Linearization Explains Fine-Tuning in Large Language Models

Zahra Rahimi Afzal[1], Tara Esmaeilbeig[1,2], Mojtaba Soltanalian[1], Mesrob I. Ohannessian[1]

[1]University of Illinois Chicago, USA
[2]Nokia Bell Labs, USA

NEURAL INFORMATION PROCESSING SYSTEMS

NOKIA BELL LABS

Motivation

- Foundational LLMs are adapted to downstream NLP tasks via fine-tuning.

Motivation

- Foundational LLMs are adapted to downstream NLP tasks via fine-tuning.
- Full fine-tuning is expensive in terms of time and other computational resources due to the large number of parameters.

Motivation

- Foundational LLMs are adapted to downstream NLP tasks via fine-tuning.
- Full fine-tuning is expensive in terms of time and other computational resources due to the large number of parameters.
- **PEFT** reduces effective trained parameters (layer selection, rank-limited updates) while preserving adaptation.

Motivation

- Foundational LLMs are adapted to downstream NLP tasks via fine-tuning.
- Full fine-tuning is expensive in terms of time and other computational resources due to the large number of parameters.
- **PEFT** reduces effective trained parameters (layer selection, rank-limited updates) while preserving adaptation.
- However, these methods often *lack a fundamental understanding* of the dynamics behind these choices, hindering informed exploration.

Motivation

- Foundational LLMs are adapted to downstream NLP tasks via fine-tuning.
- Full fine-tuning is expensive in terms of time and other computational resources due to the large number of parameters.
- **PEFT** reduces effective trained parameters (layer selection, rank-limited updates) while preserving adaptation.
- However, these methods often *lack a fundamental understanding* of the dynamics behind these choices, hindering informed exploration.

This paper: introduce **linearized fine-tuning**, a way to understand how large models adapt by viewing fine-tuning through the Neural Tangent Kernel (NTK) lens. Linearizing the fine-tuning process closely aligns it with **NTK regression**. This perspective helps us predict **model performance** based on the properties of the NTK.

Motivation

- Foundational LLMs are adapted to downstream NLP tasks via fine-tuning.
- Full fine-tuning is expensive in terms of time and other computational resources due to the large number of parameters.
- **PEFT** reduces effective trained parameters (layer selection, rank-limited updates) while preserving adaptation.
- However, these methods often *lack a fundamental understanding* of the dynamics behind these choices, hindering informed exploration.

This paper: introduce **linearized fine-tuning**, a way to understand how large models adapt by viewing fine-tuning through the Neural Tangent Kernel (NTK) lens. Linearizing the fine-tuning process closely aligns it with **NTK regression**. This perspective helps us predict **model performance** based on the properties of the NTK.

Gap addressed: Prior results establish when linearity *may* hold, but *do not quantify* closeness to linearity. We add an explicit inductive bias and prove an **upper bound** on the distance between the fine-tuned model and its linearized approximation, supporting NTK-based performance predictions.

Regularized Fine-Tuning and Linearization

Given a pretrained model $f_{\boldsymbol{\theta}_0}(\cdot)$, a target task dataset $\mathcal{D}_T = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$ for the downstream task, and a loss function $\mathcal{L}(\cdot, \cdot)$, the objective is

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n} \mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i)}_{\tilde{\mathcal{R}}(\boldsymbol{\theta})} + \frac{\lambda}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2.$$

Regularized Fine-Tuning and Linearization

Given a pretrained model $f_{\boldsymbol{\theta}_0}(\cdot)$, a target task dataset $\mathcal{D}_T = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$ for the downstream task, and a loss function $\mathcal{L}(\cdot, \cdot)$, the objective is

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n} \mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i)}_{\tilde{\mathcal{R}}(\boldsymbol{\theta})} + \frac{\lambda}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2.$$

- The regularization term enforces proximity to the pretrained parameters $\boldsymbol{\theta}_0$, promoting a lazy or linear training regime.

