss
- How to show that loss decreases → Show that the derivative of the loss function surface is negative
  - $\frac{d}{dt}\underbrace{\left\| f_{\theta_t}(x^*) - y^* \right\|^2}_{\text{MSE Loss } \tilde{R}(\theta)} \leq 0$

- (1) $\frac{d}{dt} y(t) = \nabla f_{\theta_t}(x^*) \frac{d}{dt}\theta_t \ \left(\because y(t) = f_{\theta_t}(x^*), at\ step\ t\right)$
- $= -2\nabla f_{\theta_t}(x^*)\nabla f_{\theta_t}^\top(x^*)(y(t) - y^*) - \lambda \nabla f_{\theta_t}(x^*)(\theta_t - \theta_0)$ $(\because Lemma\ 3 : \frac{d}{dt}\theta_t = -\nabla_\theta \tilde{R}(\theta_t)^\top - \lambda_t(\theta_t - \theta_0))$
- $= -2\nabla f_{\theta_t}(x^*)\nabla f_{\theta_t}^\top(x^*)(y(t) - y^*) - \lambda \nabla f_{\theta_t}(x^*)(\theta_t - \theta_0)$
- $= -(2k_t(y(t) - y^*) + \lambda \nabla f_{\theta_t}(x^*)(\theta_t - \theta_0))$ $(\because Kernel : k_t = k_t(x^*, x^*))$

Theorem 1. Monotonic Updates in Regularized Fine-tuning

- **Under the squared loss, for any $t > 0$, if $\lambda > 0$ and $\nabla_\theta \widetilde{R}(\theta_t)(\theta_t - \theta_0) \geq 0$, then $\frac{d}{dt}\left\|f_{\theta_t}(x^*) - y^*\right\|^2 \leq 0$**
- **Moreover, if $\nabla_\theta \widetilde{R}(\theta_t)(\theta_t - \theta_0) < 0$, then $\lambda = 0$ is a sufficient condition for the above equation to hold.**

- Monotonicity of updates → Training fine-tuned models decreases loss
- How to show that loss decreases → Show that the derivative of the loss function surface is negative
  - $\frac{d}{dt}\left\|f_{\theta_t}(x^*) - y^*\right\|^2 \leq 0$

    MSE Loss $\widetilde{R}(\theta)$

- (1) $\frac{d}{dt}y(t) = -(2k_t(y(t) - y^*) + \lambda \nabla f_{\theta_t}(x^*)(\theta_t - \theta_0))$
- (2) $\frac{1}{2}\frac{d}{dt}\|y(t) - y^*\|_2^2 = \frac{1}{2}\frac{d}{dt}\left((y(t) - y^*)^\top (y(t) - y^*)\right) = (y(t) - y^*)^\top \frac{d}{dt}y(t)$
- $= -2(y(t) - y^*)^\top k_t(y(t) - y^*) - \lambda(y(t) - y^*)^\top \nabla f_{\theta_t}(x^*)(\theta_t - \theta_0))$
- $= -2\langle k_t(y(t) - y^*), (y(t) - y^*)\rangle - \lambda(y(t) - y^*)^\top \nabla f_{\theta_t}(x^*)(\theta_t - \theta_0))$
- $= -2\langle k_t(y(t) - y^*), (y(t) - y^*)\rangle - \left(\frac{\lambda}{2}\right) \nabla_\theta \tilde{R}(\theta_t)(\theta_t - \theta_0)$ ($\because \nabla_\theta \tilde{R}(\theta_t) = 2(y(t) - y^*)^\top \nabla f_{\theta_t}(x^*)$)
- $\therefore \frac{1}{2}\frac{d}{dt}\|y(t) - y^*\|_2^2 = -2\langle k_t(y(t) - y^*), (y(t) - y^*)\rangle - \left(\frac{\lambda}{2}\right) \nabla_\theta \tilde{R}(\theta_t)(\theta_t - \theta_0)$

  ≥ 0 ($\because k_t = \nabla f_{\theta_t}(x^*) \nabla f_{\theta_t}^\top(x^*)$ is semidefinite)

  - $\nabla_\theta \tilde{R}(\theta_t)(\theta_t - \theta_0) \geq 0$ required to keep $\frac{1}{2}\frac{d}{dt}\|y(t) - y^*\|_2^2 \leq 0$
  - If $\nabla_\theta \tilde{R}(\theta_t)(\theta_t - \theta_0) < 0$ then $\lambda = 0$ is necessary

Theorem 1. Monotonic Updates in Regularized Fine-tuning

- **Under the squared loss, for any $t > 0$, if $\lambda > 0$ and $\nabla_\theta \tilde{R}(\theta_t)(\theta_t - \theta_0) \geq 0$, then $\frac{d}{dt}\left\|f_{\theta_t}(x^*) - y^*\right\|^2 \leq 0$**
- **Moreover, if $\nabla_\theta \tilde{R}(\theta_t)(\theta_t - \theta_0) < 0$, then $\lambda = 0$ is a sufficient condition for the above equation to hold.**

- What does this imply?
- The selective regularization scheme, where
- $\theta_t = \begin{cases} \tilde{\theta}_t - \lambda\eta\left(\tilde{\theta}_t - \theta_0\right) & if \;\; \nabla_\theta \tilde{R}(\theta_t)(\theta_t - \theta_0) \geq 0 \\ \tilde{\theta}_t & if \;\; \nabla_\theta \tilde{R}(\theta_t)(\theta_t - \theta_0) < 0 \end{cases}$
- (→ gradient descent on the regularization term $\theta_t = \tilde{\theta}_t - \lambda\eta\left(\tilde{\theta}_t - \theta_0\right)$)
- A non-increasing $\left\|f_{\theta_t}(x^*) - y^*\right\|^2$ is guaranteed at step t
- Meaning, the model is guaranteed to learn with decreasing error

Intuition : $\nabla_\theta \tilde{R}(\theta_t)(\theta_t - \theta_0)$
- How aligned the current progress is with the negative gradient update direction
- The positive inner product suggests movement toward a higher loss region, and $\tilde{\theta}_t - \lambda\eta\left(\tilde{\theta}_t - \theta_0\right)$ is applied

Lemma 1. Norm Bounds and Lipschitzness

- **If for all θ in a given range, $f_{\theta}(x^*)$ is $Lip(f) - Lipschitz$ in θ, and $\nabla f_{\theta}(x^*)$ is $Lip(\nabla f) - Lipschitz$ in θ, then $k_{\theta}(x^*, x^*) = \nabla f_{\theta}(x^*)\nabla f_{\theta}(x^*)^{\top} \in \mathbb{R}^{n\times n}$ is $Lip(k) - Lipschitz$ in θ, with**

$$Lip(k) \leq 2Lip(f)\ Lip(\nabla f)$$

Theorem 2. UB on Distance between Parameters

- **Consider the selectively regularized fine-tuning solution, under the squared loss. Denote the instantaneous value of the regularization parameter by $\boldsymbol{\lambda_t}$, which can be either 0 or $\boldsymbol{\lambda}$. If $\boldsymbol{f_\theta(x^*)}$ is $\boldsymbol{Lip(f)-Lipschitz}$ in an $\boldsymbol{l_2-ball}$ of radius $\boldsymbol{\tau}$ around pretrained parameters $\boldsymbol{\theta_0}$,**
- $\boldsymbol{\|\theta-\theta_0\| \leq 2\, Lip(f)\left\|f_{\theta_t}(x^*)-y^*\right\| \int_0^t e^{-(\Lambda_t-\Lambda_s)} ds}$, **where** $\boldsymbol{\Lambda_t = \int_0^t \lambda_s ds}$

