LoRA Training in the NTK Regime has No Spurious Local Minima

Uijeong Jang [1] **Jason D. Lee** [2] **Ernest K. Ryu** [3]

Abstract

Low-rank adaptation (LoRA) has become the standard approach for parameter-efficient fine-tuning of large language models (LLM), but our theoretical understanding of LoRA has been limited. In this work, we theoretically analyze LoRA fine-tuning in the neural tangent kernel (NTK) regime with N data points, showing: (i) full fine-tuning (without LoRA) admits a low-rank solution of rank $r \lesssim \sqrt{N}$; (ii) using LoRA with rank $r \gtrsim \sqrt{N}$ eliminates spurious local minima, allowing (stochastic) gradient descent to find the low-rank solutions; (iii) the low-rank solution found using LoRA generalizes well.

1. Introduction

The modern methodology of using large language models involves (at least) two phases: self-supervised pre-training on a large corpus followed by supervised fine-tuning to the downstream task. As large language models have grown in scale, pre-training has become out of reach for research groups without access to enormous computational resources. However, supervised fine-tuning remains feasible for such groups. One key strategy facilitating this efficient fine-tuning is Parameter-Efficient Fine-Tuning (PEFT), which freezes most of the pre-trained model's weights while selectively fine-tuning a smaller number of parameters within an adapter module. Among various PEFT methodologies, low-rank adaptation (LoRA) (Hu et al., 2021) has emerged as the standard approach. Given a pre-trained matrix $W_0 \in \mathbb{R}^{m\times n}$, LoRA trains a low-rank update such that the forward pass evaluates

$$W_0 x + \Delta W x = W_0 x + BAx$$

where $r \ll \min(m,n)$, $A \in \mathbb{R}^{r\times n}$ is initialized to be a random Gaussian, and $B \in \mathbb{R}^{m\times r}$ is initialized to be zero.

However, despite the widespread adoption of LoRA, our theoretical understanding of its mechanisms remains limited. One notable prior work is (Zeng & Lee, 2024), which analyzes the expressive power of LoRA, showing that for any given function, there exist weight configurations for LoRA that approximate it. However, their work does not address whether LoRA can efficiently learn such configurations. Additionally, Malladi et al. (2023) experimentally demonstrated that under certain conditions, LoRA fine-tuning is nearly equivalent to a kernel regression, where the A matrix provides random features and is essentially not trained. This regime neglects the possibility of the A matrix learning new features and, consequently, leads to a LoRA rank requirement of $r \geq \Theta(1/\varepsilon^2)$, where ε is an approximation tolerance, originating from the use of the Johnson–Lindenstrauss lemma (Johnson & Lindenstrauss, 1984). Crucially, LoRA's fundamental nature as a quadratic parameterization has not been considered in the prior analysis of trainability and generalizability.

Contribution. In this work, we theoretically analyze LoRA fine-tuning and present results on trainability and generalizability. We consider fine-tuning a deep (transformer) neural network with K-dimensional outputs using N training (fine-tuning) data points. Assuming that training remains under the NTK regime, which we soon define and justify in Section 2, we show the following. First, full fine-tuning (without LoRA) admits a rank-r solution such that $\frac{r(r+1)}{2} \leq KN$. Second, using LoRA with rank r such that $\frac{r(r+1)}{2} > KN$ eliminates spurious local minima, allowing (stochastic) gradient descent to find the low-rank solutions. Finally, the low-rank solution found using LoRA generalizes well.

1.1. Prior works

Theory of neural networks. The question of expressive power addresses whether certain neural networks of interest can approximate a given target function. Starting with the classical universal approximation theorems (Cybenko, 1989; Hornik et al., 1990; Barron, 1993), much research has been conducted in this direction. (Delalleau & Bengio, 2011;

[1]Department of Mathematical Sciences, Seoul National University [2]Department of Electrical and Computer Engineering, Princeton University [3]Department of Mathematics, University of California, Los Angeles. Correspondence to: Ernest Ryu <eryu@math.ucla.edu>.

Proceedings of the 41st International Conference on Machine Learning, Vienna, Austria. PMLR 235, 2024. Copyright 2024 by the author(s).

Bengio & Delalleau, 2011; Lu et al., 2017; Duan et al., 2023). These can be thought of as existence results.

The question of trainability addresses whether one can compute configurations of neural networks that approximate target functions. Ghadimi & Lan (2013); Ge et al. (2015); Du et al. (2017); Jin et al. (2017) studied general convergence results of gradient descent and stochastic gradient descent. Soltanolkotabi et al. (2018); Du & Lee (2018); Allen-Zhu et al. (2019a;b); Du et al. (2019); Zou et al. (2020) studied the loss landscape of neural networks and showed that first-order methods converge to global minima under certain conditions.

The question of generalization addresses whether neural networks trained on finite data can perform well on new unseen data. Classical learning theory (Koltchinskii & Panchenko, 2000; Bartlett et al., 2002; Bousquet & Elisseeff, 2002; Hardt et al., 2016; Bartlett et al., 2017) uses concepts such as uniform stability or the Rademacher complexities to obtain generalization bounds. Generalization bounds in the context of modern deep learning often utilize different approaches (Wu et al., 2017; Dinh et al., 2017; Zhang et al., 2021), we use the Rademacher complexity for obtaining our generalization results.

Neural tangent kernels. The theory of neural tangent kernel (NTK) concerns the training dynamics of certain infinitely wide neural networks. Jacot et al. (2018) shows that the training of an infinitely wide neural network is equivalent to training a kernel machine. Various studies such as (Arora et al., 2019; Chen et al., 2020) expand the NTK theory to more practical settings. Among these works, Wei et al. (2022a) introduced the concept of empirical NTK (eNTK) and showed that kernel regression with pretrained initialization also performs well on real datasets, providing a background to utilize NTK theory in fine-tuning.

Theory of transformers and LLMs. As the transformer architecture (Vaswani et al., 2017) became the state-of-the-art architecture for natural language processing and other modalities, theoretical investigations of transformers have been pursued. Results include that transformers are universal approximators (Yun et al., 2019), that transformers can emulate a certain class of algorithmic instructions (Wei et al., 2022b; Giannou et al., 2023), and that weight matrices in transformers increase their rank during training (Boix-Adsera et al., 2023). Also, (Zhang et al., 2020; Liu et al., 2020) presents improved adaptive optimization methods for transformers.

PEFT methods and LoRA. Low-rank adaptation (LoRA) (Hu et al., 2021) has become the standard Parameter-Efficient Fine-Tuning (PEFT) method, and many variants of LoRA have been presented (Fu et al., 2023; Dettmers et al., 2023; Lialin et al., 2023). LoRA has proven to be quite versatile and has been used for convolution layers (Yeh et al., 2024) and for diffusion models (Ryu, 2023; Smith et al., 2023; Choi et al., 2023).

Theoretically, Aghajanyan et al. (2021) found an intrinsic low-rank structure is critical for fine-tuning language models, although this finding concerns full fine-tuning, not the setting that uses LoRA. Recently, Zeng & Lee (2024) analyzed the expressive power of LoRA. However, we still lack a sufficient theoretical understanding of why LoRA is effective in the sense of optimization and generalization.

Matrix factorization. In this work, we utilize techniques developed in prior work on matrix factorization problems. Bach et al. (2008); Haeffele et al. (2014) established the sufficiency of low-rank parameterizations in matrix factorization problems, and their techniques have also been used in matrix completion (Ge et al., 2016), matrix sensing (Jin et al., 2023), and semidefinite programming (Bhojanapalli et al., 2018).

1.2. Organization

Section 2 introduces the problem setting and reviews relevant prior notions and results. Section 3 proves the existence of low-rank solutions. Section 4 proves LoRA has no spurious local minima and, therefore, establishes that (stochastic) gradient descent can find the low-rank global minima. Section 5 shows that the low-rank solution generalizes well. Finally, Section 6 presents simple experiments fine-tuning pre-trained models for different modalities. The experimental results validate our theory and provide further experimental insights.

2. Problem setting and preliminaries

We primarily consider the setup of pre-trained large language models fine-tuned with LoRA. However, our theory does generally apply to other setups that utilize pre-training and LoRA fine-tuning, such as diffusion models.

Matrix notation. For matrices A and B, let $\|A\|_*$ denote the nuclear norm, $\|A\|_F$ the Frobenius norm, and $\langle A, B\rangle = \mathbf{tr}(A^\intercal B)$ the matrix inner product. We let $\mathbb{S}^n$ and $\mathbb{S}^n_+$ for the set of $n \times n$ symmetric and positive semi-definite matrices, respectively. Let $\mathcal{R}(\cdot)$ and $\mathcal{N}(\cdot)$ respectively denote the range and the null-space of a linear operator.

Neural network. Let $f_\Theta \colon \mathcal{X} \to \mathbb{R}^K$ be a neural network (e.g., a transformer-based model) parametrized by Θ, where $\mathcal{X}$ is the set of data (e.g., natural language text) and $\mathbb{R}^K$ is the output (e.g., pre-softmax logits of tokens). K is the output dimension of f_Θ, where $K = k$ for k-class classification, $K = 1$ for binary classification, and K is the

dimension of the label Y when using mean square error loss. Assume the model has been pre-trained to $\Theta = \Theta_0$, i.e., the pre-trained model is f_{Θ_0}.

Let $\mathbf{W} = (W^{(1)}, \ldots, W^{(T)}) \subset \Theta$ be a subset of the weights (e.g., dense layers in QKV-attention) with size $W^{(i)} \in \mathbb{R}^{m_i \times n_i}$ for $i = 1, \ldots, T$ that we choose to fine-tune. Let $\mathbf{W}_0 = (W_0^{(1)}, \ldots, W_0^{(T)}) \subset \Theta_0$ be their corresponding pre-trained weights. With slight abuse of notation, write $f_{\mathbf{W}}$ to denote f_Θ, where all parameters of Θ excluding $\mathbf{W}$ are fixed to their corresponding values in Θ_0.

Fine-tuning loss. Assume we wish to fine-tune the pre-trained model with

$$\{(X_i, Y_i)\}_{i=1}^N,$$

where N is the number of (fine-tuning) training data. (In many NLP tasks, it is not uncommon to have $N < 100$.) Denote $\boldsymbol{\delta} = (\delta^{(1)}, \ldots, \delta^{(T)}) \subset \Theta$ to be the change of $\mathbf{W}$ after the fine-tuning, i.e., $f_{\mathbf{W}_0+\boldsymbol{\delta}}$ is our fine-tuned model. We use the empirical risk

$$\hat{\mathcal{L}}(\boldsymbol{\delta}) = \frac{1}{N} \sum_{i=1}^N \ell(f_{\mathbf{W}_0+\boldsymbol{\delta}}(X_i), Y_i),$$

with some loss function ℓ. We assume $\ell(x, y)$ is convex, non-negative, and twice-differentiable with respect to x for any y. (This assumption holds for the cross-entropy loss and the mean squared error loss.) The empirical risk approximates the true risk

$$\mathcal{L}(\boldsymbol{\delta}) = \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{P}} \left[\ell(f_{\mathbf{W}_0+\boldsymbol{\delta}}(X), Y)\right]$$

with some data distribution $\mathcal{P}$.

NTK regime. Under the NTK regime (also referred to as the lazy-training regime), the change of the network can be approximated by its first-order Taylor expansion

$$f_{\mathbf{W_0}+\boldsymbol{\delta}}(X) \approx f_{\mathbf{W_0}}(X) + \langle \nabla f_{\mathbf{W_0}}(X), \boldsymbol{\delta} \rangle \qquad (1)$$

sufficiently well throughout (fine-tuning) training. To clarify, $f_{\mathbf{W_0}+\boldsymbol{\delta}}(X) \in \mathbb{R}^K$, so the NTK regime requires the first-order Taylor expansion to be accurate for all coordinates:

$$f^{(j)}_{\mathbf{W_0}+\boldsymbol{\delta}}(X) \approx f^{(j)}_{\mathbf{W_0}}(X) + \langle \nabla f^{(j)}_{\mathbf{W_0}}(X), \boldsymbol{\delta} \rangle,$$

where $f^{(j)}_{\mathbf{W}}$ is the j-th coordinate of $f_{\mathbf{W}}$ for $j = 1, \ldots, K$.

The NTK regime is a reasonable assumption in fine-tuning if $\boldsymbol{\delta}$ is small, and this assertion is supported by the empirical evidence of (Malladi et al., 2023). This prior work provides extensive experiments on various NLP tasks to validate that fine-tuning happens within the NTK regime for many, although not all, NLP tasks.

Observation 2.1 (Malladi et al. (2023))**.** *When prompt-based fine-tuning (Schick & Schütze, 2021; Gao et al., 2021) is used, fine-tuning a pre-trained language model stays within the NTK regime.*

Motivated by this empirical observation, we define linearized losses

$$\hat{L}(\boldsymbol{\delta}) = \frac{1}{N} \sum_{i=1}^N \ell\left(f_{\mathbf{W}_0}(X_i) + \langle \nabla f_{\mathbf{W_0}}(X_i), \boldsymbol{\delta} \rangle, Y_i\right) \approx \hat{\mathcal{L}}(\boldsymbol{\delta})$$

and

$$L(\boldsymbol{\delta}) = \mathop{\mathbb{E}}_{(X,Y)\sim\mathcal{P}} \left[\ell\left(f_{\mathbf{W}_0}(X_i) + \langle \nabla f_{\mathbf{W_0}}(X_i), \boldsymbol{\delta} \rangle, Y_i\right)\right] \approx \mathcal{L}(\boldsymbol{\delta}).$$

LoRA. We use the low-rank parameterization

$$\delta^{(i)} = u^{(i)}(v^{(i)})^\intercal \in \mathbb{R}^{m_i \times n_i},$$

where $u^{(i)} \in \mathbb{R}^{m_i \times r}, v^{(i)} \in \mathbb{R}^{n_i \times r}$, for $i \in \{1, \cdots, T\}$. Under the NTK regime, the empirical risk can be approximated as

$$\hat{L}(\mathbf{u}\mathbf{v}^\intercal) = \frac{1}{N} \sum_{i=1}^N \ell\left(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \mathbf{u}\mathbf{v}^\intercal \rangle, Y_i\right),$$

where

$$\mathbf{u} = \begin{bmatrix} u^{(1)} \\ \vdots \\ u^{(T)} \end{bmatrix} \in \mathbb{R}^{m \times r}, \qquad \mathbf{v} = \begin{bmatrix} v^{(1)} \\ \vdots \\ v^{(T)} \end{bmatrix} \in \mathbb{R}^{n \times r}$$

with $m = \sum_{i=1}^T m_i$ and $n = \sum_{i=1}^T n_i$, and

$$\mathbf{G}(X_i) = \operatorname{diag}\left(\nabla_{W^{(1)}} f_{\mathbf{W}_0}(X_i), \ldots, \nabla_{W^{(T)}} f_{\mathbf{W}_0}(X_i)\right)$$

is an collection of K $m \times n$ block diagonal matrices. To clarify, $\mathbf{G}(X_i) \in \mathbb{R}^{K \times m \times n}$, so $\langle \mathbf{G}(X_i), \mathbf{u}\mathbf{v}^\intercal \rangle \in \mathbb{R}^K$ should be interpreted as K inner products of $m \times n$ matrices where each matrices correspond to each coordinates of f. More specifically, $\mathbf{G}^{(j)}(X_i) \in \mathbb{R}^{m \times n}$ and

$$\left(\langle \mathbf{G}(X_i), \mathbf{u}\mathbf{v}^\intercal \rangle\right)_j = \langle \mathbf{G}^{(j)}(X_i), \mathbf{u}\mathbf{v}^\intercal \rangle$$

for $j = 1, \ldots, K$. Note that $\hat{L}(\mathbf{u}\mathbf{v}^\intercal)$ under the NTK regime is non-convex in $(\mathbf{u}, \mathbf{v})$ so SGD-training does not converge to the global minimizer, in general.