Regularized Fine-Tuning and Linearization

Given a pretrained model $f_{\boldsymbol{\theta}_0}(\cdot)$, a target task dataset $\mathcal{D}_T = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$ for the downstream task, and a loss function $\mathcal{L}(\cdot,\cdot)$, the objective is

$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \underbrace{\sum_{i=1}^n \mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i)}_{\tilde{\mathcal{R}}(\boldsymbol{\theta})} + \frac{\lambda}{2}\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2.$$

- The regularization term enforces proximity to the pretrained parameters $\boldsymbol{\theta}_0$, promoting a lazy or linear training regime.
- The fine-tuning dynamics can be approximated by a first-order Taylor expansion of the model around the pretrained parameters $\boldsymbol{\theta}_0$

$$\bar{f}_{\bar{\boldsymbol{\theta}}_t}(\mathbf{x}) = f_{\boldsymbol{\theta}_0}(\mathbf{x}) + \left\langle \nabla f_{\boldsymbol{\theta}_0}(\mathbf{x}), \bar{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0 \right\rangle.$$

Regularized Fine-Tuning and Linearization

Given a pretrained model $f_{\boldsymbol{\theta}_0}(\cdot)$, a target task dataset $\mathcal{D}_T = (\mathbf{x}_i, \mathbf{y}_i)_{i=1}^n$ for the downstream task, and a loss function $\mathcal{L}(\cdot, \cdot)$, the objective is

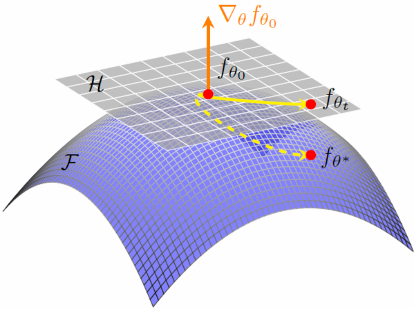
$$\boldsymbol{\theta}^\star = \underset{\boldsymbol{\theta}}{\operatorname{argmin}} \underbrace{\sum_{i=1}^{n} \mathcal{L}(f_{\boldsymbol{\theta}}(\mathbf{x}_i), \mathbf{y}_i)}_{\tilde{\mathcal{R}}(\boldsymbol{\theta})} + \frac{\lambda}{2} \|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_2^2.$$

- The regularization term enforces proximity to the pretrained parameters $\boldsymbol{\theta}_0$, promoting a lazy or linear training regime.
- The fine-tuning dynamics can be approximated by a first-order Taylor expansion of the model around the pretrained parameters $\boldsymbol{\theta}_0$

$$\bar{f}_{\bar{\boldsymbol{\theta}}_t}(\mathbf{x}) = f_{\boldsymbol{\theta}_0}(\mathbf{x}) + \left\langle \nabla f_{\boldsymbol{\theta}_0}(\mathbf{x}), \bar{\boldsymbol{\theta}}_t - \boldsymbol{\theta}_0 \right\rangle .$$

- The linearized model $\bar{f}_{\bar{\boldsymbol{\theta}}_t}(\mathbf{x})$ evolves according to Neural Tangent Kernel (NTK) dynamics.
- This makes fine-tuning theoretically equivalent to NTK regression while preserving practical accuracy.

Linearization



The NTK defines a linear function space $\mathcal{H}$ tangent to the non-linear function space $\mathcal{F}$ defined by the model. Regularized fine-tuning in the lazy regime is close to kernel regression on the tangent space. $f_{\theta^\star}(\mathbf{x})$ is the fine-tuned model obtained by empirical risk minimization. If fine-tuning remains in the linearized regime, then after T steps of training $f_{\theta^\star}(\mathbf{x}) \approx f_{\boldsymbol{\theta}_0}(\mathbf{x}) + \langle \nabla_{\boldsymbol{\theta}} f_{\boldsymbol{\theta}_0}(\mathbf{x}), \boldsymbol{\theta}_T - \boldsymbol{\theta}_0 \rangle$ is a good approximation.

Theoretical Results

We show that if $f_{\boldsymbol{\theta}}(\mathbf{x})$ is Lipschitz continuous in an ℓ_2-ball of radius r around the pretrained parameters $\boldsymbol{\theta}_0$, then we have

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\| \leq 2\,\mathrm{Lip}(f)\,\|f_{\boldsymbol{\theta}_0}(\mathbf{x}) - \mathbf{y}\|\,\frac{1 - e^{-\lambda t}}{\lambda}.$$

- The parameter deviation from initialization is bounded by the model's smoothness Lip(f), the initial prediction error, and the regularization strength.

Theoretical Results

We show that if $f_{\boldsymbol{\theta}}(\mathbf{x})$ is Lipschitz continuous in an ℓ_2-ball of radius r around the pretrained parameters $\boldsymbol{\theta}_0$, then we have

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\| \leq 2\,\mathrm{Lip}(f)\,\|f_{\boldsymbol{\theta}_0}(\mathbf{x}) - \mathbf{y}\|\,\frac{1 - e^{-\lambda t}}{\lambda}.$$

- The parameter deviation from initialization is bounded by the model's smoothness Lip(f), the initial prediction error, and the regularization strength.
- Larger $\lambda \to$ smaller deviation $\to$ training remains close to $\boldsymbol{\theta}_0$.