- $\nabla_\theta \tilde{R}(\theta_t) = 2(y(t)-y^*)^\top \nabla f_{\theta_t}(x^*)$
- Lemma 3 : $\frac{d}{dt}\theta_t = -\nabla_\theta \tilde{R}(\theta_t)^\top - \lambda_t(\theta_t-\theta_0) = -2\nabla f_{\theta_t}(x^*)^\top (y(t)-y^*) - \lambda_t(\theta_t-\theta_0)$

- $\frac{d}{dt}\theta_t = -2\nabla f_{\theta_t}(x^*)^\top(y(t)-y^*) - \lambda_t(\theta_t-\theta_0)$
- $\frac{d}{dt}u_t = -2\nabla f_{\theta_t}(x^*)^\top(y(t)-y^*) - \lambda_t u_t$ where $u_t = (\theta_t-\theta_0)$

- Given $w(t) = \|u_t\|_2, \hat{u}_t = u_t/\|u_t\|_2$ → Show the bounds of $w(t)$
- $\dot{w}(t) = \frac{u_t^\top}{\|u_t\|_2}\frac{d}{dt}u_t = -2\hat{u}_t^\top \nabla f_{\theta_t}(x^*)^\top(y(t)-y^*) - \lambda_t w(t) \quad \left(\because \left(\frac{u_t^\top}{\|u_t\|_2}\right)u_t = \left(\frac{\|u_t\|_2^2}{\|u_t\|_2}\right) = \|u_t\|_2 = w(t)\right)$

Theorem 2. UB on Distance between Parameters

- **Consider the selectively regularized fine-tuning solution, under the squared loss. Denote the instantaneous value of the regularization parameter by $\boldsymbol{\lambda_t}$, which can be either 0 or $\boldsymbol{\lambda}$. If $\boldsymbol{f_\theta(x^*)}$ is $\boldsymbol{Lip(f) - Lipschitz}$ in an $\boldsymbol{l_2 - ball}$ of radius $\boldsymbol{\tau}$ around pretrained parameters $\boldsymbol{\theta_0}$,**
- $\boldsymbol{\|\theta - \theta_0\| \le 2\, Lip(f) \left\| f_{\theta_t}(x^*) - y^* \right\| \int_0^t e^{-(\Lambda_t - \Lambda_s)} ds}$, **where** $\boldsymbol{\Lambda_t = \int_0^t \lambda_s ds}$

- $\dot{w}(t) = -2\underbrace{\hat{u}_t^\top \nabla f_{\theta_t}(x^*)^\top (y(t) - y^*)} - \lambda_t w(t)$
- $-\hat{u}_t^\top \nabla f_{\theta_t}(x^*)^\top (y(t) - y^*) \le \left\| f_{\theta_t}(x^*) \right\| \|y(t) - y^*\|_2 \le Lip(f) \|y(t) - y^*\|_2$

  $\left(\because \boldsymbol{\left\| f_{\theta_t}(x^*) \right\| \le Lip(f)}\right)$ → **Lemma 1**
- $\therefore -2\hat{u}_t^\top \nabla f_{\theta_t}(x^*)^\top (y(t) - y^*) \le 2Lip(f)\|y(t) - y^*\|_2$
- Hence, $\dot{w}(t) \le -\lambda_t w(t) + 2\, Lip(f)\|y(t) - y^*\|_2 \rightarrow \dot{w}(t) + \lambda_t w(t) \le 2\, Lip(f)\|y(t) - y^*\|_2$

Theorem 2. UB on Distance between Parameters

- **Consider the selectively regularized fine-tuning solution, under the squared loss. Denote the instantaneous value of the regularization parameter by $\boldsymbol{\lambda_t}$, which can be either 0 or $\boldsymbol{\lambda}$. If $\boldsymbol{f_\theta(x^*)}$ is $\boldsymbol{Lip(f)-Lipschitz}$ in an $\boldsymbol{l_2-ball}$ of radius $\boldsymbol{\tau}$ around pretrained parameters $\boldsymbol{\theta_0}$,**
- $\boldsymbol{\|\theta-\theta_0\| \le 2\,Lip(f)\left\|f_{\theta_t}(x^*)-y^*\right\| \int_0^t e^{-(\Lambda_t-\Lambda_s)}ds}$**, where** $\boldsymbol{\Lambda_t = \int_0^t \lambda_s ds}$

- $\dot{w}(t) + \lambda_t w(t) \le 2\,Lip(f)\|y(t)-y^*\|_2$
- Let $\Lambda_t = \int_0^t \lambda_\tau d\tau$ → Multiply by e^{Λ_t} : $\frac{d}{dt}(e^{\Lambda_t}w(t)) \le 2\,Lip(f)e^{\Lambda_t}\|y(t)-y^*\|_2$
  - $\frac{d}{dt}\left(e^{\Lambda_t}w(t)\right) = w(t)\frac{d}{dt}(e^{\Lambda_t}) + e^{\Lambda_t}\frac{d}{dt}w(t) = e^{\Lambda_t}\lambda_t w(t) + e^{\Lambda_t}\dot{w}(t)$

$\int_0^t$

- $\int_0^t \frac{d}{da}(e^{\Lambda_a}w(a))da = e^{\Lambda_t}w(t) - w(0) \le 2\,Lip(f)\int_0^t e^{\Lambda_s}\|y(s)-y^*\|_2 ds$
- $w(t) \le e^{-\Lambda_t}w(0) + 2\,Lip(f)\int_0^t e^{-(\Lambda_t-\Lambda_s)}\|y(s)-y^*\|_2 ds \le e^{-\Lambda_t}w(0) + 2\,Lip(f)\|y(0)-y^*\|_2\int_0^t e^{-(\Lambda_t-\Lambda_s)}ds$
- $(\because Theorem\ 1 : \frac{d}{dt}\left\|f_{\theta_t}(x^*)-y^*\right\|^2 \le 0 \;\rightarrow \|y(s)-y^*\|_2 \le \|y(0)-y^*\|_2, non-increasing\ error)$

Theorem 2. UB on Distance between Parameters

- **Consider the selectively regularized fine-tuning solution, under the squared loss. Denote the instantaneous value of the regularization parameter by $\boldsymbol{\lambda_t}$, which can be either 0 or $\boldsymbol{\lambda}$. If $\boldsymbol{f_\theta(x^*)}$ is $\boldsymbol{Lip(f) - Lipschitz}$ in an $\boldsymbol{l_2 - ball}$ of radius $\boldsymbol{\tau}$ around pretrained parameters $\boldsymbol{\theta_0}$,**
- $\boldsymbol{\|\theta_t - \theta_0\| \leq 2\, Lip(f) \left\| f_{\theta_t}(x^*) - y^* \right\| \int_0^t e^{-(\Lambda_t - \Lambda_s)} ds}$**, where** $\boldsymbol{\Lambda_t = \int_0^t \lambda_s ds}$

- $w(t) \leq e^{-\Lambda_t} w(0) + 2\, Lip(f) \|y(0) - y^*\|_2 \int_0^t e^{-(\Lambda_t - \Lambda_s)} ds$