Weight decay on LoRA is nuclear norm regularization. The LoRA training of optimizing $\hat{L}$ is often conducted with weight decay (Hu et al., 2021; Dettmers et al., 2023), which can be interpreted as solving

$$\underset{\mathbf{u},\,\mathbf{v}}{\text{minimize}} \quad \hat{L}(\mathbf{u}\mathbf{v}^\intercal) + \tfrac{\lambda}{2}\|\mathbf{u}\|_F^2 + \tfrac{\lambda}{2}\|\mathbf{v}\|_F^2,$$

with regularization parameter $\lambda \geq 0$. This problem is equivalent to the rank-constrained nuclear-norm regularized problem

$$\underset{\boldsymbol{\delta},\, \mathrm{rank}\boldsymbol{\delta} \leq r}{\text{minimize}} \quad \hat{L}_\lambda(\boldsymbol{\delta}) \triangleq \hat{L}(\boldsymbol{\delta}) + \lambda \|\boldsymbol{\delta}\|_*.$$

This is due to the following lemma.

Lemma 2.2 (Lemma 5.1 of (Recht et al., 2010))**.** *Let $r > 0$. For $\boldsymbol{\delta} \in \mathbb{R}^{m\times n}$ such that $\mathrm{rank}(\boldsymbol{\delta}) \leq r$,*

$$\|\boldsymbol{\delta}\|_* = \frac{1}{2} \min_{\mathbf{u}\mathbf{v}^\intercal = \boldsymbol{\delta}} \{\|\mathbf{u}\|_F^2 + \|\mathbf{v}\|_F^2 \mid \mathbf{u} \in \mathbb{R}^{m\times r},\, \mathbf{v} \in \mathbb{R}^{n\times r}\}.$$

(The connection between weight decay on Burer–Monteiro style low-rank factorization and nuclear norm regularization has been previously in different contexts not directly related to LoRA (Cabral et al., 2013; Pilanci & Ergen, 2020).)

Second-order stationary points. Let $\hat{L}\colon \mathbb{R}^{m\times n} \to \mathbb{R}$ be twice-continuously differentiable. We say $U \in \mathbb{R}^{m\times n}$ is a (first-order) *stationary* point if

$$\nabla \hat{L}(U) = \mathbf{0}.$$

We say $U \in \mathbb{R}^{m\times n}$ is a *second-order stationary point* (SOSP) if

$$\nabla \hat{L}(U) = \mathbf{0}, \qquad \nabla^2 \hat{L}(U)[V, V] \geq 0,$$

for any direction $V \in \mathbb{R}^{m\times n}$. We say U is *strict saddle* if U is a first- but not second-order stationary point. Lastly, we say $U \in \mathbb{R}^{m\times n}$ is a *local minimum* if there exists an open ball B that contains U and

$$\hat{L}(U) \leq \hat{L}(U')$$

for any $U' \in B$. It follows that a local minimum is an SOSP.

The following results, roughly speaking, establish that (stochastic) gradient descent only converges to SOSPs when a loss function is twice-continuously differentiable.

Theorem 2.3 (Theorem 4.1 of (Lee et al., 2016))**.** *Gradient descent on twice-differentiable function with random initialization, almost surely, does not converge to strict saddle points. I.e., if gradient descent converges, it converges to an SOSP, almost surely.*

Theorem 2.4 (Informal, Theorem 1 of (Ge et al., 2015))**.** *Stochastic gradient descent with noise on twice-differentiable strict saddle function (i.e., every stationary point is either a local minimum or a strict saddle) does not converge to strict saddle points with high probability. I.e., if stochastic gradient descent with noise converges, it converges to an SOSP with high probability.*

Therefore, if we can show that all SOSPs are global minima in our setup of interest, then (stochastic) gradient descent will only converge to global minima.

3. Low-rank solution exists

In this section, we show that full fine-tuning in the NTK regime admits a low-rank solution of rank $r \lesssim \sqrt{N}$. The existence of a low-rank solution provides theoretical legitimacy to using the low-rank parameterization of LoRA, which, of course, can only find low-rank solutions.

Theorem 3.1. *Let $\lambda \geq 0$. Assume $\hat{L}_\lambda(\boldsymbol{\delta})$ has a global minimizer (not necessarily unique). Then there is a rank-r solution such that $\frac{r(r+1)}{2} \leq KN$.*

The assumption that $\hat{L}_\lambda(\boldsymbol{\delta})$ has a global minimum is very mild; it is automatically satisfied if $\lambda > 0$. When $\lambda = 0$, the assumption holds if ℓ is the mean squared error loss.

The inspiration for Theorem 3.1 comes from the classical results of (Barvinok, 1995; Pataki, 1998; 2000) that establish that semi-definite programs (which have symmetric positive semi-definite matrices as optimization variables) admit low-rank solutions. We clarify that Theorem 3.1 does not require $\boldsymbol{\delta}$ to be symmetric nor any notion of "semi-definiteness" ($\boldsymbol{\delta}$ is not even square).

Proof sketch of Theorem 3.1. We quickly outline the key ideas of the proof while deferring the details to Appendix A.

We can show that finding $\boldsymbol{\delta}_\lambda^\star \in \mathrm{argmin}_{\boldsymbol{\delta}} \hat{L}_\lambda(\boldsymbol{\delta})$ with $\mathrm{rank}(\boldsymbol{\delta}_\lambda^\star) = r$ is equivalent to finding a rank-r global minimum of $F\colon \mathbb{S}_+^{(m+n)} \to \mathbb{R}$ where

$$F(Z) = \hat{L}(\bar{Z}) + \frac{\lambda}{2}\mathbf{tr}(Z)$$

and $\bar{Z} = Z[1:m, m+1:m+n] \in \mathbb{R}^{m\times n}$. I.e., $\bar{Z}$ is a off-diagonal submatrix of Z such that

$$Z = \begin{bmatrix} * & \bar{Z} \\ \bar{Z}^\intercal & * \end{bmatrix}. \tag{2}$$

Now suppose $Z^\star \in \mathbb{S}_+^{(m+n)}$ is a global minimizer of F. Define $\mathcal{S}(Z^\star) \triangleq \{Z \in \mathbb{S}^{(m+n)} \colon \mathcal{R}(Z) \subseteq \mathcal{R}(Z^\star)\}$ and a linear operator $\mathcal{A}\colon \mathbb{S}^{(m+n)} \to \mathbb{R}^{KN}$ as

$$\mathcal{A}(Z)_{ij} = \langle \mathbf{G}^{(j)}(X_i), \bar{Z}\rangle, \qquad 1 \leq i \leq N, \quad 1 \leq j \leq K.$$

Now let $\mathrm{rank}(Z^\star) = r$ and assume

$$\{\mathbf{0}\} = \mathcal{S}(Z^\star) \cap \mathcal{N}(\mathcal{A}).$$

Then by dimension counting, we have the following inequal-

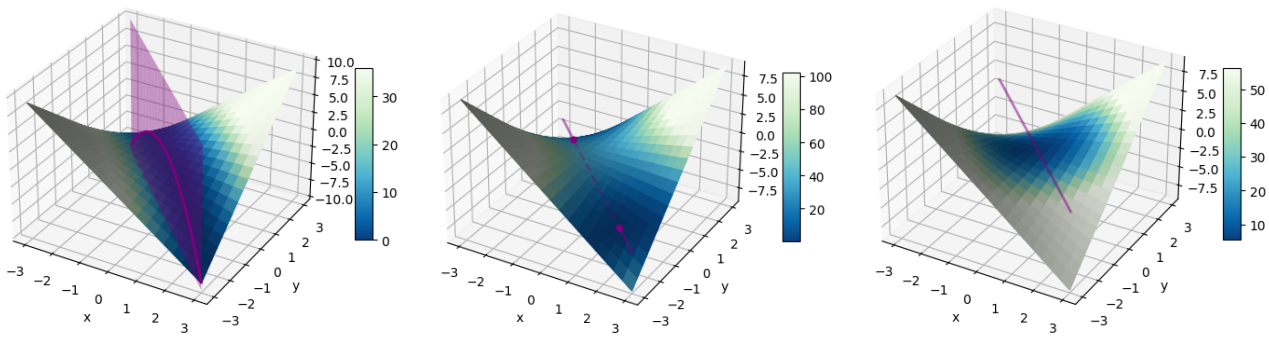


Figure 1. Geometric intuition of Theorem 3.1. The three dimensional space describes the space of 2 by 2 matrices $\begin{bmatrix} 1 & x \\ y & z \end{bmatrix}$. The surface $z = xy$ represents the rank 1 matrices. The blue region on the surface correspond to the region of smaller objective values, and the set of global minima are depicted with purple. **(Left)** Plot of (a) with $N = 1$. The set of global minima is a plane, and the intersection with the surface $z = xy$ (curve) is the set of rank-1 global minima. **(Middle)** Plot of (b) with $N = 2$. the set of global minima is a line, and the intersection with the surface (two dots) is the set of rank 1 global minima. **(Right)** Plot of (c) with $N = 3$. The set of global minima is a line, and there is no intersection with the surface, i.e., there is no global minimum of rank-1 but admits a rank-2 global minima.

ity.

$$
\begin{aligned}
0 &= \dim\mathcal{S}(Z^\star) + \dim\mathcal{N}(\mathcal{A}) - \dim(\mathcal{S}(Z^\star) + \mathcal{N}(\mathcal{A})) \\
&= \dim\mathcal{S}(Z^\star) + \dim(\mathbb{S}^{(m+n)}) - \dim\mathcal{R}(\mathcal{A}) \\
&\quad - \dim(\mathcal{S}(Z^\star) + \mathcal{N}(\mathcal{A})) \\
&= \dim\mathcal{S}(Z^\star) - KN + \dim(\mathbb{S}^{(m+n)}) \\
&\qquad - \dim(\mathcal{S}(Z^\star) + \mathcal{N}(\mathcal{A})) \\
&= \dim\mathcal{S}(Z^\star) - KN + \dim(\mathcal{S}(Z^\star)^\perp \cap \mathcal{R}(\mathcal{A})) \\
&\geq \dim\mathcal{S}(Z^\star) - KN
\end{aligned}
$$

If there exists nonzero $Z \in \mathbb{S}^{(m+n)}$ such that $Z \in \mathcal{S}(Z^\star) \cap \mathcal{N}(\mathcal{A})$, then we can show that there exists nonzero $t \in \mathbb{R}$ such that $Z^\star + tZ$ is also a global minimizer of F with strictly lower rank. Replace $Z^\star$ with $Z^\star + tZ$ and repeat this process until we find a solution $Z^\star$ with

$$\{\mathbf{0}\} = \mathcal{S}(Z^\star) \cap \mathcal{N}(\mathcal{A}).$$

Together with the fact that $\dim\mathcal{S}(Z^\star) = \frac{r(r+1)}{2}$, we have the desired result. □

Illustration of Theorem 3.1. The following toy example illustrates the geometric intuition of Theorem 3.1. Let ℓ be the mean square error loss, $K = 1$, $\boldsymbol{\delta} = \begin{bmatrix} w & x \\ y & z \end{bmatrix}$, and $\lambda = 0$ (no regularization). Then consider the following objective functions each for $N = 1$, 2, and 3:

$$\hat{L}_0(\boldsymbol{\delta}) = (x+y)^2 \tag{a}$$

$$\hat{L}_0(\boldsymbol{\delta}) = \frac{1}{2}(z+4)^2 + \frac{1}{2}(x+y)^2 \tag{b}$$

$$\hat{L}_0(\boldsymbol{\delta}) = \frac{1}{3}(w-1)^2 + \frac{1}{3}(z-4)^2 + \frac{1}{3}(\sqrt{3}x + \sqrt{3}y)^2 \tag{c}$$

The set of low-rank (rank-1) solutions for the three objectives are depicted in Figure 1.

4. GD and LoRA finds low-rank solution

In this section, we show that the optimization landscape with LoRA in the NTK regime has no spurious local minima if the LoRA parameterization uses rank $r \gtrsim \sqrt{N}$ and if we consider an ε-perturbed loss. This implies that optimizers such as stochastic gradient descent only converge to the low-rank global minimizers.

Theorem 4.1. *Let $\lambda \geq 0$. Assume $\hat{L}_\lambda(\boldsymbol{\delta})$ has a global minimizer (not necessarily unique) and $\frac{r(r+1)}{2} > KN$. Consider the perturbed loss function $\hat{L}_{\lambda,P}$ defined as*

$$\hat{L}_{\lambda,P}(\mathbf{u},\mathbf{v}) \triangleq \hat{L}(\mathbf{u}\mathbf{v}^\intercal) + \frac{\lambda}{2}\|\mathbf{u}\|_F^2 + \frac{\lambda}{2}\|\mathbf{v}\|_F^2 + \langle P, QQ^\intercal \rangle,$$

where $Q = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} \in \mathbb{R}^{(m+n)\times r}$ and $P \in \mathbb{S}_+^{(m+n)}$ is positive semi-definite. Then, for almost all nonzero P (with respect to the Lebesgue measure on $\mathbb{S}_+^{(m+n)} \subset \mathbb{S}^{(m+n)} \cong \mathbb{R}^{\frac{(m+n)(m+n+1)}{2}}$), all SOSPs of $\hat{L}_{\lambda,P}$ are global minimizers of $\hat{L}_{\lambda,P}$.

To clarify, the conclusion that 'all SOSPs are global minimizers' holds with probability 1 even if the distribution of P is supported on $\{P \in \mathbb{S}_+^{(m+n)} : \|P\| \leq \varepsilon\}$ for arbitrarily small $\varepsilon > 0$. In the practical LoRA fine-tuning setup where no perturbation is used and $P = 0$ is set deterministically, Theorem 4.1 does not apply. However, we can nevertheless interpret the result of Theorem 4.1 to show that LoRA fine-tuning *generically* has no spurious local minima.

If we do use a randomly generated small perturbation P so that Theorem 4.1 applies, the solution to the perturbed problem with small P does not differ much from that of the unperturbed problem with $P = 0$ in the following sense.