Theoretical Results

We show that if $f_{\boldsymbol{\theta}}(\mathbf{x})$ is Lipschitz continuous in an ℓ_2-ball of radius r around the pretrained parameters $\boldsymbol{\theta}_0$, then we have

$$\|\boldsymbol{\theta}_t - \boldsymbol{\theta}_0\| \leq 2\,\mathrm{Lip}(f)\,\|f_{\boldsymbol{\theta}_0}(\mathbf{x}) - \mathbf{y}\|\,\frac{1 - e^{-\lambda t}}{\lambda}.$$

- The parameter deviation from initialization is bounded by the model's smoothness Lip(f), the initial prediction error, and the regularization strength.
- Larger $\lambda \to$ smaller deviation $\to$ training remains close to $\boldsymbol{\theta}_0$.

This result serves as a building block for proving the distance between the fine-tuned model and its linearized version.

$$\|f_{\boldsymbol{\theta}_t}(\mathbf{x}) - \bar{f}_{\bar{\boldsymbol{\theta}}_t}(\mathbf{x})\| \leq 2\,\mathsf{Lip}(f)\widetilde{R}(\boldsymbol{\theta}_0)\,(2r\mathsf{Lip}(\nabla f) + \mathsf{Lip}(f))\,t.$$

Empirical Risk Bounds under the NTK Regime

We formulate the fine-tuning problem as a regularized function estimation in the RKHS, $\mathcal{H}$, generated by the NTK, $\mathbf{k}(\mathbf{x}, \mathbf{x}') = \nabla f_{\boldsymbol{\theta}_0}(\mathbf{x}) \nabla f_{\boldsymbol{\theta}_0}(\mathbf{x}')^\top$.

In the linearized regime, minimizing the empirical risk is equivalent to **kernel regression in the NTK RKHS**:

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

Empirical risk depends on NTK spectrum

$$\left(\frac{\sigma \|\mathbf{y}\|}{\sigma + \lambda_{\text{max}}(\mathbf{K})}\right)^2 \leq \mathcal{R}(\boldsymbol{\theta}) \leq \left(\frac{\sigma \|\mathbf{y}\|}{\sigma + \lambda_{\text{min}}(\mathbf{K})}\right)^2.$$

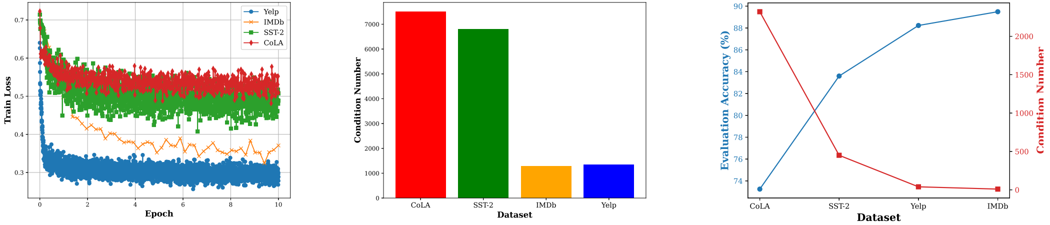
$\Rightarrow$ **Predictor:** well-conditioned NTK (smaller condition number) $\Rightarrow$ lower risk / better generalization.

Experiments

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

Table: Sweep over the hyperparameter (λ). Increasing regularization strength, i.e., larger λ, reduces the deviation between the regularized fine-tuning and linearized models at one snapshot of fine-tuning at step t. Accuracy is largely unaffected by regularization.

Experiments



(a) Train loss over 10 epochs

(b) Condition number

(c) Evaluation accuracy and Condition number

Figure: (a)-(b) Illustrate the positive correlation between the convergence rate of optimization steps of LoRA over 10 epochs and condition number of NTK at initialization. $\{\mathbf{W}_q, \mathbf{W}_v\}$ of layers $\{0, 5, 11\}$ are fine-tuned. (c) Illustrates the negative correlation between evaluation accuracy after 10 epochs of training and the condition number of NTK. LoRA with $r = 8$ is used to fine-tune $\{\mathbf{W}_k\}$ of the layers $\{0, 5, 11\}$.

Takeaways

- **Regularized fine-tuning $\Rightarrow$ linearized (NTK) regime.**
- **The NTK spectrum at initialization predicts downstream performance.**
- **Simple spectral criteria guide PEFT layer selection before training.**

Broader impact: a theory-grounded lens + practical diagnostics for efficient LLM adaptation.

NEURAL INFORMATION PROCESSING SYSTEMS

Thank you!

Email: Zrahim2@uic.edu