- Case 1: $\lambda_t = \lambda > 0 \ \rightarrow t\lambda$
- $\|u_t\|_2 \leq e^{-\Lambda_t} \|u_0\|_2 + 2\, Lip(f) \|y(0) - y^*\|_2 \int_0^t e^{-\lambda(t-s)} ds \leq 2\, Lip(f) \|y(0) - y^*\|_2 \frac{1 - e^{-\lambda t}}{\lambda}$
- $\|\theta_t - \theta_0\|_2 \leq 2\, Lip(f) \left\| f_{\theta_t}(x^*) - y^* \right\|_2 \frac{1 - e^{-\lambda t}}{\lambda}$
- $\lim_{t \to \infty} sup\ \|\theta_t - \theta_0\|_2 \leq 2\, Lip(f) \left\| f_{\theta_t}(x^*) - y^* \right\|_2 (1/\lambda)$

- Case 2 : $\lambda_t = 0$
- $\|u_t\|_2 \leq \|u_0\|_2 + 2\, Lip(f) \|y(0) - y^*\|_2 \int_0^t e^0 ds$
- $\|\theta_t - \theta_0\|_2 \leq 2\, Lip(f) \left\| f_{\theta_t}(x^*) - y^* \right\| t$

- The fine-tuning solution deviates from the origin under regularization : $\|\theta - \theta_0\|$
- But this deviation can be bounded, by the UB $2\, Lip(f) \left\| f_{\theta_t}(x^*) - y^* \right\| \int_0^t e^{-(\Lambda_t - \Lambda_s)} ds$

Theorem 3. UB on Distance between Parameters

- **Under the squared loss, if $f_{\theta_t}(x^*)$ and $\nabla f_{\theta_t}(x^*)$ are $Lip(f)-Lipschitz$ and $Lip(\nabla f)-Lipschitz$ in an l_2-ball of radius τ around pretrained parameters θ_0,**
- $\left\|f_{\theta_t}(x^*)-\bar{f}_{\bar{\theta}_t}(x^*)\right\| \le b\left(t-\frac{1-e^{-\lambda t}}{\lambda}\right)$**, where** $b=2Lip(f)^2\left\|f_{\theta_0}(x^*)-y^*\right\|(\frac{4}{\lambda}Lip(\nabla f)\left\|f_{\theta_0}(x^*)-y^*\right\|+1)$

- $y(t) = f_{\theta_t}(x^*),\ \bar{y}(t) = \bar{f}_{\bar{\theta}_t}(x^*) \rightarrow \Delta(t) = \|y(t)-\bar{y}(t)\|_2$
- $\frac{1}{2}\frac{d}{dt}\Delta(t)_2^2 = \frac{1}{2}\frac{d}{dt}\|y(t)-\bar{y}(t)\|_2^2 = \frac{1}{2}\langle y'(t)-\bar{y}'(t), y(t)-\bar{y}(t)\rangle + \frac{1}{2}\langle y(t)-\bar{y}(t), y'(t)-\bar{y}'(t)\rangle$
- $= \langle y'(t) - \bar{y}'(t), y(t)-\bar{y}(t)\rangle = \left\langle -k_t 2(y(t)-y^*) - \lambda\nabla f_{\theta_t}(x^*)(\theta_t-\theta_0) + k_0 2(\bar{y}(t)-y^*), y(t)-\bar{y}(t)\right\rangle$
  - Theorem 1 : $\frac{d}{dt}y(t) = -(2k_t(y(t)-y^*) + \lambda\nabla f_{\theta_t}(x^*)(\theta_t-\theta_0))$

    $\frac{d}{dt}\bar{y}(t) = -k_0 2(\bar{y}(t)-y^*) - \lambda\nabla\bar{f}_{\bar{\theta}_t}(x^*)(\theta_t-\theta_0))$

- Simplification : $-k_t 2(y(t)-y^*) + k_0 2(\bar{y}(t)-y^*)$
- $= -k_t 2(y(t)-y^*) + k_0 2(y(t)-y^*) - k_0 2(y(t)-y^*) + k_0 2(\bar{y}(t)-y^*)$
- $= (k_0-k_t)2(y(t)-y^*) + k_0(2(\bar{y}(t)-y^*) - 2(y(t)-y^*))$

Theorem 3. UB on Distance between Parameters

- **Under the squared loss, if $f_{\theta_t}(x^*)$ and $\nabla f_{\theta_t}(x^*)$ are $Lip(f)-Lipschitz$ and $Lip(\nabla f)-Lipschitz$ in an l_2-ball of radius τ around pretrained parameters θ_0,**
- $\|f_{\theta_t}(x^*)-\bar{f}_{\bar{\theta}_t}(x^*)\| \le b\left(t-\frac{1-e^{-\lambda t}}{\lambda}\right)$**, where** $b = 2Lip(f)^2\|f_{\theta_0}(x^*)-y^*\|(\frac{4}{\lambda}Lip(\nabla f)\|f_{\theta_0}(x^*)-y^*\|+1)$

- $\frac{1}{2}\frac{d}{dt}\Delta(t)_2^2 = \left\langle -k_t 2(y(t)-y^*) - \lambda\nabla f_{\theta_t}(x^*)(\theta_t-\theta_0) + k_0 2(\bar{y}(t)-y^*), y(t)-\bar{y}(t)\right\rangle$
- $-k_t 2(y(t)-y^*) + k_0 2(\bar{y}(t)-y^*) = (k_0-k_t)2(y(t)-y^*) + k_0(2(\bar{y}(t)-y^*)-2(y(t)-y^*))$

- Substitution : $\frac{1}{2}\frac{d}{dt}\Delta(t)_2^2 = \left\langle (k_0-k_t)2(y(t)-y^*) - \lambda\nabla f_{\theta_t}(x^*)(\theta_t-\theta_0), y(t)-\bar{y}(t)\right\rangle + \langle k_0(2(\bar{y}(t)-y^*) - 2(y(t)-y^*)), y(t)-\bar{y}(t)\rangle$
  - $\langle k_0(2(\bar{y}(t)-\cancel{y^*})-2(y(t)-\cancel{y^*})), y(t)-\bar{y}(t)\rangle = -2\big(y(t)-\bar{y}(t)\big)^\top k_0(y(t)-\bar{y}(t)) \le 0$
  - ($\because$ $\nabla f_{\theta_0}(x^*)^\top \nabla f_{\theta_0}(x^*)^\top$ is positive semidefinite)

- $\therefore \frac{1}{2}\frac{d}{dt}\Delta(t)_2^2 \le \left\langle (k_0-k_t)2(y(t)-y^*) - \lambda\nabla f_{\theta_t}(x^*)(\theta_t-\theta_0), y(t)-\bar{y}(t)\right\rangle$

Theorem 3. UB on Distance between Parameters

- Under the squared loss, if $f_{\theta_t}(x^*)$ and $\nabla f_{\theta_t}(x^*)$ are $Lip(f) - Lipschitz$ and $Lip(\nabla f) - Lipschitz$ in an $l_2 - ball$ of radius τ around pretrained parameters θ_0,
- $\|f_{\theta_t}(x^*) - \bar{f}_{\bar{\theta}_t}(x^*)\| \leq b\left(t - \frac{1-e^{-\lambda t}}{\lambda}\right)$, where $b = 2Lip(f)^2\|f_{\theta_0}(x^*) - y^*\|(\frac{4}{\lambda}Lip(\nabla f)\|f_{\theta_0}(x^*) - y^*\| + 1)$