Corollary 4.2. *Consider the setup of Theorem 4.1 and let $\varepsilon > 0$. Assume $\boldsymbol{\delta}_\lambda^\star \in \operatorname{argmin}_{\boldsymbol{\delta}} \hat{L}_\lambda(\boldsymbol{\delta})$. Assume P is randomly sampled with a probability distribution supported in*

$$\{P \in \mathbb{S}_+^{(m+n)} : \|P\|_F < \varepsilon\}$$

and is absolutely continuous with respect to the Lebesgue measure on $\mathbb{S}^{(m+n)} \cong \mathbb{R}^{\frac{(m+n)(m+n+1)}{2}}$. Then for any SOSP $(\hat{\mathbf{u}}, \hat{\mathbf{v}})$ of $\hat{L}_{\lambda,P}$

$$\begin{aligned}\hat{L}_\lambda(\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal) &\le \hat{L}(\boldsymbol{\delta}_\lambda^\star) + \lambda\|\boldsymbol{\delta}_\lambda^\star\|_* + 2\varepsilon\|\boldsymbol{\delta}_\lambda^\star\|_* \\ &= \min_{\boldsymbol{\delta}} \hat{L}_\lambda(\boldsymbol{\delta}) + 2\varepsilon\|\boldsymbol{\delta}_\lambda^\star\|_*.\end{aligned}$$

I.e., if $(\hat{\mathbf{u}}, \hat{\mathbf{v}})$ is an SOSP (and thus a global minimizer by Theorem 4.1) of the perturbed loss $\hat{L}_{\lambda,P}$, then it is an ε-approximate minimizer of the unperturbed loss $\hat{L}_\lambda$.

So if $\frac{r(r+1)}{2} > KN$, then Theorem 2.3, Theorem 2.4, and Corollary 4.2 together establish that (stochastic) gradient descent finds a $\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal$ such that its unperturbed empirical risk is ε-close to the the minimum unperturbed empirical risk.

4.1. Proof outlines

The proof is done by continuing our analysis of global minimum of $\hat{L}_\lambda(\boldsymbol{\delta})$. Given that low-rank solution exists, which we proved in the previous section, recall that LoRA training with weight decay is equivalent to solving

$$\operatorname*{argmin}_{\mathbf{u},\mathbf{v}} \hat{L}(\mathbf{u}\mathbf{v}^\intercal) + \frac{\lambda}{2}\|\mathbf{u}\|_F^2 + \frac{\lambda}{2}\|\mathbf{v}\|_F^2.$$

In this section, we relate SOSPs with global minimum, which opens the chance to find a global minimum by using gradient-based optimization methods. We start the analysis from the following lemma, which is a prior characterization of SOSPs in the matrix factorization.

Lemma 4.3. *(Theorem 2 of (Haeffele et al., 2014)) Let $G\colon \mathbb{S}_+^{(m+n)} \to \mathbb{R}$ be a twice differentiable convex function with compact level sets, $H\colon \mathbb{S}_+^{(m+n)} \to \mathbb{R}$ be a proper convex lower semi-continuous function, and $r > 0$. If the function $F\colon U \mapsto G(UU^\intercal) + H(UU^\intercal)$ defined over matrices $U \in \mathbb{R}^{(m+n)\times r}$ has a second order staionary point at a rank-deficient matrix U, then $UU^\intercal$ is a global minimum of $G + H$.*

We build our analysis upon Lemma 4.3. However, Lemma 4.3 is not directly applicable to our setting since it requires that the SOSP must be rank-deficient. However, this can be effectively circumvented by employing a perturbed empirical risk:

$$\operatorname*{minimize}_{\mathbf{u},\mathbf{v}} \hat{L}(\mathbf{u}\mathbf{v}^\intercal) + \frac{\lambda}{2}\|\mathbf{u}\|_F^2 + \frac{\lambda}{2}\|\mathbf{v}\|_F^2 + \langle P, QQ^\intercal\rangle,$$

where $Q = \begin{bmatrix}\mathbf{u}\\ \mathbf{v}\end{bmatrix}$, and P is a positive semi-definite matrix. Now we get the following lemma by applying Lemma 4.3 to the perturbed empricial risk.

Lemma 4.4. *Fix $\lambda \ge 0$. Assume $\hat{L}_\lambda(\boldsymbol{\delta})$ has a global minimum (not necessarily unique), $P \in \mathbb{S}_+^{(m+n)}$ is nonzero positive semi-definite, and $r > 0$. If $\hat{Q} = \begin{bmatrix}\hat{\mathbf{u}}\\ \hat{\mathbf{v}}\end{bmatrix} \in \mathbb{R}^{(m+n)\times r}$ is a rank deficient SOSP of*

$$\hat{L}_{\lambda,P}(\mathbf{u},\mathbf{v}) = \hat{L}(\mathbf{u}\mathbf{v}^\intercal) + \frac{\lambda}{2}\|\mathbf{u}\|_F^2 + \frac{\lambda}{2}\|\mathbf{v}\|_F^2 + \langle P, QQ^\intercal\rangle,$$

then $\hat{Q}$ is a global minimum of $\hat{L}_{\lambda,P}(\mathbf{u},\mathbf{v})$.

Proof. Define $G, H : \mathbb{S}_+^{(m+n)} \to \mathbb{R}$ to be

$$G(X) = \frac{\lambda}{2}\mathbf{tr}(X) + \langle P, X\rangle, \quad H(X) = \hat{L}(\bar{X})$$

where $\bar{X}$ is the off-diagonal submatrix of X defined in (2). Note that G has compact level set for every $\lambda \ge 0$ since $\mathbf{tr}(X) \ge 0$ and P, X are positive semi-definite, concluding that $\hat{Q}_{\lambda,P}$ is a global minimum of $F(Q) \triangleq G(QQ^\intercal) + H(QQ^\intercal) = \hat{L}_{\lambda,P}(\mathbf{u},\mathbf{v})$. □

We now give a detailed analysis of the proof of Theorem 4.1. The structure of the proof is inspired by the original work of Pataki (1998) and followed by Burer & Monteiro (2003); Boumal et al. (2016); Du & Lee (2018). The proof uses an application of Sard's theorem of differential geometry. The argument is captured in Lemma 4.5, and its proof is deferred to Appendix B.

Lemma 4.5. *Let $\mathcal{M}$ be m-dimensional smooth manifold embedded in $\mathbb{R}^d$ and V be a linear subspace of $\mathbb{R}^d$ with dimension n. If $m + n < d$, then the set*

$$\mathcal{M} + V = \{p + v : p \in \mathcal{M}, v \in V\}$$

has Lebesgue measure zero in $\mathbb{R}^d$.

Proof of Theorem 4.1. We show that second-order stationary point $\hat{Q}_{\lambda,P} = \begin{bmatrix}\hat{\mathbf{u}}\\ \hat{\mathbf{v}}\end{bmatrix}$ is rank-deficient for almost all positive semi-definite P, then use Lemma 4.4 to complete the proof. Denote $f^{(j)}$ for the j-th coordinate of f. For simplicity of notations, define

$$\hat{Y}_i^{(j)} \triangleq f_{\mathbf{W}_0}^{(j)}(X_i) + \langle \mathbf{G}^{(j)}(X_i), \mathbf{u}\mathbf{v}^\intercal\rangle,$$

and

$$v_i^{(j)} \triangleq \frac{1}{N}\frac{\partial}{\partial \hat{Y}_i^{(j)}}\ell(\hat{Y}_i, Y_i)$$

for $1 \le i \le N$ and $1 \le j \le K$, which depends on $\mathbf{u}$ and $\mathbf{v}$. Then for $v = \{v_i^{(j)}\} \in \mathbb{R}^{KN}$ define

$$S(v) \triangleq \sum_{i=1}^{N}\sum_{j=1}^{K} v_i^{(j)}\mathbf{G}^{(j)}(X_i) \in \mathbb{R}^{m\times n}.$$

Then by first-order gradient condition, we have

$$\Bigg(\underbrace{\begin{bmatrix}\mathbf{0} & S(v)\\ S(v)^\intercal & \mathbf{0}\end{bmatrix} + \lambda I + P}_{\triangleq M}\Bigg)\hat{Q}_{\lambda,P} = \mathbf{0}$$

We observe that the range of $\hat{Q}_{\lambda,P} \in \mathbb{R}^{(m+n)\times r}$ is in the nullspace of $M \in \mathbb{S}^{(m+n)}$. We now suppose $\hat{Q}_{\lambda,P}$ has full rank, i.e., $\mathrm{rank}(\hat{Q}_{\lambda,P}) = r$. Hence, we have the following inequality:

$$r = \mathrm{rank}(\hat{Q}_{\lambda,P}) \le \dim \mathcal{N}(M) \le m+n$$

Now for $r \le s \le m+n$ and $s \in \mathbb{Z}$, define

$$\mathcal{A}_s = \Big\{P : P = M - \lambda I, M \in \mathbb{S}^{(m+n)}, \dim\mathcal{N}(M) = s\Big\}.$$

Then from Proposition 2.1 of (Helmke & Shayman, 1995), $\mathcal{A}_s$ is a smooth manifold embedded in $\mathbb{R}^{\frac{(m+n)(m+n+1)}{2}} \cong \mathbb{S}^{(m+n)}$ with dimension

$$\dim\mathcal{A}_s = \frac{(m+n+1)(m+n)}{2} - \frac{s(s+1)}{2}.$$

Now by definition of P, we know that

$$P \in \bigcup_{s=r}^{m+n}(\mathcal{A}_s + \mathcal{R}(S))$$

where " $+$ " is the set-sum (Minkowski sum) and $\mathcal{R}(S)$ is the range of $S(v)$ in $\mathbb{R}^{\frac{(m+n)(m+n+1)}{2}}$ for any $v \in \mathbb{R}^{KN}$. The dimensions can be bounded by

$$\dim\mathcal{A}_s \le \frac{(m+n)(m+n+1)}{2} - \frac{r(r+1)}{2}$$

for $r \le s \le m+n$ and

$$\dim\mathcal{R}(S) \le KN.$$

Therefore given that $\frac{r(r+1)}{2} > KN$, we have

$$\dim\mathcal{A}_s + \dim\mathcal{R}(S) < \frac{(m+n)(m+n+1)}{2}.$$

Then, by Lemma 4.5, which is effectively an application of Sard's theorem, we can conclude $\mathcal{A}_s + \mathcal{R}(S)$ is a measure-zero set, and the finite union of such measure-zero sets is measure-zero. This implies that every P that makes $\hat{Q}_{\lambda,P}$ to be of full rank must be chosen from measure-zero subset of $\mathbb{S}_+^{(m+m)} \subset \mathbb{S}^{(m+n)}$. Therefore we may conclude that $\mathrm{rank}(\hat{Q}_{\lambda,P}) < r$ for almost every nonzero positive semi-definite P. □

Proof of Corollary 4.2. Assume $\boldsymbol{\delta}_\lambda^\star \in \mathrm{argmin}_{\boldsymbol{\delta}}\, \hat{L}_\lambda(\boldsymbol{\delta})$. We observe the following chain of inequalities.

$$\begin{aligned}\hat{L}(\hat{\boldsymbol{\delta}}) + \lambda\|\hat{\boldsymbol{\delta}}\|_* &\le \hat{L}(\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal) + \frac{\lambda}{2}\|\hat{\mathbf{u}}\|_F^2 + \frac{\lambda}{2}\|\hat{\mathbf{v}}\|_F^2\\ &\le \hat{L}(\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal) + \frac{\lambda}{2}\|\hat{\mathbf{u}}\|_F^2 + \frac{\lambda}{2}\|\hat{\mathbf{v}}\|_F^2 + \langle P, \hat{Q}\hat{Q}^\intercal\rangle\\ &= \hat{L}_{\lambda,P}(\hat{\mathbf{u}}, \hat{\mathbf{v}}),\end{aligned}$$

where the first inequality of is from Lemma 2.2, the second is from P and $\hat{Q}\hat{Q}^\intercal$ being positive semi-definite. On the other hand, we can find $\mathbf{u}^\star$ and $\mathbf{v}^\star$ such that $\boldsymbol{\delta}_\lambda^\star = \mathbf{u}^\star\mathbf{v}^{\star\intercal}$ and $\|\boldsymbol{\delta}_\lambda^\star\|_* = \frac{1}{2}(\|\mathbf{u}^\star\|_F^2 + \|\mathbf{v}^\star\|_F^2)$ by using Lemma 2.2. Now take $Q^\star = \begin{bmatrix}\mathbf{u}^\star\\ \mathbf{v}^\star\end{bmatrix}$, then we get

$$\begin{aligned}\hat{L}_{\lambda,P}(\mathbf{u}^\star, \mathbf{v}^\star) &= \hat{L}(\boldsymbol{\delta}_\lambda^\star) + \lambda\|\boldsymbol{\delta}_\lambda^\star\|_* + \langle P, Q^\star Q^{\star\intercal}\rangle\\ &\le \hat{L}(\boldsymbol{\delta}_\lambda^\star) + \lambda\|\boldsymbol{\delta}_\lambda^\star\|_* + \varepsilon\|Q^\star Q^{\star\intercal}\|_F\\ &\le \hat{L}(\boldsymbol{\delta}_\lambda^\star) + \lambda\|\boldsymbol{\delta}_\lambda^\star\|_* + \varepsilon\|Q^\star\|_F^2\\ &= \hat{L}(\boldsymbol{\delta}_\lambda^\star) + \lambda\|\boldsymbol{\delta}_\lambda^\star\|_* + \varepsilon\|\mathbf{u}^\star\|_F^2 + \varepsilon\|\mathbf{v}^\star\|_F^2\\ &= \hat{L}(\boldsymbol{\delta}_\lambda^\star) + \lambda\|\boldsymbol{\delta}_\lambda^\star\|_* + 2\varepsilon\|\boldsymbol{\delta}_\lambda^\star\|_*,\end{aligned}$$

where the first inequality is Cauchy–Schwartz inequality, and the second inequality is from sub-multiplicativity of $\|\cdot\|_F$. Moreover by Theorem 4.1,

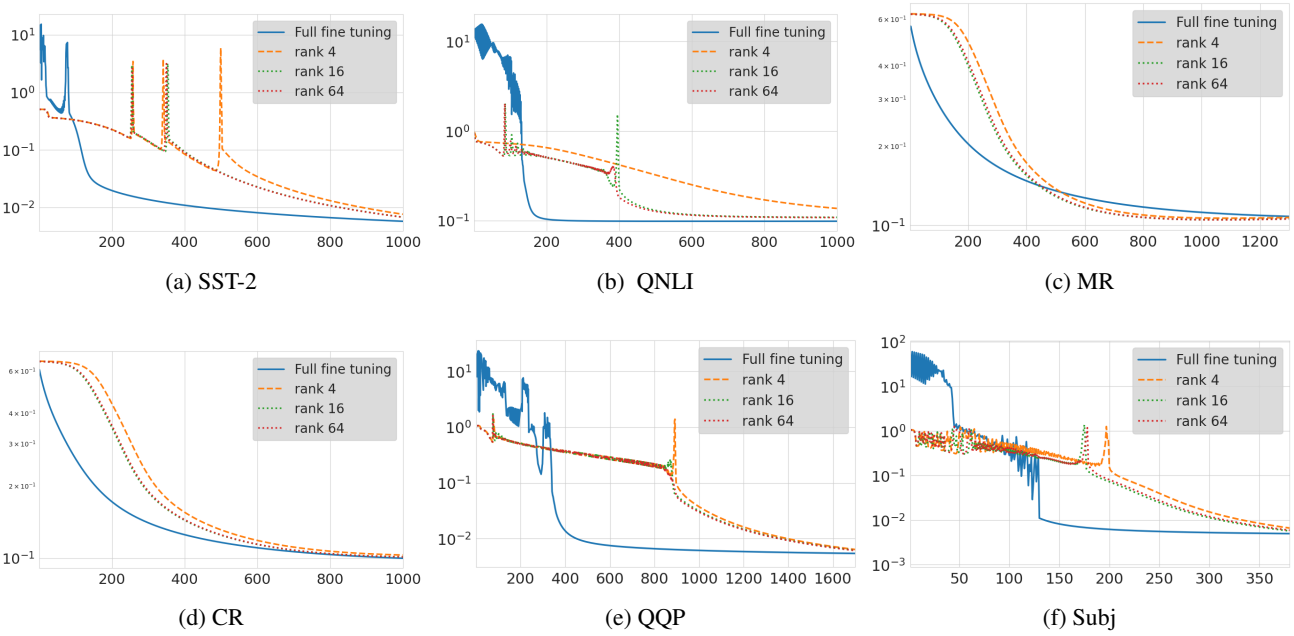
$$\hat{L}_{\lambda,P}(\hat{\mathbf{u}}^\star, \hat{\mathbf{v}}^\star) \le \hat{L}_{\lambda,P}(\mathbf{u}^\star, \mathbf{v}^\star),$$

and this happens for almost sure, since we sampled P from a probability distribution which is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}^{\frac{(m+n)(m+n+1)}{2}} \cong \mathbb{S}^{(m+n)}$. □

5. Low-rank LoRA solution generalizes well

In this section, we establish a generalization guarantee for the low-rank solution obtained by minimizing the perturbed loss $\hat{L}_{\lambda,P}$ of Theorem 4.1. For simplicity, we restrict the following main result to the cross-entropy loss. Generalization guarantees for general convex, non-negative, and twice continuously differentiable losses, are provided as Theorem C.6 in Appendix C.

Theorem 5.1. *Assume ℓ is cross-entropy loss. Assume the population risk L has a minimizer (not necessarily unique) and denote it as $\boldsymbol{\delta}_{\mathrm{true}}^\star \in \mathrm{argmin}_{\boldsymbol{\delta}}\, L(\boldsymbol{\delta})$. Assume $\boldsymbol{\delta}_{\mathrm{true}}^\star \neq \mathbf{0}$. For $1 \le j \le K$, suppose $\|\mathbf{G}^{(j)}(X)\|_F \le R$ almost surely*



(a) SST-2 (b) QNLI (c) MR

(d) CR (e) QQP (f) Subj

Figure 2. Training curves (training loss vs. epochs) on different NLP tasks.

with respect to the random data $X \sim \mathcal{P}$. Let $\varepsilon > 0$, $\eta \in (0,1)$, and

$$\lambda = \frac{2(2+\varepsilon)\sqrt{K}R}{\sqrt{N}}\left(2+\sqrt{\log\frac{1}{\eta}}\right).$$

Write $\boldsymbol{\delta}_\lambda^\star$ to denote a minimizer (not necessarily unique) of $\hat{L}_\lambda(\boldsymbol{\delta})$. Consider the setup of Corollary 4.2 with P randomly sampled with a probability distribution supported in

$$\left\{P \in \mathbb{S}_+^{(m+n)} : \|P\|_F < \frac{\varepsilon\lambda\|\boldsymbol{\delta}_{\text{true}}^\star\|_*}{2\|\boldsymbol{\delta}_\lambda^\star\|_*}\right\}$$

and is absolutely continuous with respect to the Lebesgue measure on $\mathbb{S}^{(m+n)} \cong \mathbb{R}^{\frac{(m+n)(m+n+1)}{2}}$. Let $(\hat{\mathbf{u}}, \hat{\mathbf{v}})$ be an SOSP of $\hat{L}_{\lambda,P}$. Then with probability greater than $1-\eta$,

$$L(\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal) - L(\boldsymbol{\delta}_{\text{true}}^\star) < \|\boldsymbol{\delta}_{\text{true}}^\star\|_* \frac{2(2+\varepsilon)^2\sqrt{K}R}{\sqrt{N}}\left(2+\sqrt{\log\frac{1}{\eta}}\right)$$

In the context of fine-tuning, where the target task is closely related to the pre-training task, it is natural to assume that $\boldsymbol{\delta}_{\text{true}}^\star$ in Theorem 5.1 is "small". The proof, deferred to Appendix C, utilizes standard arguments with Rademacher complexity.

6. Experiments

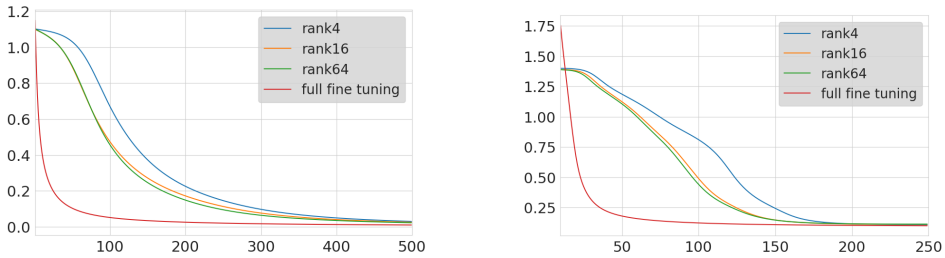
In this section, we conduct simple experiments on fine-tuning linearized pre-trained models to validate our theory.[1]

[1]Code available at https://github.com/UijeongJang/LoRA-NTK.

Experimental setup on NLP tasks. We use prompt-based fine-tuning (Schick & Schütze, 2021; Gao et al., 2021) and consider the same architecture and dataset as in (Malladi et al., 2023), which empirically verifies that with prompt-based fine-tuning, the fine-tuning dynamics stay within the NTK regime. We present the results of six NLP tasks that were also considered in (Malladi et al., 2023): sentiment analysis (SST-2, MR, CR), natural language inference (QNLI), subjectivity (Subj), and paraphrase detection (QQP). We optimize a linearized RoBERTa-base (Liu et al., 2019) model with dataset of size 32 ($N = 32$) with two labels ($K = 2$) using cross entropy loss. With LoRA rank $r \geq 11$, our theory guarantees that no spurious local minima exist. For a baseline comparison, we also perform full fine-tuning (without LoRA) on the linearized model. The training curves are presented in Figure 2, and additional details are provided in Appendix D. Results showing test accuracy are also presented in Appendix D.

Experimental setup on image and speech classification tasks. We use a pre-trained vision transformer (Dosovitskiy et al., 2021) and fine-tune it on the bean disease dataset (Makerere AI Lab, 2020) to perform an image classification task with 3 labels. We use dataset of size 48 with three labels. Similar to our experiments on NLP tasks, we find that training curves converge to the same loss value, where the rates of convergence differ.

For speech classification, we use a pre-trained wav2vec2 (Baevski et al., 2020) model and fine-tune it on a SUPERB dataset (Yang et al., 2021) to perform a speech classification



(a) Image classification

(b) Speech classification

Figure 3. Training curves (training loss vs. epochs) on image and speech classification tasks.

task with 4 labels. We use a dataset of size 64 with four labels. We also find that the training curves converge to the same loss value. The details are the same as with the image classification task.

The training curves of both image and speech data are presented in Figure 3, and additional details are provided in Appendix D.

Empirical observation. The experiments validate our theory as the training curves converge to the same globally optimal loss value. However, we do observe that the *rates* of convergence differ. When the LoRA rank is higher or when full fine-tuning is performed and LoRA is not used, fine-tuning converges faster. Indeed, our theory ensures that spurious local minima do not exist, but it says nothing about how convex or favorable the landscape may or may not be. Our intuitive hypothesis is that using lower LoRA rank creates unfavorable regions of the loss landscape, such as plateaus or saddle points, and they slow down the gradient descent dynamics.

If this hypothesis is generally true, we face an interesting tradeoff: lower LoRA rank reduces memory cost and per-iteration computation cost but increases the number of iterations needed for convergence. Then, using a very low LoRA rank may be suboptimal not due to representation power, presence of spurious local minima, or poor generalization guarantees, but rather due to unfavorable flat training landscapes slowing down convergence. Exploring this phenomenon and designing remedies is an interesting direction for future work.

7. Conclusion

In this work, we present theoretical guarantees on the trainability and generalization capabilities of LoRA fine-tuning of pre-trained models. Together with the work of Zeng & Lee (2024), our results represent a first step in theoretically analyzing the LoRA fine-tuning dynamics of pre-trained models by presenting guarantees (upper bounds). For future work, carrying out further refined analyses under more specific assumptions, relaxing the linearization/NTK regime assumption through a local analysis, better understanding the minimum rank requirement through lower bounds, and, motivated by the observation of Section 6, analyzing the tradeoff between training rate and LoRA rank are exciting directions.

Acknowledgments

UJ and EKR were supported by the Samsung Science and Technology Foundation (Project Number SSTF-BA2101-02) and the National Research Foundation of Korea (NRF) Grant funded by the Korean Government (MSIP) [NRF-2022R1C1C1010010]. JDL acknowledges support of the NSF CCF 2002272, NSF IIS 2107304, and NSF CAREER Award 2144994. We thank Jungsoo Kang for the discussion on the proof of Lemma 4.5. We also thank Jisun Park for providing valuable feedback.

Impact statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none which we feel must be specifically highlighted here.

References

Aghajanyan, A., Zettlemoyer, L., and Gupta, S. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *Association for Computational Linguistics*, 2021.

Allen-Zhu, Z., Li, Y., and Song, Z. A convergence theory for deep learning via over-parameterization. *International Conference on Machine Learning*, 2019a.

Allen-Zhu, Z., Li, Y., and Song, Z. On the convergence rate of training recurrent neural networks. *Neural Information Processing Systems*, 2019b.

Arora, S., Du, S. S., Hu, W., Li, Z., Salakhutdinov, R. R., and Wang, R. On exact computation with an infinitely

wide neural net. *Neural Information Processing Systems*, 2019.

Bach, F. *Learning Theory from First Principles*. Draft, 2023.

Bach, F., Mairal, J., and Ponce, J. Convex sparse matrix factorizations. *arXiv preprint arXiv:0812.1869*, 2008.

Baevski, A., Zhou, Y., Mohamed, A., and Auli, M. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Neural Information Processing Systems*, 2020.

Barron, A. R. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information theory*, 39(3):930–945, 1993.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Bartlett, P. L., Boucheron, S., and Lugosi, G. Model selection and error estimation. *Machine Learning*, 48:85–113, 2002.

Bartlett, P. L., Bousquet, O., and Mendelson, S. Local rademacher complexities. *The Annals of Statistics*, 33(4): 1497–1537, 2005.

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. Spectrally-normalized margin bounds for neural networks. *Neural Information Processing Systems*, 2017.

Barvinok, A. I. Problems of distance geometry and convex properties of quadratic maps. *Discrete & Computational Geometry*, 13:189–202, 1995.

Bengio, Y. and Delalleau, O. On the expressive power of deep architectures. *Algorithmic Learning Theory*, 2011.

Bhojanapalli, S., Boumal, N., Jain, P., and Netrapalli, P. Smoothed analysis for low-rank solutions to semidefinite programs in quadratic penalty form. *Conference On Learning Theory*, 2018.

Boix-Adsera, E., Littwin, E., Abbe, E., Bengio, S., and Susskind, J. Transformers learn through gradual rank increase. *Neural Information Processing Systems*, 2023.

Boumal, N., Voroninski, V., and Bandeira, A. The non-convex Burer–Monteiro approach works on smooth semidefinite programs. *Neural Information Processing Systems*, 29, 2016.

Bousquet, O. and Elisseeff, A. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

Burer, S. and Monteiro, R. D. A nonlinear programming algorithm for solving semidefinite programs via low-rank factorization. *Mathematical Programming*, 95(2):329–357, 2003.

Cabral, R., De la Torre, F., Costeira, J. P., and Bernardino, A. Unifying nuclear norm and bilinear factorization approaches for low-rank matrix decomposition. *International Conference on Computer Vision*, 2013.

Chen, Z., Cao, Y., Gu, Q., and Zhang, T. A generalized neural tangent kernel analysis for two-layer neural networks. *Neural Information Processing Systems*, 2020.

Choi, J. Y., Park, J., Park, I., Cho, J., No, A., and Ryu, E. K. LoRA can replace time and class embeddings in diffusion probabilistic models. *NeurIPS 2023 Workshop on Diffusion Models*, 2023.

Cybenko, G. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Delalleau, O. and Bengio, Y. Shallow vs. deep sum-product networks. *Neural Information Processing Systems*, 2011.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: efficient finetuning of quantized llms. *Neural Information Processing Systems*, 2023.

Dinh, L., Pascanu, R., Bengio, S., and Bengio, Y. Sharp minima can generalize for deep nets. *International Conference on Machine Learning*, 2017.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *International Conference on Learning Representations*, 2021.

Du, S. and Lee, J. On the power of over-parametrization in neural networks with quadratic activation. *International Conference on Machine Learning*, 2018.

Du, S., Lee, J., Li, H., Wang, L., and Zhai, X. Gradient descent finds global minima of deep neural networks. *International Conference on Machine Learning*, 2019.

Du, S. S., Jin, C., Lee, J. D., Jordan, M. I., Singh, A., and Poczos, B. Gradient descent can take exponential time to escape saddle points. *Neural Information Processing Systems*, 2017.

Duan, Y., Ji, G., Cai, Y., et al. Minimum width of leaky-relu neural networks for uniform universal approximation. *International Conference on Machine Learning*, 2023.

Fu, Z., Yang, H., So, A. M.-C., Lam, W., Bing, L., and Collier, N. On the effectiveness of parameter-efficient fine-tuning. *AAAI Conference on Artificial Intelligence*, 2023.

Gao, T., Fisch, A., and Chen, D. Making pre-trained language models better few-shot learners. *Association for Computational Linguistics*, 2021.

Ge, R., Huang, F., Jin, C., and Yuan, Y. Escaping from saddle points—online stochastic gradient for tensor decomposition. *Conference on Learning Theory*, 2015.

Ge, R., Lee, J. D., and Ma, T. Matrix completion has no spurious local minimum. *Neural Information Processing Systems*, 2016.

Ghadimi, S. and Lan, G. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

Giannou, A., Rajput, S., Sohn, J.-y., Lee, K., Lee, J. D., and Papailiopoulos, D. Looped transformers as programmable computers. *International Conference on Machine Learning*, 2023.

Haeffele, B., Young, E., and Vidal, R. Structured low-rank matrix factorization: optimality, algorithm, and applications to image processing. *International Conference on Machine Learning*, 2014.