- $\frac{d}{dt}\Delta(t) = \|(k_0 - k_t)2(y(t) - y^*) - \lambda\nabla f_{\theta_t}(x^*)(\theta_t - \theta_0), y(t) - \bar{y}(t)\|$
- $\leq \|(k_0 - k_t)2(y(t) - y^*)\| - \lambda\|\nabla f_{\theta_t}(x^*)(\theta_t - \theta_0), y(t) - \bar{y}(t)\|$ $\because Triangular\ \ inequality)$
- $\leq \|(k_0 - k_t)2(y(t) - y^*)\| - \lambda\|\nabla f_{\theta_t}(x^*)(\theta_t - \theta_0), y(t) - \bar{y}(t)\|$
- $\leq Lip(k)\|(\theta_t - \theta_0)\|\|2(y(t) - y^*)\| + \lambda Lip(f)\|\theta_t - \theta_0\|$ ($\because k_\theta(x^*, x^*)$ is $Lip(k) - Lipschitz$)
- $\leq 2Lip(k)\|(\theta_t - \theta_0)\|\|(y(0) - y^*)\| + \lambda Lip(f)\|\theta_t - \theta_0\|$
- $\leq 4Lip(k)Lip(f)\|(y(0) - y^*)\|^2\frac{(1-e^{-\lambda t})}{\lambda} + 2Lip^2(f)\|y(0) - y^*\|(1 - e^{-\lambda t})$ ($\because Theorem\ 2$)
- $\leq -2Lip(f)\|(y(0) - y^*)\|\left(\frac{2}{\lambda}Lip(k)\|y(0) - y^*\| + Lip(f)\right)(e^{-\lambda t})$
- $+2Lip(f)\|(y(0) - y^*)\|\left(\frac{2}{\lambda}Lip(k)\|y(0) - y^*\| + Lip(f)\right)$

Theorem 3. UB on Distance between Parameters

- **Under the squared loss, if $f_{\theta_t}(x^*)$ and $\nabla f_{\theta_t}(x^*)$ are $\boldsymbol{Lip(f) - Lipschitz}$ and $\boldsymbol{Lip(\nabla f) - Lipschitz}$ in an $\boldsymbol{l_2 - ball}$ of radius $\boldsymbol{\tau}$ around pretrained parameters $\boldsymbol{\theta_0}$,**
- $\boldsymbol{\left\|f_{\theta_t}(x^*) - \bar{f}_{\bar{\theta}_t}(x^*)\right\| \le b\left(t - \frac{1-e^{-\lambda t}}{\lambda}\right)}$**, where** $\boldsymbol{b = 2Lip(f)^2\left\|f_{\theta_0}(x^*) - y^*\right\|(\frac{4}{\lambda}Lip(\nabla f)\left\|f_{\theta_0}(x^*) - y^*\right\| + 1)}$

- $\frac{d}{dt}\Delta(t) \le -2Lip(f)\|(y(0) - y^*)\|\left(\frac{2}{\lambda}Lip(k)\|y(0) - y^*\| + Lip(f)\right)(1 - e^{-\lambda t})$
  - $-2Lip(f)\|(y(0) - y^*)\|\left(\frac{2}{\lambda}Lip(k)\|y(0) - y^*\| + Lip(f)\right) \le -2Lip(f)\|(y(0) - y^*)\|\left(\frac{4}{\lambda}Lip(\nabla f)Lip(f)\|y(0) - y^*\| + Lip(f)\right)$ $(\because Lemma\ 1)$

$$= 2Lip(f)^2\|y(0) - y^*\|\left(\frac{4}{\lambda}Lip(\nabla f)\|y(0) - y^*\| + 1\right)(1 - e^{-\lambda t})$$

- *Using Lemma 1 :* $\frac{d}{dt}\Delta(t) \le b - be^{-\lambda t}$, *where* $b = 2Lip(f)^2\|y(0) - y^*\|\left(\frac{4}{\lambda}Lip(\nabla f)\|y(0) - y^*\| + 1\right)$
- $\therefore \Delta(t) \le b(t + \frac{1}{\lambda}e^{-\lambda t} - \frac{1}{\lambda})$

- Intuition : For a proper choice of the regularization parameter λ, linearization of the fine-tuning only depends on the local properties of $f_{\theta_t}(x^*)$ around θ_0

Theorem 4.

- Under the squared loss, if $f_{\theta_t}(x^*)$ and $\nabla f_{\theta_t}(x^*)$ are $Lip(f)-Lipschitz$ and $Lip(\nabla f)-Lipschitz$ in an l_2-ball of radius r around pretrained parameters θ_0.
- Define $\lambda_\circ = \frac{2\left\|f_{\theta_0}(x^*)-y^*\right\| Lip(f)}{r}$
- If $\lambda \geq \lambda_\circ$, then for all t, the following holds
- If $\lambda < \lambda_\circ$, then the following holds for $t \leq \frac{1}{\lambda} ln(\frac{1}{1-\lambda/\lambda_\circ})$
- In particular, for $\lambda \geq \lambda_\circ$, the bound from theorem 3 always holds and simplifies to
- $\left\|f_{\theta_t}(x^*) - \bar{f}_{\bar{\theta}_t}(x^*)\right\| \leq 2Lip(f)\tilde{R}(\theta_0)\big(2rLip(\nabla f) + Lip(f)\big)t$

- The bound "$2Lip(f)\tilde{R}(\theta_0)\big(2rLip(\nabla f) + Lip(f)\big)t$"
  - Holds when λ is large → But it also implies that r needs to be small, where the parameters need to be close to the pretrained parameters
  - If λ is small, the bound holds when the time step is smaller than a certain value

The authors provide with a guide to deciding an appropriate regularization parameter $\lambda_\circ$
- However, certain inquiries arise with respect to the bounding value

Section 4 : Fine-Tuning Meets NTK Regression

Fine-tuning can be defined as RKHS

- **As so far, we showed that regularized fine-tuning is in the linearized regime.**
- **Linearized regime? NTK!**

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\text{minimize}}\, \widetilde{\mathcal{R}}(\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2, \quad \text{where}$$

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

$$\Rightarrow$$

$$\underset{\alpha}{\text{minimize}}\; \frac{1}{n}\sum_i \left|a^\top K(:, x^*) - y^*\right|^2 + \sigma\, \alpha^\top K \alpha$$

Regularized fine-tuning → Kernel ridge regression problem

Fine-tuning can be defined as RKHS

- **As so far, we showed that regularized fine-tuning is in the linearized regime.**
- **Linearized regime? NTK!**

If the model is linear, the problem is equivalent to solve ridge regression problem with feature $\phi(x)$

Original problem : $min_\theta \frac{1}{n}\sum_i | f_\theta(x_i) - y_i|^2 + \frac{\lambda}{2} | \theta - \theta_0|^2$

We have : $f_\theta(x) \approx f_{\theta_0}(x) + \nabla_\theta \mathrm{f}_{\theta_0}(\mathrm{x})^\top(\theta - \theta_0) = f_{\theta_0}(x) + \phi(x)^\top w,\ w := \theta - \theta_0$

The problem is approximately equivalent to solve

$$min_w \frac{1}{n}\sum_i (\phi(x_i)^\top w - \tilde{y}_i)^2 + \sigma|w|^2$$

where $\tilde{y}_i = y_i - f_{\theta_0}(x_i)$, and $\sigma = \lambda/2$. This is actually ridge regression problem with feature $\phi(x_i)$.