Hardt, M., Recht, B., and Singer, Y. Train faster, generalize better: stability of stochastic gradient descent. *International Conference on Machine Learning*, 2016.

Helmke, U. and Shayman, M. A. Critical points of matrix least squares distance functions. *Linear Algebra and its Applications*, 215:1–19, 1995.

Hornik, K., Stinchcombe, M., and White, H. Universal approximation of an unknown mapping and its derivatives using multilayer feedforward networks. *Neural Networks*, 3(5):551–560, 1990.

Hu, E. J., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. LoRA: low-rank adaptation of large language models. *International Conference on Learning Representations*, 2021.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: convergence and generalization in neural networks. *Neural Information Processing Systems*, 2018.

Jin, C., Ge, R., Netrapalli, P., Kakade, S. M., and Jordan, M. I. How to escape saddle points efficiently. *International Conference on Machine Learning*, 2017.

Jin, J., Li, Z., Lyu, K., Du, S. S., and Lee, J. D. Understanding incremental learning of gradient descent: A fine-grained analysis of matrix sensing. *International Conference on Machine Learning*, 2023.

Johnson, W. and Lindenstrauss, J. Extensions of lipschitz maps into a hilbert space. *Contemporary Mathematics*, 26:189–206, 1984.

Koltchinskii, V. and Panchenko, D. Rademacher processes and bounding the risk of function learning. In Giné, E., Mason, D. M., and Wellner, J. A. (eds.), *High Dimensional Probability II*, pp. 443–457. Springer, 2000.

Lee, J. D., Simchowitz, M., Jordan, M. I., and Recht, B. Gradient descent only converges to minimizers. *Conference on Learning Theory*, 2016.

Lialin, V., Muckatira, S., Shivagunde, N., and Rumshisky, A. ReLoRA: high-rank training through low-rank updates. *Workshop on Advancing Neural Network Training (WANT): Computational Efficiency, Scalability, and Resource Optimization*, 2023.

Liu, L., Liu, X., Gao, J., Chen, W., and Han, J. Understanding the difficulty of training transformers. *Empirical Methods in Natural Language Processing*, 2020.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. RoBERTa: a robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Lu, Z., Pu, H., Wang, F., Hu, Z., and Wang, L. The expressive power of neural networks: a view from the width. *Neural Information Processing Systems*, 2017.

Makerere AI Lab. Bean disease dataset, 2020. URL https://github.com/AI-Lab-Makerere/ibean/.

Malladi, S., Wettig, A., Yu, D., Chen, D., and Arora, S. A kernel-based view of language model fine-tuning. *International Conference on Machine Learning*, 2023.

Maurer, A. A vector-contraction inequality for rademacher complexities. *Algorithmic Learning Theory*, 2016.

McDiarmid, C. et al. On the method of bounded differences. *Surveys in Combinatorics*, 141(1):148–188, 1989.

Pataki, G. On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of Operations Research*, 23(2):339–358, 1998.

Pataki, G. The geometry of semidefinite programming. In Wolkowicz, H., Saigal, R., and Vandenberghe, L. (eds.), *Handbook of Semidefinite Programming: Theory, Algorithms, and Applications*, pp. 29–65. Springer, 2000.

Pilanci, M. and Ergen, T. Neural networks are convex regularizers: exact polynomial-time convex optimization formulations for two-layer networks. *International Conference on Machine Learning*, 2020.

Polyak, B. T. *Introduction to Optimization*. New York, Optimization Software, 1987.

Recht, B., Fazel, M., and Parrilo, P. A. Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501, 2010.

Ryu, S. Low-rank adaptation for fast text-to-image diffusion fine-tuning, 2023. URL https://github.com/cloneofsimo/lora.

Schick, T. and Schütze, H. Exploiting cloze questions for few shot text classification and natural language inference. *Association for Computational Linguistics*, 2021.

Smith, J. S., Hsu, Y.-C., Zhang, L., Hua, T., Kira, Z., Shen, Y., and Jin, H. Continual diffusion: continual customization of text-to-image diffusion with c-lora. *arXiv preprint arXiv:2304.06027*, 2023.

Soltanolkotabi, M., Javanmard, A., and Lee, J. D. Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*, 65(2):742–769, 2018.

Sridharan, K., Shalev-Shwartz, S., and Srebro, N. Fast rates for regularized objectives. *Neural Information Processing Systems*, 21, 2008.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Neural Information Processing Systems*, 2017.

Wei, A., Hu, W., and Steinhardt, J. More than a toy: random matrix models predict how real-world neural representations generalize. *International Conference on Machine Learning*, 2022a.

Wei, C., Chen, Y., and Ma, T. Statistically meaningful approximation: a case study on approximating turing machines with transformers. *Neural Information Processing Systems*, 2022b.

Wu, L., Zhu, Z., et al. Towards understanding generalization of deep learning: perspective of loss landscapes. *arXiv preprint arXiv:1706.10239*, 2017.

Yang, S.-w., Chi, P.-H., Chuang, Y.-S., Lai, C.-I. J., Lakhotia, K., Lin, Y. Y., Liu, A. T., Shi, J., Chang, X., Lin, G.-T., et al. Superb: Speech processing universal performance benchmark. *Interspeech*, 2021.

Yeh, S.-Y., Hsieh, Y.-G., Gao, Z., Yang, B. B., Oh, G., and Gong, Y. Navigating text-to-image customization: from LyCORIS fine-tuning to model evaluation. *International Conference on Learning Representations*, 2024.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S. J., and Kumar, S. Are transformers universal approximators of sequence-to-sequence functions? *International Conference on Learning Representations*, 2019.

Zeng, Y. and Lee, K. The expressive power of low-rank adaptation. *International Conference on Learning Representations*, 2024.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. Understanding deep learning (still) requires rethinking generalization. *Communications of the ACM*, 64(3):107–115, 2021.

Zhang, J., Karimireddy, S. P., Veit, A., Kim, S., Reddi, S., Kumar, S., and Sra, S. Why are adaptive methods good for attention models? *Neural Information Processing Systems*, 2020.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. Stochastic gradient descent optimizes over-parameterized deep relu networks. *Machine Learning*, 109(3):467–492, 2020.

A. Omitted proof of Theorem 3.1

Here, we explain the details in the proof of Theorem 3.1. We first prove the equivalence of

$$\underset{\boldsymbol{\delta}\in\mathbb{R}^{m\times n}}{\text{minimize}} \quad \hat{L}(\boldsymbol{\delta}) + \lambda\|\boldsymbol{\delta}\|_* \tag{P}$$

and

$$\underset{Z\in\mathbb{S}_+^{(m+n)}}{\text{minimize}} \quad \hat{L}(\bar{Z}) + \frac{\lambda}{2}\mathbf{tr}(Z) \tag{Q}$$

where $\bar{Z} = Z[1:m, m+1:m+n] \in \mathbb{R}^{m\times n}$. I.e., $\bar{Z}$ is a off-diagonal submatrix of X such that

$$Z = \begin{bmatrix} * & \bar{Z} \\ \bar{Z}^\intercal & * \end{bmatrix}.$$

Lemma A.1. *The following two statements hold.*

1. *Fix $\lambda \geq 0$ and suppose (P) has a global minimizer (not necessarily unique). Let $\boldsymbol{\delta}_\lambda^\star \in \mathbb{R}^{m\times n}$ be a global minimizer of (P). Then there exists an $Z_\lambda^\star \in \mathbb{S}_+^{(m+n)}$ induced from $\boldsymbol{\delta}_\lambda^\star$ such that $Z_\lambda^\star$ is a global minimizer of (Q), $\mathrm{rank}(Z_\lambda^\star) = \mathrm{rank}(\boldsymbol{\delta}_\lambda^\star)$, and has same objective value.*

2. *Fix $\lambda \geq 0$ and suppose (Q) has a global minimizer (not necessarily unique). Let $Z_\lambda^\star \in \mathbb{S}_+^{(m+n)}$ be a global minimum of (Q). Then $\bar{Z}_\lambda^\star \in \mathbb{R}^{m\times n}$ is a global minimizer of (P) such that $\mathrm{rank}(\bar{Z}_\lambda^\star) = \min(m, n, \mathrm{rank}(Z_\lambda^\star))$ and has same objective value.*

Proof. We prove the two statements at once. Let $\boldsymbol{\delta}_\lambda^\star \in \mathbb{R}^{m\times n}$ be a global minimizer of (P) and let $r = \mathrm{rank}(\boldsymbol{\delta}_\lambda^\star)$. Then by Lemma 2.2, there exists $\mathbf{u} \in \mathbb{R}^{m\times r}$ and $\mathbf{v} \in \mathbb{R}^{n\times r}$ such that $\|\boldsymbol{\delta}_\lambda^\star\|_* = \frac{1}{2}(\|\mathbf{u}\|_F^2 + \|\mathbf{v}\|_F^2)$ and $\mathbf{u}\mathbf{v}^\intercal = \boldsymbol{\delta}_\lambda^\star$. Take

$$Z_\lambda^\star = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} \begin{bmatrix} \mathbf{u}^\intercal & \mathbf{v}^\intercal \end{bmatrix} = \begin{bmatrix} \mathbf{u}\mathbf{u}^\intercal & \mathbf{u}\mathbf{v}^\intercal \\ \mathbf{v}\mathbf{u}^\intercal & \mathbf{v}\mathbf{v}^\intercal \end{bmatrix} \in \mathbb{S}_+^{(m+n)}.$$

Then since

$$\mathbf{tr}(Z_\lambda^\star) = \|Z_\lambda^\star\|_* = \left\| \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} \right\|_F^2 = \|\mathbf{u}\|_F^2 + \|\mathbf{v}\|_F^2 = 2\|\boldsymbol{\delta}_\lambda^\star\|_*,$$

(Q) with $Z_\lambda^\star$ has the same objective value with (P) with $\boldsymbol{\delta}_\lambda^\star$ and $\mathrm{rank}(\boldsymbol{\delta}_\lambda^\star) = \mathrm{rank}(Z_\lambda^\star) = r$. Conversely, let $Z_\lambda^\star \in \mathbb{S}_+^{(m+n)}$ be a global minimizer of (Q) and let $\mathrm{rank}(Z_\lambda^\star) = r$. Note that r may be larger than m or n. Then there exists $Q = \begin{bmatrix} \mathbf{u} \\ \mathbf{v} \end{bmatrix} \in \mathbb{R}^{(m+n)\times r}$ such that $QQ^\intercal = Z_\lambda^\star$. Then since

$$\mathbf{tr}(Z_\lambda^\star) = \|Z_\lambda^\star\|_* = \|Q\|_F^2 = \|\mathbf{u}\|_F^2 + \|\mathbf{v}\|_F^2 \geq 2\|\mathbf{u}\mathbf{v}^\intercal\|_* = 2\|\bar{Z}_\lambda^\star\|_*,$$

the objective value of (P) with $\bar{Z}_\lambda^\star \in \mathbb{R}^{m\times n}$ has less than or equal to minimum objective value of (Q) and $\mathrm{rank}(\bar{Z}_\lambda^\star) = \min(m, n, r)$.

If there exists $m \times n$ matrix whose objective value of (P) is strictly less than the minimum objective value of (Q), then we repeat the same step that was applied on $\boldsymbol{\delta}_\lambda^\star$ to induce a solution of (Q) with strictly less objective value, which is a contradiction. Conversely, if there exists positive semi-definite matrix of size $m+n$ whose objective value of (Q) is strictly less than the minimum objective value of (P), then we repeat the same step applied on $Z_\lambda^\star$ to induce a solution of (P) with strictly less objective value, which is also a contradiction. Therefore if one of (P) and (Q) has a global minimizer, the other must have a global minimizer with same objective value. □

Next lemma states that if the rank of the global minimizer of (Q) is sufficiently large, then we can find an another solution with strictly less rank.

Lemma A.2. *Suppose $X \in \mathbb{S}_+^n$ and let $Z \in \mathbb{S}^n$ be a nonzero symmetric matrix such that $\mathcal{R}(Z) \subseteq \mathcal{R}(X)$. Then there exists nonzero $t^* \in \mathbb{R}$ such that $X + t^*Z$ is positive semi-definite and $\mathrm{rank}(X + t^*Z) < \mathrm{rank}(X)$.*

Proof. Let $r = \mathrm{rank}(X)$. Suppose $Q \in \mathbb{R}^{n\times r}$ is a matrix where its columns are basis to $\mathcal{R}(X)$. Now suppose $\mu_1(Q^\intercal(X + tZ)Q) > 0$ for all $t \in \mathbb{R}$ where $\mu_1(\cdot)$ denotes the smallest eigenvalue (note that $\mu_1(\cdot)$ is continuous). Then $Q^\intercal(X+tZ)Q \in \mathbb{S}^r$ should be positive definite for all t. For contradiction, take $v \in \mathcal{R}(Z) \subseteq \mathcal{R}(X) = \mathcal{R}(Q)$ to be an eigenvector of nonzero eigenvalue of Z. Since $v^\intercal Xv > 0$ and $v^\intercal Zv \neq 0$, there exists some t such that $v^\intercal(X + tZ)v < 0$. Now take $w \in \mathbb{R}^r$ such that $Qw = v$. Then it follows that

$$w^\intercal(Q^\intercal(X + tZ)Q)w < 0,$$

which is a contradiction. This implies that there exists $t^\star \neq 0$ such that

$$\mu_1(Q^\intercal(X + t^*Z)Q) = 0,$$

Hence we have

$$r > \mathrm{rank}(Q^\intercal(X + t^*Z)Q) = \mathrm{rank}(X + t^*Z)$$

and $Q^\intercal(X + t^*Z)Q$ is positive semi-definite. To show that $X + t^*Z$ is positive semi-definite, take any $x \in \mathbb{R}^n$ and consider the decomposition $x = Qy + z$ where $y \in \mathbb{R}^r$ and $z \in \mathcal{N}(Q) = \mathcal{N}(X) \subseteq \mathcal{N}(Z)$. Then, we have

$$\begin{aligned} y^\intercal(X + t^\star Z)y &= (y^\intercal Q^\intercal + z^\intercal)(X + t^*Z)(Qy + z) \\ &= y^\intercal Q^\intercal(X + t^*Z)Qy \geq 0. \end{aligned}$$

□

Finally, the following lemma and its proof are similar to the previous one, but we state it separately for the sake of clarity. It will be used in the proof of Theorem 3.1.

Lemma A.3. *Suppose $X \in \mathbb{S}^n_+$ which is nonzero and let $Z \in \mathbb{S}^n$ be a nonzero symmetric matrix such that $\mathcal{R}(Z) \subseteq \mathcal{R}(X)$. Then there exists $t^* > 0$ such that $X \pm t^*Z$ is positive semi-definite.*

Proof. Let $\mathrm{rank}(X) = r$ and $\{y_1, \dots, y_r\}$ be orthonormal eigenvectors of nonzero eigenvalues of X. Since $y_i^\intercal Xy_i > 0$ for all y_i, $i = 1, \cdots, r$, there exists an interval $(-a_i, a_i)$ for $a_i > 0$ such that $y_i^\intercal(X \pm tZ)y_i \geq 0$ for $t \in (-a_i, a_i)$. Take $t^* = \min\{a_1, \dots, a_r\}$. Then t^* satisfies the statement of the theorem. □

Now we provide the complete proof of Theorem 3.1.