Optimal parameter of MSE + weight decay objective

Let's consider a MSE+weight decay optimization problem of model $\boldsymbol{f_w(x) = \phi(x)^\top w}$, where $\boldsymbol{\phi(x)}$ is a feature vector of input $\boldsymbol{x}$.

$$min_w \frac{1}{n} \sum_i (\phi(x_i)^\top w - \tilde{y}_i)^2 + \sigma |w|^2$$

Then, the optimal parameter is on the space spanned by the feature vectors $\boldsymbol{\phi(x_i)}$,

$$\boldsymbol{w^* = \sum_i a_i \phi(x_i)}$$

1. The parameters can be divided into $w = w_{||} + w_\perp$, where $w_{||}$ is on the space spanned by features ϕ, and $w_\perp$ is perpendicular to that space.

2. $\phi(x_i)^\top w = \phi(x_i)^\top (w_{||} + w_\perp) = \phi(x_i)^\top w_{||}$, so the $w_\perp$ does not affect to the model output.

3. In optimal parameter, the regularizer make $|w_\perp|^2 = 0$. If not, it is not optimal.

4. Therefore, the optimal parameter is spanned by feature vectors, so $w^* = \sum_i a_i \phi(x_i)$

Fine-tuning can be defined as RKHS

Consider

$$\boldsymbol{w}^* = \sum_{\boldsymbol{i}} \boldsymbol{a_i \phi(x_i)}\ , \boldsymbol{f}(\cdot) = \boldsymbol{\phi(x)}^\top \boldsymbol{w}$$

the optimal function value is

$$f^*(\cdot) = \phi(\cdot) \sum_i a_i \phi(x_i) = \nabla \mathrm{f}_{\theta_0}(\cdot) \sum_i a_i \nabla \mathrm{f}_{\theta_0}(\mathrm{x_i})^\top = \sum_i a_i k(\cdot, x_i) = a^\top K(\cdot, x^*)$$

Where $K(\cdot, x^*) = [k(\cdot, x_1), k(\cdot, x_2), k(\cdot, x_3), \ldots k(\cdot, x_n)] \in R^{1 \times n}$

Substituting original equation yields…

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\text{minimize}}\ \widetilde{\mathcal{R}}(\boldsymbol{\theta}) + \frac{\lambda}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2, \quad \text{where}$$

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

$$\Longrightarrow \quad \underset{\alpha}{\text{minimize}}\ \frac{1}{n} \sum_i \left|a^\top K(:, x^*) - y^*\right|^2 + \sigma\, \alpha^\top K \alpha$$

Fine-tuning can be defined as RKHS

Now, we have to solve linear regression problem with regularizer.

$$\boldsymbol{min_{\alpha} \frac{1}{n} \sum_{i} |a^{\top} K(\cdot, x^*) - y^*|^2 + \sigma\, \alpha^{\top} K \alpha}$$

The solution: $\boldsymbol{\alpha}^* = [\boldsymbol{K}(\boldsymbol{x}^*, \boldsymbol{x}^*) + \boldsymbol{\sigma I}]^{-1} \boldsymbol{y}^*$

Also, the optimal function is

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

where $\boldsymbol{x}^* = [\boldsymbol{x_1}, \boldsymbol{x_2}, \ldots, \boldsymbol{x_n}]^{\top}$, and $\boldsymbol{y}^* = [\boldsymbol{y_1}, \boldsymbol{y_2}, \ldots, \boldsymbol{y_n}]^{\top}$

As the model is linear in regularized fine-tuning, we can acquire optimal function w/o SGD like optimization process

NTK directly affects the empirical risk

Theorem 5: The empirical risk is bounded as

$$\left(\frac{\boldsymbol{\sigma}|\boldsymbol{y}^*|}{\boldsymbol{\sigma} + \boldsymbol{\lambda}_{\boldsymbol{max}}(\boldsymbol{K})}\right)^{\mathbf{2}} \leq \boldsymbol{R}(\boldsymbol{\theta}) \leq \left(\frac{\boldsymbol{\sigma}|\boldsymbol{y}^*|}{\boldsymbol{\sigma} + \boldsymbol{\lambda}_{\boldsymbol{min}}(\boldsymbol{K})}\right)^{\mathbf{2}}$$

where $\boldsymbol{\lambda}_{\boldsymbol{min}}(\boldsymbol{K})$ and $\boldsymbol{\lambda}_{\boldsymbol{max}}(\boldsymbol{K})$ are the minimum and maximum eigenvalues of $\boldsymbol{K}(\boldsymbol{x}^*, \boldsymbol{x}^*)$, repectively.

- The regularized condition number as at-initialization metric for predicting the performance of fine-tuning.

- Let condition number as $\kappa(K + \sigma I) = \frac{\lambda_{max}(K)+\sigma}{\lambda_{min}(K)+\sigma}$. For given $\lambda_{max}(K)$, If κ is small, the risk can be small, and the risk is high otherwise.

- It can be useful in fine-tuning
  - ex) selecting what subset of parameters to tune.
  - We will show this later.

Afzal, Zahra Rahimi, et al. "Linearization Explains Fine-Tuning in Large Language Models." *The 39th Annual Conference on NeurIPS*.

Proof of theorem 5

Let $U\Sigma U^\top$ denote the eigenvalue decomposition of K.

$$
\begin{aligned}
\mathcal{R}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\big(f_\theta(x_i), y_i\big) \approx \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}\big(f_{\bar{\theta}}(x_i), y_i\big) \\
&= \frac{1}{n}\sum_{i=1}^{n}\left\|y_i - \mathbf{K}(x_i, x^*)\left[\mathbf{K}(x^*, x^*) + \sigma\mathbf{I}\right]^{-1}\mathbf{y}^*\right\|_2^2 \\
&= \frac{1}{n}\left\|\mathbf{y}^* - \mathbf{K}(x^*, x^*)\big(\mathbf{K}(x^*, x^*) + \sigma\mathbf{I}\big)^{-1}\mathbf{y}^*\right\|_2^2 \\
&= \frac{1}{n}\left\|\big(\mathbf{I} - \mathbf{K}(x^*, x^*)(\mathbf{K}(x^*, x^*) + \sigma\mathbf{I})^{-1}\big)\mathbf{y}^*\right\|_2^2 \\
&= \frac{1}{n}\left\|\big(\mathbf{I} - \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^\top(\mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^\top + \sigma\mathbf{I})^{-1}\big)\mathbf{y}^*\right\|_2^2 \\
&= \frac{1}{n}\left\|\big(\mathbf{I} - \mathbf{U}\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \sigma\mathbf{I})^{-1}\mathbf{U}^\top\big)\mathbf{y}^*\right\|_2^2 \\
&= \frac{1}{n}\left\|\mathbf{U}\big(\mathbf{I} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \sigma\mathbf{I})^{-1}\big)\mathbf{U}^\top\mathbf{y}^*\right\|_2^2 \\
&= \frac{1}{n}\left\|\big(\mathbf{I} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \sigma\mathbf{I})^{-1}\big)\mathbf{U}^\top\mathbf{y}^*\right\|_2^2.
\end{aligned}
$$

Since $B = \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \boldsymbol{\sigma}\mathbf{I})^{-1}$ is diagonal matrix, we have

$$
\lambda_{min}(I - B)^2\|U^\top y^*\|^2 < R(\theta) < \lambda_{max}(I - B)^2\|U^\top y^*\|^2
$$

Also, Note that

$$
\begin{aligned}
\lambda_{\min}\left(\mathbf{I} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \sigma\mathbf{I})^{-1}\right) &= \frac{\sigma}{\sigma + \lambda_{\max}(\mathbf{K})}, \\
\lambda_{\max}\left(\mathbf{I} - \boldsymbol{\Sigma}(\boldsymbol{\Sigma} + \sigma\mathbf{I})^{-1}\right) &= \frac{\sigma}{\sigma + \lambda_{\min}(\mathbf{K})},
\end{aligned}
$$

Therefore,

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

Section 5 : Spectral Perturbation of Layers

So far…

1. Regularized fine-tuning around initial point is **approximately a linear model**.

2. As it is a linear model, we can **apply NTK kernel ridge regression**.

3. The empirical risk is bounded with NTK gram matrix's (K) eigen values, therefore we can utilize condition number of matrix K **to anticipate performance**.