Proof of Theorem 3.1. Suppose $Z^\star_\lambda \in \mathbb{S}^{(m+n)}_+$ is a global minimizer of

$$F(Z) = \hat{L}(\bar{Z}) + \frac{\lambda}{2}\mathbf{tr}(Z) = \frac{1}{N}\sum_{i=1}^{N} \ell\left(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \bar{Z}\rangle, Y_i\right) + \frac{\lambda}{2}\mathbf{tr}(Z)$$

which is induced from $\boldsymbol{\delta}^\star_\lambda \in \mathbb{R}^{m\times n}$ by Lemma A.1. Suppose there exists nonzero symmetric matrix Z such that $Z \in \mathcal{S}(Z^\star_\lambda) \triangleq \{Z \in \mathbb{S}^{(m+n)} : \mathcal{R}(Z) \subseteq \mathcal{R}(Z^\star_\lambda)\}$ and $\langle \mathbf{G}(X_i), Z\rangle = \mathbf{0}$ for $1 \leq i \leq N$. In other words, $Z \in \mathcal{S}(Z^\star_\lambda) \cap \mathcal{N}(\mathcal{A})$ where $\mathcal{A}\colon \mathbb{S}^{(m+n)} \to \mathbb{R}^{KN}$ is a linear operator defined as

$$\mathcal{A}(Z)_{ij} = \langle \mathbf{G}^{(j)}(X_i), \bar{Z}\rangle, \qquad 1 \leq i \leq N, \quad 1 \leq j \leq K.$$

Then there exists $t > 0$ such that $Z^\star_\lambda \pm tZ$ is positive semi-definite by Lemma A.3, since $Z^\star$ must be nonzero. Therefore $\mathbf{tr}(Z) = 0$, otherwise it will contradict the minimality of $Z^\star_\lambda$. Also we know that there exists nonzero $t^* \in \mathbb{R}$ such that $Z^\star_\lambda + t^*Z$ is also positive semi-definite with strictly lower rank by Lemma A.2. Since $\mathbf{tr}(Z) = 0$, $Z^\star_\lambda + t^*Z$ is also a global minimizer of F. Replace $Z^\star_\lambda$ with $Z^\star_\lambda + tZ$ and repeat this process until we find a solution $Z^\star_\lambda$ with

$$\{\mathbf{0}\} = \mathcal{S}(Z^\star_\lambda) \cap \mathcal{N}(\mathcal{A}).$$

Now we let $\mathrm{rank}(Z^\star_\lambda) = r$. Then by dimension counting, we have the following inequality.

$$\begin{aligned} 0 &= \mathrm{dim}\mathcal{S}(Z^\star_\lambda) + \mathrm{dim}\mathcal{N}(\mathcal{A}) - \mathrm{dim}(\mathcal{S}(Z^\star_\lambda) + \mathcal{N}(\mathcal{A})) \\ &= \mathrm{dim}\mathcal{S}(Z^\star_\lambda) + \mathrm{dim}(\mathbb{S}^{(m+n)}) - \mathrm{dim}\mathcal{R}(\mathcal{A}) - \mathrm{dim}(\mathcal{S}(Z^\star_\lambda) + \mathcal{N}(\mathcal{A})) \\ &= \mathrm{dim}\mathcal{S}(Z^\star_\lambda) - KN + \mathrm{dim}(\mathbb{S}^{(m+n)}) - \mathrm{dim}(\mathcal{S}(Z^\star_\lambda) + \mathcal{N}(\mathcal{A})) \\ &= \mathrm{dim}\mathcal{S}(Z^\star) - KN + \mathrm{dim}(\mathcal{S}(Z^\star)^\perp \cap \mathcal{R}(\mathcal{A})) \\ &\geq \mathrm{dim}\mathcal{S}(Z^\star_\lambda) - KN \end{aligned}$$

Now we prove that $\dim\mathcal{S}(Z_\lambda^\star) = \frac{r(r+1)}{2}$ to complete the proof. Consider the diagonalization $Z_\lambda^\star = U\Lambda U^\intercal$ where U is a orthogonal matrix. Since the dimension of the subspace is invariant under orthogonal transformations, we have

$$\dim\mathcal{S}(Z_\lambda^\star) = \dim\mathcal{S}(\Lambda) = \dim\{Z \in \mathbb{S}^{(m+n)} : \mathcal{R}(Z) \subseteq \mathcal{R}(\Lambda)\}$$

where Λ is diagonal matrix with nontrivial entries in the leading principle minor of size $r \times r$. This restricts the symmetric matrix Z to have nontrivial entries only in the leading $r \times r$ block. Hence, $\dim\mathcal{S}(Z_\lambda^\star) = \frac{r(r+1)}{2}$. □

B. Omitted proof of Lemma 4.5

We prove Lemma 4.5 in this section.

Proof of Lemma 4.5. Let $\Pi_{V^\perp} : \mathbb{R}^d \to V^\perp$ be the orthogonal projection onto the orthogonal complement of V in $\mathbb{R}^d$. Then, $\Pi_{V^\perp}|_{\mathcal{M}} : \mathcal{M} \to V^\perp$ is a smooth mapping between manifolds. Since

$$\dim V^\perp = d - n > m = \dim\mathcal{M},$$

p is singular for all $p \in \mathcal{M}$. Therefore $\Pi_{V^\perp}(\mathcal{M})$ has measure zero in $\mathbb{R}^{d-n}$ by Sard's theorem. Note that $\mathcal{M} + V \subseteq \Pi_{V^\perp}(\mathcal{M}) + V$ and the measure of $\Pi_{V^\perp}(\mathcal{M}) + V$ in $\mathbb{R}^d$ is zero. This concludes that $\mathcal{M} + V$ is measure-zero in $\mathbb{R}^d$. □

As a remark, the prior works of (Boumal et al., 2016; Du & Lee, 2018) also use dimension-counting arguments that would warrant the use of Lemma 4.5, but they do not provide a precise justification. Our Theorem 4.1 makes a similar argument, but does so fully rigorous through Lemma 4.5.

C. Generalization guarantee

In this section, let $\ell(\cdot,\cdot)$ be our loss function which is convex, non-negative, and twice-differentiable on the first argument. Then, our empirical risk is

$$\hat{L}(\boldsymbol{\delta}) = \frac{1}{N}\sum_{i=1}^{N} \ell\left(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta}\rangle, Y_i\right).$$

We start the analysis from this non-regularized risk and expand it to regularized ones. We assume that our model is class of affine predictors $X \mapsto f_{\mathbf{W}_0}(X) + \langle \mathbf{G}(X), \boldsymbol{\delta}\rangle$ for given data X. Now we apply the theory of Rademacher complexity to derive the upper bound of the generalization bound. To begin with, we start with introducing the classical result in probability theory from (McDiarmid et al., 1989) without proof.

Lemma C.1. *(McDiarmid inequality) Let $X_1, \ldots, X_N \in \mathcal{X}$ be i.i.d N random samples from dataset $\mathcal{X}$. Let $g : \mathcal{X}^N \to \mathbb{R}$ be a function satisfying the following property with $c > 0$:*

$$|g(X_1, \ldots, X_{i-1}, X_i, X_{i+1}, \ldots, X_N) - g(X_1, \ldots, X_{i-1}, X_i', X_{i+1}, \ldots, X_N)| \le c$$

for all $X_1, \ldots, X_N, X_i' \in \mathcal{X}$. Then, for all $\varepsilon > 0$,

$$\mathbb{P}\left(|g(X_1, \ldots, X_N) - \mathbb{E}[g(X_1, \ldots, X_N)]| \ge \varepsilon\right) \le \exp\left(-\frac{2\varepsilon^2}{Nc^2}\right).$$

Now, we define the *Rademacher complexity* of the class of functions $\mathcal{H}$ from $\mathcal{X}$ to $\mathbb{R}$:

$$R_N(\mathcal{H}) = \mathbb{E}_{\varepsilon,\mathcal{D}}\left(\sup_{h\in\mathcal{H}} \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i h(X_i)\right),$$

where $\{\varepsilon_i\}_{1\le i\le N}$ are independent Rademacher random variables, and $\mathcal{D} = \{X_1, \ldots, X_N\}$ is N random samples from $\mathcal{X}$. In our analysis, we will focus on class of affine predictors $X_i \mapsto f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta}\rangle$ and composition of affine predictors with loss $X_i \mapsto \ell(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta}\rangle, Y_i)$. Rademacher complexities are closely related to upper bounds on generalization bound due to the following lemma.

Lemma C.2. *Let $R_N(\mathcal{H})$ be the Rademacher complexity of the class of functions $\mathcal{H}$ from $\mathcal{X}$ to $\mathbb{R}$ and $X_1, \ldots, X_N$ are N samples from $\mathcal{X}$. Then the following inequality holds.*

$$\mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\frac{1}{N}\sum_{i=1}^{N} h(X_i) - \mathbb{E}[h(X)]\right)\right] \leq 2R_N(\mathcal{H}), \quad \mathbb{E}\left[\sup_{h\in\mathcal{H}}\left(\mathbb{E}[h(X)] - \frac{1}{N}\sum_{i=1}^{N} h(X_i)\right)\right] \leq 2R_N(\mathcal{H}).$$

Proof. The proof is by using standard symmetrization arguments. We defer its proof to Theorem 8 of (Bartlett & Mendelson, 2002), or Section 4.5 of (Bach, 2023). □

The next lemma uses a contraction property to reduce the Rademacher complexity of losses to linear predictors. These type of results are widely used in Rademacher analysis and we use the following specific version of contraction, which was originally introduced in Corollary 4 of (Maurer, 2016) and adapted to our setting. Write $\|\cdot\|_2$ for Euclidean vector norm.

Lemma C.3. *Let $\mathcal{A}$ be the class of functions $a : \mathcal{X} \to \mathbb{R}^K$. For $1 \leq i \leq N$, let $\ell_i \colon \mathbb{R}^K \to \mathbb{R}$ be G-Lipschitz continuous on $\mathcal{A}$ with respect to the Euclidean norm in the sense that the following holds:*

$$|\ell_i(a(X_1)) - \ell_i(a'(X_2))| \leq G\|a(X_1) - a'(X_2)\|_2 \qquad \text{for any } a, a' \in \mathcal{A}, \quad X_1, X_2 \in \mathcal{X}.$$

Then we have the following inequality for independent Rademacher random variables $\{\sigma_i\}_{1\leq i\leq N}$ and $\{\varepsilon_{ij}\}_{1\leq i\leq N, 1\leq j\leq K}$:

$$\mathbb{E}_{\sigma,\mathcal{D}}\left[\sup_{a\in\mathcal{A}} \frac{1}{N}\sum_{i=1}^{N} \sigma_i \ell_i(a(X_i))\right] \leq \sqrt{2}G \cdot \mathbb{E}_{\varepsilon,\mathcal{D}}\left[\sup_{a\in\mathcal{A}} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} \varepsilon_{ij} a_j(X_i)\right],$$

where a_j denotes the j-th coordinate of a and $\mathcal{D} = \{(X_i, Y_i)\}_{i\in\{1,\ldots,N\}}$ are i.i.d N random samples sampled from $\mathcal{X}$.

Proof. We defer the proof to the Section 5 of (Maurer, 2016). □

In Lemma C.3, if we sample $\mathcal{D}$ from a probability distribution $\mathcal{P}$, we can relax the Lipschitz continuity condition to hold for $\mathcal{P}$- almost surely. In other words,

$$|\ell(a(X_1)) - \ell(a'(X_2))| \leq G\|a(X_1) - a'(X_2)\|_2 \qquad \text{for any } a, a' \in \mathcal{A}, \quad X_1, X_2 \subseteq \mathcal{D} \sim \mathcal{P}.$$

The next lemma states that the Rademacher complexity of class of bounded affine predictors decays at most $\mathcal{O}(\frac{1}{\sqrt{N}})$ rate.

Lemma C.4. *Assume $\mathcal{D} = \{(X_i, Y_i)\}_{i\in\{1,\ldots,N\}}$ is i.i.d N random samples sampled from probability distribution $\mathcal{P}$. Assume $\mathcal{A}_D = \{X_i \mapsto f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta}\rangle \in \mathbb{R}^K : \|\boldsymbol{\delta}\|_* \leq D, \boldsymbol{\delta} \in \mathbb{R}^{m\times n}\}$ is class of affine predictors with bounded nuclear norm $D > 0$. Suppose $\|\mathbf{G}^{(j)}(X_i)\|_F \leq R$ almost surely with respect to the random data $X_i \sim \mathcal{P}$. Then,*

$$\mathbb{E}_{\varepsilon,\mathcal{D}}\left[\sup_{a\in\mathcal{A}} \frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K} \varepsilon_{ij} a_j(X_i)\right] \leq \frac{RD\sqrt{K}}{\sqrt{N}}$$

where $\{\varepsilon_{ij}\}_{1\leq i\leq N, 1\leq j\leq K}$ are i.i.d Rademacher random variables.