Let's consider we want to tune only a subset of layers, not the entire model.
Then, how can we select the target layer?
=> NTK approach can give a hint for it.

Afzal, Zahra Rahimi, et al. "Linearization Explains Fine-Tuning in Large Language Models." *The 39th Annual Conference on NeurIPS*.

NTK Eigenvalue Stability under Layer Addition

Theorem 6: Let K be the NTK with respect to the set of selected fine-tuning parameters, and S be the kernel with respect to the parameters of the candidate layers, to add to the fine-tuning parameters. Then,

$$(1-\eta)\lambda_i(K) \leq \lambda_i(K+S) \leq (1+\eta)\lambda_i(K)$$

where $\eta = |K^{-1/2} S K^{-1/2}|$, $[S_l]_{i,j} = \nabla_{\theta_l} f_\theta(x_i) \nabla_{\theta_l} f_\theta(x_j)^\top$

Proof: skip

⇒ Each eigenvalue of K+S is **bounded at most $[1 - \eta, 1 + \eta]$**
⇒ We can get useful insight if we combine this with previous finding:
- ⇒ **If η is small**, we don't have to select that layer, since it **doesn't affect empirical risk much.**
- ⇒ **If η is large** enough and doesn't mass up the condition number, **it is worth to tune.**

Predictive Risk Bound for Layer-wise Fine-tuning

Theorem 7: Let K be the NTK induced by the trainable parameters in θ, then if $\kappa(K+\sigma I) \leq c$, we have

$$\frac{\lambda_{max}(K+S+\sigma I)}{a\,\lambda_{max}(K+\sigma I)} \leq \left(\frac{R(\theta \cup \hat{\theta})}{R(\theta)}\right)^{\frac{1}{2}} \leq \frac{a\lambda_{max}(K+S+\sigma I)}{\lambda_{max}(K+\sigma I)}$$

where $a = \frac{c}{(1-\eta)^2}, \eta = |K^{-1/2}SK^{-1/2}|$ and S is the kernel induced by θ with $[S]_{i,j} = \nabla_{\hat{\theta}} f_{\theta}(x_i) \nabla_{\hat{\theta}} f(x_j)^{\top}$

If we add more candidate layers, how much fine-tuning risk can be improved?

⇒ With theorem 7, we can **bound risk improvement** through maximum eigenvalue of K.

Proof of theorem 7

From Theorem 5, we already know

$$\frac{(\lambda_{\min}(\mathbf{K})+\sigma)^2}{\sigma^2\|\mathbf{y}^*\|^2} \le \frac{1}{\mathcal{R}(\boldsymbol{\theta})} \le \frac{(\lambda_{\max}(\mathbf{K})+\sigma)^2}{\sigma^2\|\mathbf{y}^*\|^2}$$

Also from theorem 6, we know how to bound the existing eigenvalues from K+S matrix's eigenvalue. Therefore, we have

$$\lambda_{\max}(\mathbf{K})+\sigma = \lambda_{\max}(\mathbf{K}+\sigma\mathbf{I}) \le \frac{\lambda_{\max}(\mathbf{K}+\mathbf{S}+\sigma\mathbf{I})}{1-\eta}$$

and

$$\lambda_{\min}(\mathbf{K})+\sigma = \frac{\lambda_{\max}(\mathbf{K}+\sigma\mathbf{I})}{\kappa(\mathbf{K}+\sigma\mathbf{I})} \ge \frac{\lambda_{\max}(\mathbf{K}+\mathbf{S}+\sigma\mathbf{I})}{\kappa(\mathbf{K}+\sigma\mathbf{I})(1+\eta)}$$

by combining the above inequalities, we have

$$\frac{\lambda_{\max}(\mathbf{K}+\mathbf{S}+\sigma\mathbf{I})}{\sigma\|\mathbf{y}^*\|\kappa(\mathbf{K}+\sigma\mathbf{I})(1+\eta)} \le \frac{1}{\mathcal{R}(\boldsymbol{\theta})^{\frac{1}{2}}} \le \frac{\lambda_{\max}(\mathbf{K}+\mathbf{S}+\sigma\mathbf{I})}{\sigma\|\mathbf{y}^*\|(1-\eta)}.$$

Proof of theorem 7

We will do the same thing on K+S case. From theorem 5, we have

$$\frac{\sigma\|\mathbf{y}^*\|}{\lambda_{\max}(\mathbf{K}+\mathbf{S})+\sigma} \leq \mathcal{R}(\boldsymbol{\theta}\cup\hat{\boldsymbol{\theta}})^{\frac{1}{2}} \leq \frac{\sigma\|\mathbf{y}^*\|}{\lambda_{\min}(\mathbf{K}+\mathbf{S})+\sigma}$$

and similarly, from theorem 6 we have

$$\begin{aligned}\lambda_{\min}(\mathbf{K}+\mathbf{S})+\sigma &\geq (1-\eta)\lambda_{\min}(\mathbf{K}+\sigma\mathbf{I}) \\ &= \frac{(1-\eta)\lambda_{\max}(\mathbf{K}+\sigma\mathbf{I})}{\kappa(\mathbf{K}+\sigma\mathbf{I})} \\ \lambda_{\max}(\mathbf{K}+\mathbf{S})+\sigma &\leq (1+\eta)\lambda_{\max}(\mathbf{K}+\sigma\mathbf{I}).\end{aligned}$$

Therefore, combining these we have

$$\frac{\sigma\|\mathbf{y}^*\|}{(1+\eta)\lambda_{\max}(\mathbf{K}+\sigma\mathbf{I})} \leq \mathcal{R}(\boldsymbol{\theta}\cup\hat{\boldsymbol{\theta}})^{\frac{1}{2}} \leq \frac{\sigma\|\mathbf{y}^*\|\kappa(\mathbf{K}+\sigma\mathbf{I})}{(1-\eta)\lambda_{\max}(\mathbf{K}+\sigma\mathbf{I})}$$

Proof of theorem 7

By combining two inequalities, we can obtain

$$\left(\frac{R(\theta \cup \hat{\theta})}{R(\theta)}\right)^{1/2} \geq \frac{\lambda_{max}(K+S+\sigma I)}{\kappa(K+\sigma I)(1+\eta)^2 \lambda_{max}(K+\sigma I)}$$

and

?

$$\left(\frac{R(\theta \cup \hat{\theta})}{R(\theta)}\right)^{1/2} \leq \frac{\lambda_{max}(K+S+\sigma I)\kappa(K+\sigma I)}{(1-\eta)^2 \lambda_{max}(K+\sigma I)}$$

Applying the $\kappa(K+\sigma I) \leq c$, and suppose that $0 \leq \eta \leq 1$, then $(1-\eta)^2 \leq (1+\eta)^{-2}$, by defining $a = \frac{c}{(1-\eta)^2}$, we can have the desired form:

$$\frac{\lambda_{\max}(\mathbf{K}+\mathbf{S}+\sigma\mathbf{I})}{a\lambda_{\max}(\mathbf{K}+\sigma\mathbf{I})} \leq \left(\frac{\mathcal{R}(\boldsymbol{\theta} \cup \hat{\boldsymbol{\theta}})}{\mathcal{R}(\boldsymbol{\theta})}\right)^{\frac{1}{2}} \leq \frac{a\lambda_{\max}(\mathbf{K}+\mathbf{S}+\sigma\mathbf{I})}{\lambda_{\max}(\mathbf{K}+\sigma\mathbf{I})}$$

Section 6 : Experiments

Experiments

What they show:

- **Does the regularizer in fact make the model linear?**
- **Does the condition number on matrix K really bound the empirical risk?**

Experiment setup:

- **RoBERTa base fine-tuning task with LoRA**
- **Dataset: Binary classification (SST-2, CoLA, IMDb, Yelp)**

Some notes:

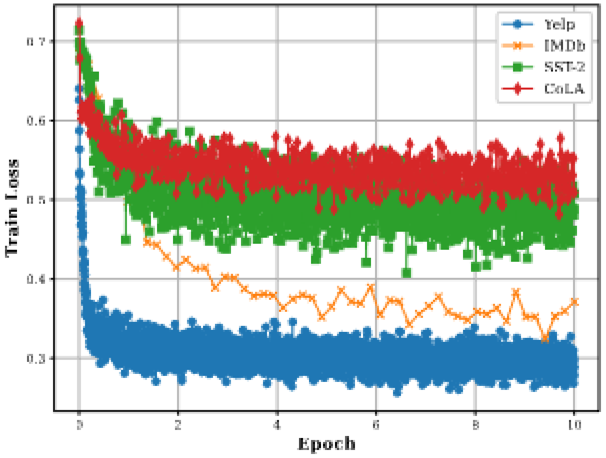
- **They transfer their findings into MSE to cross-entropy, and GD to AdamW.**
- **To fair comparison with other tasks, they converted Yelp dataset into binary classification**

Does the regularizer in fact make the model linear?

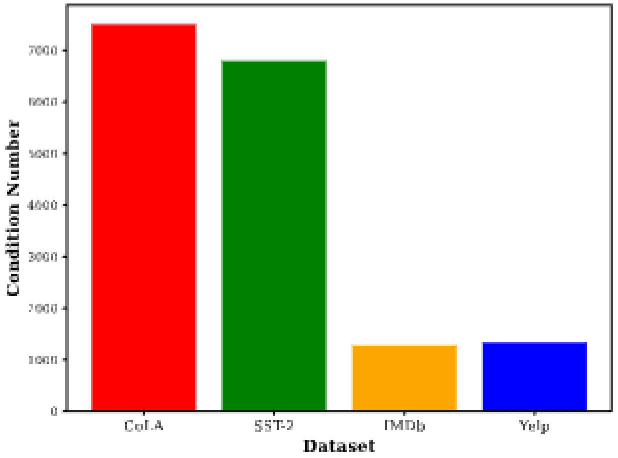
Dataset	Hyper-Parameter λ	50	10	5	2	1	0.5	0.1	0.0
CoLA	$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2$	0.280	0.350	0.404	0.5263	0.6148	0.6946	0.8223	0.960
	$\|f_{\boldsymbol{\theta}_t}(\mathbf{x}^*) - \bar{f}_{\bar{\boldsymbol{\theta}}_t}(\mathbf{x}^*)\|_2$	1.06	1.12	1.39	1.25	1.27	1.32	1.28	1.47
	KL Divergence	0.1060	0.1377	0.200	0.1613	0.1788	0.1961	0.1599	0.210
	Evaluation Accuracy of $f_{\boldsymbol{\theta}_t}(\mathbf{x})$	74.59	79.57	80.44	79.38	80.24	80.15	80.15	79.67
SST-2	$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\|_2$	0.292	0.336	0.369	0.424	0.520	0.700	1.589	2.519
	$\|f_{\boldsymbol{\theta}_t}(\mathbf{x}^*) - \bar{f}_{\bar{\boldsymbol{\theta}}_t}(\mathbf{x}^*)\|_2$	1.712	2.303	2.635	2.957	3.217	3.331	3.397	2.791
	KL Divergence	0.320	0.433	0.476	0.517	0.545	0.560	0.578	0.540
	Evaluation Accuracy of $f_{\boldsymbol{\theta}_t}(\mathbf{x})$	0.893	0.912	0.915	0.924	0.928	0.930	0.924	0.916

As the regularization strength increases, the model behaves more like a linear model.

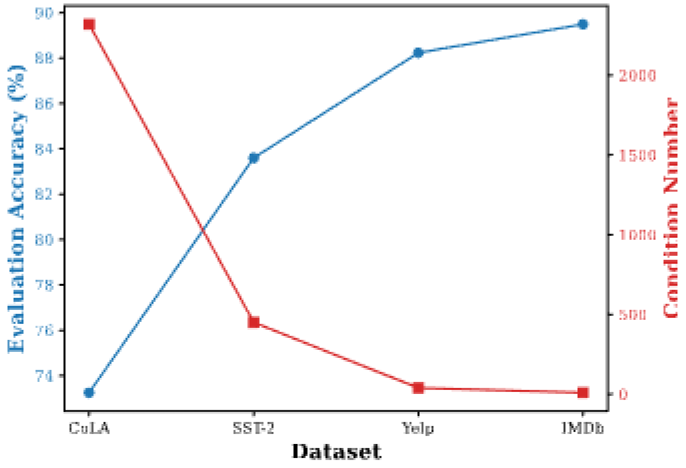
Does the condition number on matrix K really bound the empirical risk?



(a) Train loss over 10 epochs



(b) Condition number



(c) Evaluation accuracy and Condition number

Higher condition number → Higher train loss

Limitation

Limitation

1. **Discrepancy between theoretical findings and experimental setup**

2. **Weak validation of the main assumption: fine-tuned parameters remain close to the initialization.**

3. **Unclear and weak support for the layer selection algorithm they propose.**

End of Document

Lemma 3. Gradient Flow

- Let θ_t be the gradient flow limit of the regularized fine-tuning gradient descent described above. If we assume that λ switches at most countably often and denote instantaneous value by λ_t, the θ_t satisfies the following differential equation
$$\frac{d}{dt}\theta_t = -\nabla_\theta \tilde{R}(\theta_t)^\top - \lambda_t(\theta_t - \theta_0)$$