Proof.

$$
\begin{aligned}
\mathbb{E}_\varepsilon\left[\sup_{a\in\mathcal{A}}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\varepsilon_{ij}a_j(X_i)\right] &= \mathbb{E}_\varepsilon\left[\sup_{\|\boldsymbol{\delta}\|_*\le D}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\varepsilon_{ij}\left(f^{(j)}_{\mathbf{W_0}}(X_i)+\langle\mathbf{G}^{(j)}(X_i),\boldsymbol{\delta}\rangle\right)\right]\\
&= \mathbb{E}_\varepsilon\left[\sup_{\|\boldsymbol{\delta}\|_*\le D}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\varepsilon_{ij}\langle\mathbf{G}^{(j)}(X_i),\boldsymbol{\delta}\rangle+\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\varepsilon_{ij}f^{(j)}_{\mathbf{W_0}}(X_i)\right]\\
&= \mathbb{E}_\varepsilon\left[\sup_{\|\boldsymbol{\delta}\|_*\le D}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\varepsilon_{ij}\langle\mathbf{G}^{(j)}(X_i),\boldsymbol{\delta}\rangle\right]\\
&\le \mathbb{E}_\varepsilon\left[\sup_{\|\boldsymbol{\delta}\|_F\le D}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\varepsilon_{ij}\langle\mathbf{G}^{(j)}(X_i),\boldsymbol{\delta}\rangle\right]\\
&= \mathbb{E}_\varepsilon\left[\frac{D}{N}\sup_{\|\boldsymbol{\delta}\|_F\le 1}\sum_{i=1}^{N}\sum_{j=1}^{K}\varepsilon_{ij}\langle\mathbf{G}^{(j)}(X_i),\boldsymbol{\delta}\rangle\right]\\
&= \frac{D}{N}\mathbb{E}_\varepsilon\left\|\sum_{i=1}^{N}\sum_{j=1}^{K}\varepsilon_{ij}\mathbf{G}^{(j)}(X_i)\right\|_F.
\end{aligned}
$$

The inequality is from the fact that $\|\cdot\|_F\le\|\cdot\|_*$, hence $\{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_*\le D\}\subset\{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|_F\le D\}$. The last equality is from the fact that $\|\cdot\|_F$ is self-dual. Next, we can bound $\mathbb{E}_\varepsilon\left\|\sum_{i=1}^{N}\sum_{j=1}^{K}\varepsilon_{ij}\mathbf{G}^{(j)}(X_i)\right\|_F$ by the following inequalities.

$$
\begin{aligned}
\mathbb{E}_\varepsilon\left\|\sum_{i=1}^{N}\sum_{j=1}^{K}\varepsilon_{ij}\mathbf{G}^{(j)}(X_i)\right\|_F &\le \sqrt{\mathbb{E}_\varepsilon\left\|\sum_{i=1}^{N}\sum_{j=1}^{K}\varepsilon_{ij}\mathbf{G}^{(j)}(X_i)\right\|_F^2}\\
&= \sqrt{\mathbb{E}_\varepsilon\sum_{i=1}^{N}\sum_{j=1}^{K}\left\|\varepsilon_{ij}\mathbf{G}^{(j)}(X_i)\right\|_F^2}\\
&= \sqrt{\sum_{i=1}^{N}\sum_{j=1}^{K}\left\|\mathbf{G}^{(j)}(X_i)\right\|_F^2}\\
&\le R\sqrt{NK}. \quad \text{a.s.}
\end{aligned}
$$

The first inequality is from Jensen's inequality, the equalities are from i.i.d assumption of ε_{ik}. We combine the results and take expectation with respect to $\mathcal{D}$ to get

$$
\mathbb{E}_{\varepsilon,\mathcal{D}}\left[\sup_{a\in\mathcal{A}}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\varepsilon_{ij}a_k(X_i)\right]\le\frac{D}{N}\cdot R\sqrt{NK}=\frac{RD\sqrt{K}}{\sqrt{N}}.
$$

□

We then combine the previous results to get the following Lemma.

Lemma C.5. *Assume $\mathcal{D}=\{(X_i,Y_i)\}_{i\in\{1,\dots,N\}}$ is i.i.d N random samples sampled from probability distribution $\mathcal{P}$. Let $\hat{L}$ is non-regularized empirical risk defined as*

$$
\hat{L}(\boldsymbol{\delta})=\frac{1}{N}\sum_{i=1}^{N}\ell\left(f_{\mathbf{W}_0}(X_i)+\langle\mathbf{G}(X_i),\boldsymbol{\delta}\rangle,Y_i\right)
$$

and $\mathcal{A}_D = \{X_i \mapsto f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta}\rangle \in \mathbb{R}^K : \|\boldsymbol{\delta}\|_* \leq D, \boldsymbol{\delta} \in \mathbb{R}^{m\times n}\}$ *is class of affine predictors with bounded nuclear norm* D*. For* $1 \leq j \leq K$*, suppose* $\|\mathbf{G}^{(j)}(X)\|_F \leq R$ *almost surely with respect to the random data* $X_i \sim \mathcal{P}$*. For* $1 \leq i \leq N$*, suppose* $\ell_i \triangleq \ell(\cdot, Y_i)$ *is* G*-Lipschitz continuous on* $\mathcal{A}$ *on the first argument (with respect to the Euclidean norm) for almost surely with respect to the random data* $X_i \subseteq \mathcal{D} \sim \mathcal{P}$*. That is,*

$$|\ell_i(a(X_1)) - \ell_i(a'(X_2))| \leq G\|a(X_1) - a'(X_2)\|_2 \qquad \text{for any } a, a' \in \mathcal{A}, \quad X_1, X_2 \subseteq \mathcal{D} \sim \mathcal{P}.$$

Then for any $\|\boldsymbol{\delta}\|_* \leq D$*, fixed* $\boldsymbol{\delta}_0$ *such that* $\|\boldsymbol{\delta}_0\|_* \leq D$*, and* $\eta \in (0,1)$*, the following inequality holds with probability greater than* $1-\eta$*:*

$$\hat{L}(\boldsymbol{\delta}_0) - \hat{L}(\boldsymbol{\delta}) - L(\boldsymbol{\delta}_0) + L(\boldsymbol{\delta}) < \frac{\sqrt{2K}GRD}{\sqrt{N}}\left(2 + \sqrt{\log\frac{1}{\eta}}\right).$$

Proof. Take g of Lemma C.1 to be $g = \sup_{\|\boldsymbol{\delta}\|_* \leq D}(\hat{L}(\boldsymbol{\delta}_0) - \hat{L}(\boldsymbol{\delta}) - L(\boldsymbol{\delta}_0) + L(\boldsymbol{\delta}))$, which is a function of $X_1, \ldots, X_N$. Since $\|\boldsymbol{\delta}\|_* \leq D$ implies $\|\boldsymbol{\delta}\|_F \leq D$ and by the Lipschitz continuity of $\ell(\cdot, Y_i)$, we have the following for any $(X_i, Y_i) \in \mathcal{D}$:

$$\begin{aligned}
|\ell\left(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta}_0\rangle, Y_i\right) - \ell\left(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta}\rangle, Y_i\right)| &\leq G\|\langle \boldsymbol{\delta}_0 - \boldsymbol{\delta}, \mathbf{G}(X_i)\rangle\|_2 \\
&\leq G\sqrt{\sum_{j=1}^{K}\|\boldsymbol{\delta}_0 - \boldsymbol{\delta}\|_F^2\|\mathbf{G}^{(j)}(X_i)\|_F^2} \\
&\leq G\sqrt{\sum_{j=1}^{K}\|\boldsymbol{\delta}_0 - \boldsymbol{\delta}\|_*^2\|\mathbf{G}^{(j)}(X_i)\|_F^2} \\
&\leq G\sqrt{\sum_{j=1}^{K}4D^2 \cdot R^2} \\
&= 2GRD\sqrt{K}.
\end{aligned}$$

Hence if we change only one data point (X_i, Y_i) of g to $(X_i^{'}, Y_i^{'})$, the deviation of $\hat{L}(\boldsymbol{\delta}_0) - \hat{L}(\boldsymbol{\delta})$ is at most $\frac{2GRD\sqrt{K}}{N}$. Then by Lemma C.1, we have

$$\sup_{\|\boldsymbol{\delta}\|_* \leq D}(\hat{L}(\boldsymbol{\delta}_0) - \hat{L}(\boldsymbol{\delta}) - L(\boldsymbol{\delta}_0) + L(\boldsymbol{\delta})) < \mathbb{E}\left[\sup_{\|\boldsymbol{\delta}\|_* \leq D}(\hat{L}(\boldsymbol{\delta}_0) - \hat{L}(\boldsymbol{\delta}) - L(\boldsymbol{\delta}_0) + L(\boldsymbol{\delta}))\right] + \frac{t\sqrt{2K}GRD}{\sqrt{N}}$$

with probability greater than $1 - e^{-t^2}$. The expectation on the right hand side can be reduced to

$$\begin{aligned}
\mathbb{E}_{\mathcal{D}}\left[\sup_{\|\boldsymbol{\delta}\|_* \leq D}(\hat{L}(\boldsymbol{\delta}_0) - \hat{L}(\boldsymbol{\delta}) - L(\boldsymbol{\delta}_0) + L(\boldsymbol{\delta}))\right] &= \mathbb{E}_{\mathcal{D}}\left[\sup_{\|\boldsymbol{\delta}\|_* \leq D}(-\hat{L}(\boldsymbol{\delta}) + L(\boldsymbol{\delta})) + \hat{L}(\boldsymbol{\delta}_0) - L(\boldsymbol{\delta}_0)\right] \\
&= \mathbb{E}_{\mathcal{D}}\left[\sup_{\|\boldsymbol{\delta}\|_* \leq D}(L(\boldsymbol{\delta}) - \hat{L}(\boldsymbol{\delta}))\right]
\end{aligned}$$

Note that

$$\begin{aligned}
L(\boldsymbol{\delta}) - \hat{L}(\boldsymbol{\delta}) &= L(\boldsymbol{\delta}) - \frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta}\rangle, Y_i) \\
&= \mathbb{E}\Big[\ell(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta}\rangle, Y_i)\Big] - \frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta}\rangle, Y_i),
\end{aligned}$$

where the expectation is taken over $X_i \sim \mathcal{P}$. Now apply Lemma C.2 to get

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[\sup_{\|\boldsymbol{\delta}\|_* \le D}(L(\boldsymbol{\delta}) - \hat{L}(\boldsymbol{\delta}))\right] &= \mathbb{E}_{\mathcal{D}}\left[\sup_{\|\boldsymbol{\delta}\|_* \le D}(\mathbb{E}\Big[\ell(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta}\rangle, Y_i)\Big]\right] \\ &\quad - \mathbb{E}_{\mathcal{D}}\left[\frac{1}{N}\sum_{i=1}^{N}\ell(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta}\rangle, Y_i))\right] \\ &\le 2\mathbb{E}_{\sigma,\mathcal{D}}\left[\sup_{\|\boldsymbol{\delta}\|_* \le D}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\ell(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta}\rangle, Y_i))\right]\end{aligned}$$

where $\{\sigma\}_{1\le i\le N}$ are i.i.d Rademacher variables. Then apply Lemma C.3 to get

$$\begin{aligned}&2\mathbb{E}_{\sigma,\mathcal{D}}\left[\sup_{\|\boldsymbol{\delta}\|_* \le D}\frac{1}{N}\sum_{i=1}^{N}\sigma_i\ell(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta}\rangle, Y_i))\right] \\ &= 2\sqrt{2}G\mathbb{E}_{\varepsilon,\mathcal{D}}\left[\sup_{\|\boldsymbol{\delta}\|_* \le D}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\varepsilon_{ij}\left(f^j_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}^j(X_i), \boldsymbol{\delta}\rangle\right)\right]\end{aligned}$$

where $\{\varepsilon_{ij}\}_{1\le i\le N, 1\le j\le K}$ are i.i.d Rademacher random variables. Finally, use Lemma C.4 to get

$$2\sqrt{2}G\mathbb{E}_{\varepsilon,\mathcal{D}}\left[\sup_{\|\boldsymbol{\delta}\|_* \le D}\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{K}\varepsilon_{ij}\left(f^j_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}^j(X_i), \boldsymbol{\delta}\rangle\right)\right] \le 2\sqrt{2}G\cdot\frac{RD\sqrt{K}}{\sqrt{N}}.$$

Therefore, we conclude that

$$\hat{L}(\boldsymbol{\delta}_0) - \hat{L}(\boldsymbol{\delta}) - L(\boldsymbol{\delta}_0) + L(\boldsymbol{\delta}) < \frac{\sqrt{2K}GRD}{\sqrt{N}}(2+t).$$

for $\|\boldsymbol{\delta}\|_* \le D$ with probability greater than $1 - e^{-t^2}$. By reparametrization, we get

$$\hat{L}(\boldsymbol{\delta}_0) - \hat{L}(\boldsymbol{\delta}) - L(\boldsymbol{\delta}_0) + L(\boldsymbol{\delta}) < \frac{\sqrt{2K}GRD}{\sqrt{N}}\left(2 + \sqrt{\log\frac{1}{\eta}}\right).$$

for $\|\boldsymbol{\delta}\|_* \le D$ with probability greater than $1 - \eta$. □

Now we can extend this generalization guarantee of constrained optimization to regularized optimization, which aligns with our problem of interest. For notational convenience, let

$$L_\lambda(\boldsymbol{\delta}) = L(\boldsymbol{\delta}) + \lambda\|\boldsymbol{\delta}\|_*, \quad \hat{L}_\lambda(\boldsymbol{\delta}) = \hat{L}(\boldsymbol{\delta}) + \lambda\|\boldsymbol{\delta}\|_*$$

We follow the proof structure of (Bach, 2023), which was motivated by (Bartlett et al., 2005) and (Sridharan et al., 2008).

Theorem C.6. *Fix $\varepsilon > 0$ and let $0 \ne \boldsymbol{\delta}^\star_{\text{true}} \in \operatorname{argmin}_{\boldsymbol{\delta}} L(\boldsymbol{\delta})$ be the true optimum of the population risk and consider the setup of Lemma C.5 with $D = (2+\varepsilon)\|\boldsymbol{\delta}^\star_{\text{true}}\|_*$, which is the upper bound on the nuclear norm of the predictors. Let $\eta \in (0,1)$ and*

$$\lambda = \frac{(2+\varepsilon)\sqrt{2K}GR}{\sqrt{N}}\left(2 + \sqrt{\log\frac{1}{\eta}}\right).$$

Write $\boldsymbol{\delta}^\star_\lambda$ to denote a minimizer (not necessarily unique) of $\hat{L}_\lambda(\boldsymbol{\delta})$.Consider the setup of Corollary 4.2 with P randomly sampled with a probability distribution supported in

$$\left\{P \in \mathbb{S}^{(m+n)}_+ : \|P\|_F < \frac{\varepsilon\lambda\|\boldsymbol{\delta}^\star_{\text{true}}\|_*}{2\|\boldsymbol{\delta}^\star_\lambda\|_*}\right\}$$

and is absolutely continuous with respect to the Lebesgue measure on $\mathbb{S}^{(m+n)} \cong \mathbb{R}^{\frac{(m+n)(m+n+1)}{2}}$. Let $(\hat{\mathbf{u}}, \hat{\mathbf{v}})$ be an SOSP of $\hat{L}_{\lambda,P}$. Then with probability greater than $1-\eta$,

$$L(\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal) - L(\boldsymbol{\delta}^\star_{\text{true}}) < \|\boldsymbol{\delta}^\star_{\text{true}}\|_*\frac{(2+\varepsilon)^2\sqrt{2K}GR}{\sqrt{N}}\left(2 + \sqrt{\log\frac{1}{\eta}}\right).$$

Proof. Let $\tilde{\varepsilon} = \frac{\varepsilon\lambda\|\boldsymbol{\delta}^\star_{\text{true}}\|_*}{2\|\boldsymbol{\delta}^\star_\lambda\|_*}$ and consider the convex set

$$C = \left\{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_* \le 2\|\boldsymbol{\delta}^\star_{\text{true}}\|_* + \frac{2\tilde{\varepsilon}}{\lambda}\|\boldsymbol{\delta}^\star_\lambda\|_*, L_\lambda(\boldsymbol{\delta}) - L_\lambda(\boldsymbol{\delta}^\star_{\text{true}}) \le \lambda\|\boldsymbol{\delta}^\star_{\text{true}}\|_* + 2\tilde{\varepsilon}\|\boldsymbol{\delta}^\star_\lambda\|_*\right\}.$$

Then for $\|\boldsymbol{\delta}\|_* = 2\|\boldsymbol{\delta}^\star_{\text{true}}\|_* + \frac{2\tilde{\varepsilon}}{\lambda}\|\boldsymbol{\delta}^\star_\lambda\|_*$, $\boldsymbol{\delta} \notin \text{int}C$ since the following inequalities hold.