- Gradient descent on regularization term : $\theta_t = \tilde{\theta}_t - \lambda\eta(\tilde{\theta}_t - \theta_0) \rightarrow \theta_t - \theta_0 = \tilde{\theta}_t - \theta_0 - \lambda\eta(\tilde{\theta}_t - \theta_0)$
- Modified regularization step : $\theta_t = \begin{cases} \tilde{\theta}_t - \lambda\eta(\tilde{\theta}_t - \theta_0) & if\ \nabla_\theta \tilde{R}(\theta_t)(\theta_t - \theta_0) \geq 0 \\ \tilde{\theta}_t & if\ \nabla_\theta \tilde{R}(\theta_t)(\theta_t - \theta_0) < 0 \end{cases}$
  - $\theta_t = \theta_0 + (1 - \lambda_t\eta)(\tilde{\theta}_t - \theta_0)\ where\ \lambda_t = \begin{cases} \lambda & if\ \nabla_\theta \tilde{R}(\theta_t)(\theta_t - \theta_0) \geq 0 \\ 0 & if\ \nabla_\theta \tilde{R}(\theta_t)(\theta_t - \theta_0) < 0 \end{cases}$
- $\theta_{t+1} = \theta_t - \eta\nabla\tilde{R}(\theta_t)^\top - \lambda_{t+1}\eta(\tilde{\theta}_{t+1} - \theta_0)$
  $= \theta_t - \eta\nabla\tilde{R}(\theta_t)^\top - \frac{\lambda_{t+1}\eta}{1-\lambda_{t+1}\eta}(\theta_{t+1} - \theta_0)\ (\because \tilde{\theta}_{t+1} - \theta_0 = \frac{\lambda_{t+1}\eta}{1-\lambda_{t+1}\eta}(\theta_{t+1} - \theta_0))$
- $\theta_{t+1} = (1 - \lambda_{t+1}\eta)\left(\theta_t - \eta\nabla\tilde{R}(\theta_t)^\top + \frac{\lambda_{t+1}\eta}{1-\lambda_{t+1}\eta}\theta_0\right) = (1 - \lambda_{t+1}\eta)\theta_t - \eta(1 - \lambda_{t+1}\eta)\nabla\tilde{R}(\theta_t)^\top + \lambda_{t+1}\eta\theta_0$
- $\frac{\theta_{t+1} - \theta_t}{\eta} = -\lambda_{t+1}\theta_t - (1 - \lambda_{t+1}\eta)\nabla\tilde{R}(\theta_t)^\top + \lambda_{t+1}\eta\theta_0 \rightarrow \lim_{\eta\rightarrow 0}\frac{\theta_{t+1} - \theta_t}{\eta} = \frac{d}{dt}\theta_t = -\nabla_\theta\tilde{R}(\theta_t)^\top - \lambda_t(\theta_t - \theta_0)$

Lemma 1. Norm Bounds and Lipschitzness

- Consider the selectively regularized fine-tuning solution, under the squared loss. Denote the instantaneous value of the regularization parameter by λ_t, which can be either 0 or λ. If $f_\theta(x^*)$ is $Lip(f) - Lipschitz$ in an $l_2 - ball$ of radius τ around pretrained parameters θ_0,
- $\|\theta - \theta_0\| \leq 2\, Lip(f) \left\| f_{\theta_t}(x^*) - y^* \right\| \int_0^t e^{-(\Lambda_t - \Lambda_s)} ds$, where $\Lambda_t = \int_0^t \lambda_s ds$

- $\nabla_\theta \widetilde{R}(\theta_t) = 2(y(t) - y^*)^\top \nabla \mathrm{f}_{\theta_t}(x^*)$
- $\frac{d}{dt}\theta_t = -\nabla_\theta \widetilde{R}(\theta_t)^\top - \lambda_t(\theta_t - \theta_0) = -2\nabla \mathrm{f}_{\theta_t}(x^*)^\top (y(t) - y^*) - \lambda_t(\theta_t - \theta_0)$
- $\frac{d}{dt}u_t = -2\nabla \mathrm{f}_{\theta_t}(x^*)^\top (y(t) - y^*) - \lambda_t u_t$ where $u_t = (\theta_t - \theta_0)$

Theorem 4.

- Under the squared loss, if $f_{\theta_t}(x^*)$ and $\nabla f_{\theta_t}(x^*)$ are $Lip(f) - Lipschitz$ and $Lip(\nabla f) - Lipschitz$ in an $l_2 - ball$ of radius τ around pretrained parameters θ_0.
- *Define* $\lambda_\circ = \frac{2\left\|f_{\theta_0}(x^*) - y^*\right\| Lip(f)}{r}$
- *If $\lambda \geq \lambda_\circ$, then for all t, the following holds*
- *If $\lambda < \lambda_\circ$, then the following holds for* $t \leq \frac{1}{\lambda} ln(\frac{1}{1-\lambda/\lambda_\circ})$
- *In particular, for $\lambda \geq \lambda_\circ$, the bound from theorem 3 always holds and simplifies to*
- $\left\|f_{\theta_t}(x^*) - \bar{f}_{\bar{\theta}_t}(x^*)\right\| \leq 2Lip(f)\widetilde{R}(\theta_0)\big(2rLip(\nabla f) + Lip(f)\big)t$

- Theorem 2 : $\|\theta_t - \theta_0\| \leq \frac{2\|y(0) - y^*\| Lip(f)}{\lambda}\left(1 - e^{-\lambda t}\right)$
- In order for the regularizer to satisfy the Lipschitz continuity assumptions, mainly that $\|\theta_t - \theta_0\| \leq r$, this shows that there are two phases of behavior depending on how large λ is, where the threshold is given by $\lambda_\circ = \frac{2\left\|f_{\theta_0}(x^*) - y^*\right\| Lip(f)}{r}$
- In particular, if $\lambda \geq \lambda_\circ$, then θ_t remains in the r -ball around θ_0 for all t. Otherwise, it remains in this ball only as long as $t \leq \frac{1}{\lambda} ln(\frac{1}{1-\lambda/\lambda_\circ})$

Theorem 4.

- Under the squared loss, if $f_{\theta_t}(x^*)$ and $\nabla f_{\theta_t}(x^*)$ are $Lip(f) - Lipschitz$ and $Lip(\nabla f) - Lipschitz$ in an $l_2 - ball$ of radius τ around pretrained parameters θ_0.
- *Define* $\lambda_\circ = \frac{2\left\| f_{\theta_0}(x^*) - y^* \right\| Lip(f)}{r}$
- *If* $\lambda \geq \lambda_\circ$, *then for all* t, *the following holds*
- *If* $\lambda < \lambda_\circ$, *then the following holds for* $t \leq \frac{1}{\lambda} ln(\frac{1}{1-\lambda/\lambda_\circ})$
- *In particular, for* $\lambda \geq \lambda_\circ$, *the bound from theorem 3 always holds and simplifies to*
- $\left\| f_{\theta_t}(x^*) - \bar{f}_{\bar{\theta}_t}(x^*) \right\| \leq 2Lip(f)\widetilde{R}(\theta_0)\big(2rLip(\nabla f) + Lip(f)\big)t$

- Based on Theorem 3, we get that
- $\Delta(t) \leq 2Lip(f)^2 \left\| f_{\theta_0}(x^*) - y^* \right\| (\frac{4}{\lambda} Lip(\nabla f) \left\| f_{\theta_0}(x^*) - y^* \right\| + 1)(t + \frac{1}{\lambda} e^{-\lambda t} - \frac{1}{\lambda})$
- When $\lambda \geq \lambda_\circ$, $\Delta(t)$ grows linearly, with coefficient given by
- $\Delta(t) \leq 2Lip(f) \left\| y(0) - y^* \right\| \big(2rLip(\nabla f) + Lip(f)\big)t$