$$L_\lambda(\boldsymbol{\delta}) - L_\lambda(\boldsymbol{\delta}^\star_{\text{true}}) = L(\boldsymbol{\delta}) - L(\boldsymbol{\delta}^\star_{\text{true}}) + \lambda\|\boldsymbol{\delta}\|_* - \lambda\|\boldsymbol{\delta}^\star_{\text{true}}\|_* \ge \lambda\|\boldsymbol{\delta}\|_* - \lambda\|\boldsymbol{\delta}^\star_{\text{true}}\|_* = \lambda\|\boldsymbol{\delta}^\star_{\text{true}}\|_* + 2\tilde{\varepsilon}\|\boldsymbol{\delta}^\star_\lambda\|_*.$$

Therefore the boundary ∂C of C should be

$$\partial C = \left\{\boldsymbol{\delta} : \|\boldsymbol{\delta}\|_* \le 2\|\boldsymbol{\delta}^\star_{\text{true}}\|_* + \frac{2\tilde{\varepsilon}}{\lambda}\|\boldsymbol{\delta}^\star_\lambda\|_*, L_\lambda(\boldsymbol{\delta}) - L_\lambda(\boldsymbol{\delta}^\star_{\text{true}}) = \lambda\|\boldsymbol{\delta}^\star_{\text{true}}\|_* + 2\tilde{\varepsilon}\|\boldsymbol{\delta}^\star_\lambda\|_*\right\}.$$

Now suppose $\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal \notin C$. Then since $\boldsymbol{\delta}^\star_{\text{true}} \in C$, there exists $\boldsymbol{\delta}$ in the segment $[\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal, \boldsymbol{\delta}^\star_{\text{true}}]$ such that $\boldsymbol{\delta} \in \partial C$. By the convexity of $\hat{L}_\lambda$, we have

$$\hat{L}_\lambda(\boldsymbol{\delta}) \le \max\left(\hat{L}_\lambda(\boldsymbol{\delta}^\star_{\text{true}}), \hat{L}_\lambda(\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal)\right).$$

Then we get

$$\hat{L}_\lambda(\boldsymbol{\delta}^\star_{\text{true}}) - \hat{L}_\lambda(\boldsymbol{\delta}) \ge -2\tilde{\varepsilon}\|\boldsymbol{\delta}^\star_\lambda\|_*$$

by Corollary 4.2. Therefore,

$$\begin{aligned}\hat{L}(\boldsymbol{\delta}^\star_{\text{true}}) - \hat{L}(\boldsymbol{\delta}) - L(\boldsymbol{\delta}^\star_{\text{true}}) + L(\boldsymbol{\delta}) &= \hat{L}_\lambda(\boldsymbol{\delta}^\star_{\text{true}}) - \hat{L}_\lambda(\boldsymbol{\delta}) - L_\lambda(\boldsymbol{\delta}^\star_{\text{true}}) + L_\lambda(\boldsymbol{\delta}) \\ &\ge L_\lambda(\boldsymbol{\delta}) - L_\lambda(\boldsymbol{\delta}^\star) - 2\tilde{\varepsilon}\|\boldsymbol{\delta}^\star_\lambda\|_* \\ &= \lambda\|\boldsymbol{\delta}^\star_{\text{true}}\|_*\end{aligned} \tag{3}$$

Note that $\|\boldsymbol{\delta}\|_* \le 2\|\boldsymbol{\delta}^\star_{\text{true}}\|_* + \frac{2\tilde{\varepsilon}}{\lambda}\|\boldsymbol{\delta}^\star_\lambda\|_* < (2+\varepsilon)\|\boldsymbol{\delta}^\star_{\text{true}}\|_*$ and

$$\lambda\|\boldsymbol{\delta}^\star_{\text{true}}\|_* = \|\boldsymbol{\delta}^\star_{\text{true}}\|_*\frac{(2+\varepsilon)\sqrt{2K}GR}{\sqrt{N}}\left(2 + \sqrt{\log\frac{1}{\eta}}\right).$$

Then by Lemma C.5, (3) should happen with probability less than η. Then with probability greater than $1-\eta$, $\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal \in C$. In other words,

$$L_\lambda(\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal) - L_\lambda(\boldsymbol{\delta}^\star_{\text{true}}) < \lambda\|\boldsymbol{\delta}^\star_{\text{true}}\|_* + 2\tilde{\varepsilon}\|\boldsymbol{\delta}^\star_\lambda\|_*.$$

Hence,

$$\begin{aligned}L(\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal) + \lambda\|\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal\|_* &< L_\lambda(\boldsymbol{\delta}^\star_{\text{true}}) + \lambda\|\boldsymbol{\delta}^\star_{\text{true}}\|_* + 2\tilde{\varepsilon}\|\boldsymbol{\delta}^\star_\lambda\|_* \\ &= L(\boldsymbol{\delta}^\star_{\text{true}}) + 2\lambda\|\boldsymbol{\delta}^\star_{\text{true}}\|_* + 2\tilde{\varepsilon}\|\boldsymbol{\delta}^\star_\lambda\|_* \\ &\le L(\boldsymbol{\delta}^\star_{\text{true}}) + 2\lambda\|\boldsymbol{\delta}^\star_{\text{true}}\|_* + \varepsilon\lambda\|\boldsymbol{\delta}^\star_{\text{true}}\|_*.\end{aligned}$$

Finally, we get

$$L(\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal) - L(\boldsymbol{\delta}^\star_{\text{true}}) < \|\boldsymbol{\delta}^\star_{\text{true}}\|_*\frac{(2+\varepsilon)^2\sqrt{2K}GR}{\sqrt{N}}\left(2 + \sqrt{\log\frac{1}{\delta}}\right).$$

□

By using the fact that ℓ^{CE} is Lipschitz continuous, we can reduce Theorem C.6 to Theorem 5.1. Note that the loss function ℓ may not be Lipschitz continuous in general. However, Lipschitz continuity is a mild assumption when the domain is restricted to a bounded class of predictors $\mathcal{A}_D$ of Lemma C.5.

Proof of Theorem 5.1. If $\ell(\cdot, Y)\colon \mathbb{R}^K \to \mathbb{R}$ is cross entropy loss defined as

$$\ell(X, Y) = \ell^{CE}(X, Y) = -\log\left(\frac{\exp X^{(j)}}{\sum_{i=1}^K \exp X^{(i)}}\right) = -X^{(j)} + \log\left(\sum_{i=1}^K \exp X^{(i)}\right)$$

with true label $Y = j$, we have

$$\nabla \ell^{CE}(X, Y)_j = -1 + \frac{\exp X^{(j)}}{\sum_{i=1}^{K} \exp X^{(i)}} = -\frac{\sum_{i \neq j} \exp X^{(Y)}}{\sum_{i=1}^{K} \exp X^{(i)}}$$

and for $k \neq j$,

$$\nabla \ell^{CE}(X, Y)_k = \frac{\exp X^{(k)}}{\sum_{i=1}^{K} \exp X^{(Y)}}$$

Then we can bound the Euclidean norm of the gradient as follows.

$$\|\nabla \ell^{CE}(X, Y)\|_2^2 = \frac{\left(\sum_{i \neq j} \exp X^{(i)}\right)^2}{\left(\sum_{i=1}^{K} \exp X^{(i)}\right)^2} + \frac{\sum_{i \neq j} \exp 2X^{(k)}}{\left(\sum_{i=1}^{K} \exp X^{(i)}\right)^2} \leq 1 + 1 = 2.$$

Hence the gradient of the cross entropy loss is bounded by $\sqrt{2}$ and we may replace G in Theorem C.6 with $\sqrt{2}$ to get

$$L(\hat{\mathbf{u}}\hat{\mathbf{v}}^\intercal) - L(\boldsymbol{\delta}^\star_{\text{true}}) < \|\boldsymbol{\delta}^\star_{\text{true}}\|_* \frac{2(2+\varepsilon)^2 \sqrt{K} R}{\sqrt{N}} \left(2 + \sqrt{\log \frac{1}{\delta}}\right).$$

□

D. Details of experiments

Optimizing nuclear norm. Recall that SGD or GD on the loss function with weight decay and with regularization parameter λ is equivalent to minimizing

$$\frac{1}{N} \sum_{i=1}^{N} \ell\left(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \mathbf{u}\mathbf{v}^\intercal \rangle, Y_i\right) + \frac{\lambda}{2}\|\mathbf{u}\|_F^2 + \frac{\lambda}{2}\|\mathbf{v}\|_F^2,$$

with respect to $\mathbf{u}$ and $\mathbf{v}$. In full fine-tuning however, this is equivalent to minimize the following with respect to $\boldsymbol{\delta}$:

$$\frac{1}{N} \sum_{i=1}^{N} \ell\left(f_{\mathbf{W}_0}(X_i) + \langle \mathbf{G}(X_i), \boldsymbol{\delta} \rangle, Y_i\right) + \lambda \|\boldsymbol{\delta}\|_*.$$

The problem here is that gradient methods no longer apply since the nuclear norm is non-differentiable. Therefore, we use the proximal gradient method:

$$\boldsymbol{\delta}_{t+1} = \mathbf{prox}_{\alpha\lambda\|\cdot\|_*}(\boldsymbol{\delta}_t - \alpha \nabla \hat{L}(\boldsymbol{\delta}_t))$$

where

$$\mathbf{prox}_{\alpha\lambda\|\cdot\|_*}(\boldsymbol{\delta}) = \underset{\boldsymbol{\delta}'}{\operatorname{argmin}} \left(\lambda \|\boldsymbol{\delta}'\|_* + \frac{1}{2\alpha}\|\boldsymbol{\delta}' - \boldsymbol{\delta}\|_F^2\right).$$

It is well known that the proximal gradient method on convex objective converges to a global minimum (Polyak, 1987).

Hyperparameters on NLP tasks For NLP tasks, we use full batch to perform GD on training. We only train the query (W_q) and value (W_v) weights of the RoBERTa-base model, which was empirically shown to have good performance (Hu et al., 2021). Furthermore, calculating the proximal operator of a nuclear norm is a computational bottleneck during the training of all W_q and W_v matrices. Therefore, we limit our training to only the last layer of W_q and W_v. To ensure a fair comparison, we apply the same approach to the LoRA updates. Additional information is in Table 1.

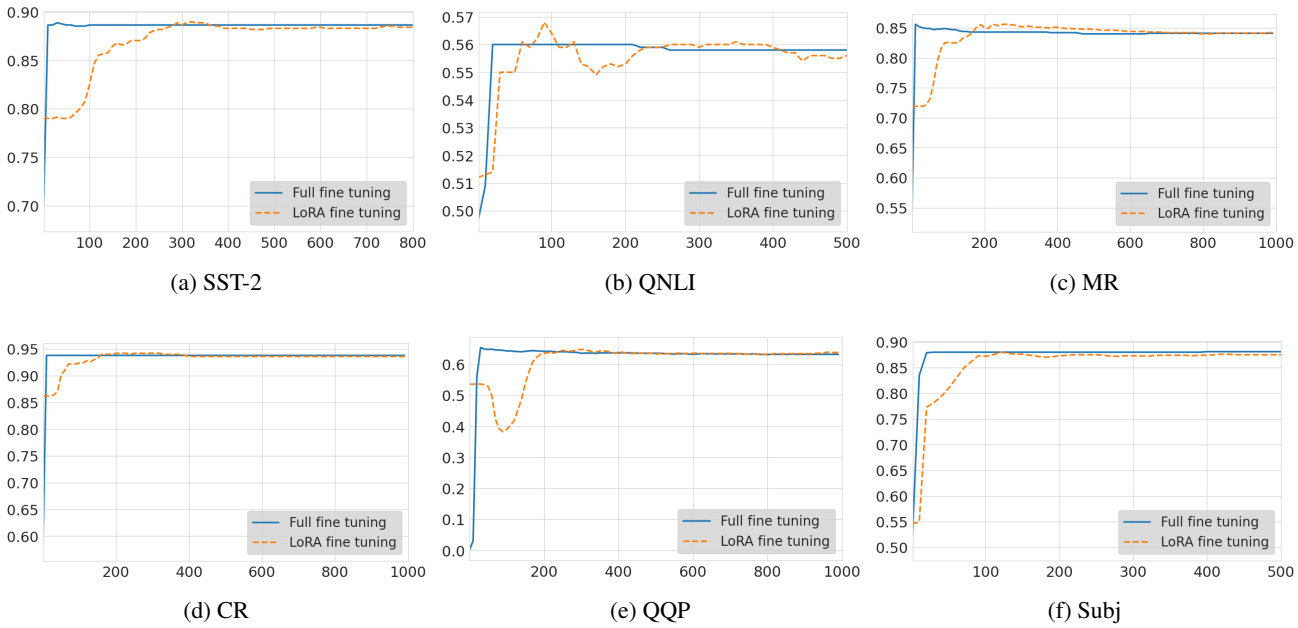
Hyperparameters on image and speech classification tasks Similar to NLP tasks, we train the last attention layers. Further details are in Table 2.

Task	**SST-2,QNLI**	**MR,CR,QQP,Subj**
Batch size	32	32
Learning rate (Full, LoRA fine tuning)	0.0005	0.001
Trained layer	W_q, W_v (last layer only)	W_q, W_v (last layer only)
Weight decay	0.01	0.01

Table 1. Hyperparameters on experiment in Section 6 (NLP tasks)

Task	**Image classification**	**Speech classification**
Batch size	16	16
Learning rate (Full, LoRA fine tuning)	0.005	0.005
Trained layer	W_q, W_v (last layer only)	W_q, W_v (last layer only)
Weight decay	0	0.001

Table 2. Hyperparameters on experiment in Section 6 (Image and speech classification tasks)



(a) SST-2 (b) QNLI (c) MR

(d) CR (e) QQP (f) Subj

Figure 4. Test curves (accuracy vs. epochs) on different NLP tasks. We used the LoRA rank of 16.

Test accuracy. For the setting of Section 6 on NLP tasks, we additionally conduct evaluations on a test set of 1000 samples during training and present the results in Figure 4. We observed that in most tasks the performance using LoRA eventually converges a test accuracy that matches that of full fine-tuning, although the rates of convergence sometimes differ. We list the hyperparameters in Table 3

Task	**SST-2,QQP,MR,CR**	**Subj**	**QNLI**
Batch size	32	32	24
Learning rate (Full, LoRA fine tuning)	0.0001	0.001	0.0005
Trained layer	W_q, W_v (all layers)	W_q, W_v (all layers)	W_q, W_v (all layers)
Weight decay	0.005	0.005	0.005

Table 3. Hyperparameters on experiment in Figure 4

For image and speech classification tasks, we also validate the performance of our linearized update to confirm that the accuracy is on par with actual LoRA updates. Accuracies are averaged over 3 runs (See Table 4).

Task	**Image classification**	**Speech classification**
Accuracy (actual / linearized)	86.20 / 87.00	74.67 / 73.67

Table 4. Accuaricies of LoRA updates on vision and speech classification tasks